

# Index of Examples

## CHAPTER 1

1. Three coins
2. Four of a kind
3. Prisoner's dilemma
4. Coin tossing game: HHH vs. TTHH
5. Geometric distribution
6. Three way duel

## CHAPTER 2

7. Fair price interpretation of expectations
8. Expected values for coin tossing game
9. Expected value for geometric distribution
10. Coupon collector problem
11. Urn experiment
12. Gambler's ruin: fair coin
13. Gambler's ruin: biased coin
14. Big pills, little pills

## CHAPTER 3

15. Binomials from coin tossing
16. Expected value of the Binomial distribution
17. Where to get mugged
18. Bayesian example
19. Bad checks

## CHAPTER 4

20. Polya urn model
21. The game of Bet Red
22. The ballot theorem

## CHAPTER 5

23. Independent versus uncorrelated
24. The Tchebychev inequality
25. Concentration of sample means
26. Variability of a sample average
27. Decomposition of variance

## CHAPTER 6

28. Functions of random variables with continuous distributions
29. Order statistics from the uniform distribution
30. The beta distribution
31. Binomial tail probabilities
32. Expectation of a random variable with a continuous distribution

## CHAPTER 7

33. Expected value and variance of the Binomial distribution

- 34. Normal approximation to the Binomial
- 35. The  $N(\mu, \sigma^2)$  distribution has expected value  $\mu$  and variance  $\sigma^2$ .
- 36. Standardization of the normal distribution
- 37. The box plot
- 38. Normal approximations in sampling

#### CHAPTER 8

- 39. Poisson( $np$ ) approximation to Binomial( $n, p$ )
- 40. Properties of the Poisson distribution
- 41. Poisson approximation with dependence

#### CHAPTER 9

- 42. Gamma distributions from Poisson process
- 43. Facts about the gamma function and gamma distribution
- 44. Gammas from normals
- 45. Conditioning on a rv with a continuous distribution
- 46. A queuing problem

#### CHAPTER 10

- 47. Joint densities for independent random variables
- 48. Joint densities for linear combinations
- 49. Linear combinations of independent normals
- 50. Betas from gammas
- 51. Sums of independent gamma random variables

#### CHAPTER 11

- 52. Conditional density from joint density
- 53. Transformation to polar coordinates

#### CHAPTER 12

- 54. The standard bivariate normal with correlation  $\rho$
- 55. Regression to the mean
- 56. Rotation of coordinate axes
- 57. Rotation to new coordinates: multivariate case
- 58. Independence of sample mean and sample variance

#### CHAPTER 13

- 59. Continuous analog of gambler's ruin problem
- 60. A model for stock prices
- 61. The Black-Scholes differential equation

## Chapter 1

# Probabilities and random variables

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

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.

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. (See Example 14 for an alternative interpretation.)

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) :  $0 \leq \mathbb{P}E \leq 1$  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  $E_1, E_2, \dots$  then  $\mathbb{P}E = \sum_i \mathbb{P}E_i$ .




---

**Example 1:** Find  $\mathbb{P}\{\text{at least two heads}\}$  for the tossing of three coins.

---

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

$$\mathbb{P}\{\text{at least two heads} \mid \text{coins fair, “independence,” } \dots\} = \dots$$

to indicate that the assignment is conditional on certain information (or assumptions). The vertical bar stands for the word *given*; that is, we read the symbol as *probability of at least two heads given that* ...

For fixed conditioning information, the **conditional probabilities**  $\mathbb{P}\{\dots \mid \text{info}\}$  satisfy rules (P1) through (P4). For example,  $\mathbb{P}(\emptyset \mid \text{info}) = 0$ , 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 the 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:

$$\mathbb{P}(\dots \mid \text{event } A \text{ AND event } B \text{ have occurred AND previous info})$$

or

$$\mathbb{P}(\dots \mid A \cap B \cap \text{previous info}) \quad \text{where } \cap \text{ means intersection}$$

or

$$\mathbb{P}(\dots \mid A, B, \text{previous info})$$

or

$$\mathbb{P}(\dots \mid A \cap B) \quad \text{if the “previous info” is understood.}$$

or

$$\mathbb{P}(\dots \mid AB) \quad \text{where } AB \text{ is an abbreviation for } A \cap B.$$

The commas in the third expression are open to misinterpretation, but convenience recommends the more concise notation.

REMARK. I must confess to some inconsistency in my use of parentheses and braces. If the “...” is a description in words, then  $\{\dots\}$  denotes the subset of  $S$  on which the description is true, and  $\mathbb{P}\{\dots\}$  or  $\mathbb{P}\{\dots \mid \text{info}\}$  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  $\mathbb{P}(A \cap B \mid \text{info})$  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  $\mathbb{P}(A \cap B \mid \text{info}) = \mathbb{P}(A \mid \text{info}) \cdot \mathbb{P}(B \mid A, \text{info})$ .

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)

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

Then, for big  $N$ ,

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

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 the event  $B$  also occurs. Thus  $\mathbb{P}(B \mid A, \text{info}) \approx N_{A \cap B}/N_A$ . The rest is division.

In my experience, conditional probabilities provide a more reliable method for solving problems traditionally handled by counting arguments (Combinatorics). I find it hard to be consistent about how I count, to make sure every case is counted once and only once, to decide whether order should matter, and so on. The next Example illustrates my point.

---

**Example 2:** What is the probability that a hand of 5 cards contains four of a kind?

---

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 *Fifty Challenging Problems in Probability* by Frederick Mosteller. This little book is one of my favourite sources for elegant examples. One could learn a lot of probability by trying to solve all fifty problems. The underlying problem has resurfaced in recent years in various guises, including a variation in the highly publicized “Ask Marylyn” incident.

---

**Example 3:** 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.

---

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, construction of a sample space would be a nontrivial exercise. The Example shows how conditioning can break a complex random mechanism into a sequence of simpler stages.

---

**Example 4:** 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, and R waits for tthh. The one whose pattern appears first is the winner. What is the probability that M wins?

---

In both Examples 3 and 4 we had situations where particular pieces of information could be ignored in the calculation of some conditional probabilities,

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

Both situations are instances of a property called **independence**.

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

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

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

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

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.

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 the concept. 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.

---

**Example 5:** Suppose a coin has probability  $p$  of landing heads on any particular toss, independent of the outcomes of other tosses. In a sequence of such tosses, what is the probability that the first head appears on the  $k$ th toss (for  $k = 1, 2, \dots$ )?

---

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

$$\mathbb{P}\{X = k\} = (1 - p)^{k-1}p \quad \text{for } k = 1, 2, \dots$$

The  $X$  is an example of a **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.

For example, we say that a random variable  $Z$  has a **geometric( $p$ ) distribution** if it can take values  $1, 2, 3, \dots$  with probabilities

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

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

REMARK. 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,

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

Notice, in particular, that if  $0 < p \leq 1$ , 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.

---

Example 6: 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?

---

### Things to remember

-  ,  , and the five rules for manipulating (conditional) probabilities.
- Conditioning is often easier, or at least more reliable, than counting.
- Conditional independence is a major simplifying assumption of probability theory.
- What is a random variable? What is meant by the distribution of a random variable?
- What is the geometric( $p$ ) distribution?

### EXAMPLE 1: THREE COINS

Find  $\mathbb{P}\{\text{at least two heads}\}$  for the tossing of three coins. Use the sample space

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

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

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

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

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

□



## EXAMPLE 2: FOUR OF A KIND

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.

I will consider the ordering of the cards within the hand as significant. For example,  $(7\clubsuit, 3\diamond, 2\heartsuit, K\heartsuit, 8\heartsuit)$  will be a different hand from  $(K\heartsuit, 7\clubsuit, 3\diamond, 2\heartsuit, 8\heartsuit)$ .

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.

$$\mathbb{P}\{\text{four of a kind}\} = \sum_{i=1}^{13} \mathbb{P}F_i = 13\mathbb{P}F_1 \quad \text{by rule P4.}$$

Now break  $F_1$  into simpler pieces,

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

where  $F_{1j} = \{\text{four aces with } j\text{th card not an ace}\}$ . 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

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

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

$$\begin{aligned} \mathbb{P}F_{1,1} &= \mathbb{P}N_1 \mathbb{P}(A_2 | N_1) \mathbb{P}(A_3 | N_1 A_2) \dots \mathbb{P}(A_5 | N_1 A_2 A_3 A_4) \\ &= \frac{48}{52} \times \frac{4}{51} \times \frac{3}{50} \times \frac{2}{49} \times \frac{1}{48}. \end{aligned}$$

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? □

Which sample space was I using, implicitly? How would the argument be affected if we took  $S$  as the set of all of all  $\binom{52}{5}$  distinct subsets of size 5, with equal probability on each sample point? That is, would it matter if we ignored ordering of cards within hands?

### EXAMPLE 3: PRISONER'S DILEMMA

(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

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

There are three events to be considered

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

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

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,

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

What is the split between  $\boxed{Bc}$  and  $\boxed{bC}$ ? 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

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

from which we deduce (via rule P5) that

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

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

Notice that

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

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

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

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

he would calculate a conditional probability

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

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.

#### EXAMPLE 4: COIN TOSSING GAME: HHH vs. TTHH

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 ththttthh... would result in a win for R, but ththhthhh... would result in a win for M.

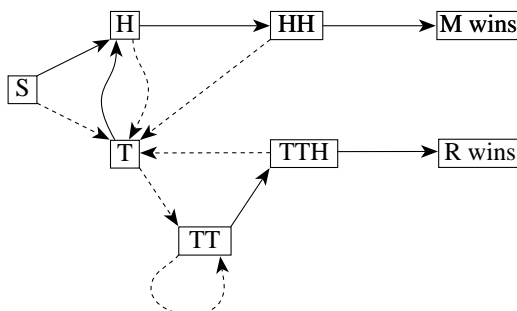
At first thought one might imagine that M has the advantage. After all, surely it must be easier to get a pattern of length 3 than a pattern of length 4. You'll discover that the solution is not that straightforward.

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
<b>S</b>	—	—
<b>H</b>	h	—
<b>T</b>	—	t
<b>TT</b>	—	tt
<b>HH</b>	hh	—
<b>TTH</b>	h	tth
<b>M wins</b>	hhh	?
<b>R wins</b>	?	tthh

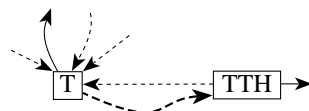
By claiming that these states summarize the game I am tacitly assuming that the coin has no “memory”, in the sense that the conditional probability of a head given any particular past sequence of heads and tails is  $1/2$  (for a fair coin). The past history leading to a particular state does not matter; 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 **S** to **H** to **HH** to **M wins** correspond to heads; the game would progress in exactly that way if the first three tosses gave hhh. Similarly the arrows from **S** to **T** to **TT** correspond to tails.

The arrow looping from  $\boxed{TT}$  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  $\boxed{T}$  to  $\boxed{TT}$  the step into  $\boxed{TTH}$  becomes inevitable. Indeed, for the purpose of calculating the probability that M wins, we could replace the side branch by:



The new arrow from  $\boxed{T}$  to  $\boxed{TTH}$  would correspond to a sequence of tails followed by a head. With the state  $\boxed{TT}$  removed, the diagram would become almost symmetric with respect to M and R. The arrow from  $\boxed{HH}$  back to  $\boxed{T}$  would show that R actually has an advantage: the first h in the thhh 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  $\boxed{S}$  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:

$$P_S = \mathbb{P}\{\text{reach } \boxed{\text{M wins}} \mid \text{start at } \boxed{S}\}$$

$$P_T = \mathbb{P}\{\text{reach } \boxed{\text{M wins}} \mid \text{start at } \boxed{T}\}$$

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.

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

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  $\boxed{S}$ , heads" has been turned into "start at  $\boxed{H}$ " and so on. We have our first equation:

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

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

$$P_H = \frac{1}{2}P_T + \frac{1}{2}P_{HH}$$

$$P_{HH} = \frac{1}{2} + \frac{1}{2}P_T$$

$$P_T = \frac{1}{2}P_H + \frac{1}{2}P_{TT}$$

$$P_{TT} = \frac{1}{2}P_{TT} + \frac{1}{2}P_{TTH}$$

$$P_{TTH} = \frac{1}{2}P_T + 0.$$

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

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

This simple elimination of the  $P_{TT}$  contribution corresponds to the excision of the  $\boxed{TT}$  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.  $\square$

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:

$$P = \begin{matrix} & \begin{matrix} \boxed{S} & \boxed{H} & \boxed{T} & \boxed{HH} & \boxed{TT} & \boxed{TTH} & \boxed{M \text{ wins}} & \boxed{R \text{ wins}} \end{matrix} \\ \begin{matrix} \boxed{S} \\ \boxed{H} \\ \boxed{T} \\ \boxed{HH} \\ \boxed{TT} \\ \boxed{TTH} \\ \boxed{M \text{ wins}} \\ \boxed{R \text{ wins}} \end{matrix} & \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

If we similarly define a column vector,

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

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

$$P\pi = \pi,$$

with the boundary conditions  $P_{M \text{ wins}} = 1$  and  $P_{R \text{ wins}} = 0$ . I didn't bother adding the equations  $P_{M \text{ wins}} = 1$  and  $P_{R \text{ wins}} = 0$  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  $P_{TTH}$ .

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  $\boxed{T}$ , gives transition probabilities from state  $\boxed{T}$ . 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  $\boxed{T}$  and column  $\boxed{TTH}$  is given by

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

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

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

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

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

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.

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.

# EXAMPLE 5: GEOMETRIC DISTRIBUTION

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  $k$ th toss (for  $k = 1, 2, \dots$ )?

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

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

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

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

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

$$\mathbb{P}\{\text{first head on } k\text{th toss}\} = (1 - p)^{k-1} p \quad \text{for } k = 1, 2, \dots .$$

□

# EXAMPLE 6: THREE WAY DUEL

(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:

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

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

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

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

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

In summary:

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

$$\mathbb{P}(\text{A survives} \mid \text{he kills B first}) \approx .23$$

$$\mathbb{P}(\text{A survives} \mid \text{he misses with first shot}) = .3$$

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



## Chapter 2

# Expectations

Just as events have (conditional) probabilities attached to them, with possible interpretation as a long-run frequency, so too do random variables have a number interpretable as a long-run average attached to them. Given a particular piece of information, the symbol

$$\mathbb{E}(X \mid \text{information})$$

denotes the **(conditional) expected value** or **(conditional) expectation** of the random variable  $X$  (given that information). When the information is taken as understood, the expected value is abbreviated to  $\mathbb{E}X$ .

Expected values are not restricted to lie in the range from zero to one.

As with conditional probabilities, there are convenient abbreviations when the conditioning information includes something like {event  $F$  has occurred}:

$$\mathbb{E}(X \mid \text{information and “}F\text{ has occurred”})$$

$$\mathbb{E}(X \mid \text{information, } F)$$

Unlike many authors, I will take the expected value as a primitive concept, not one to be derived from other concepts. All of the methods that those authors use to *define* expected values will be *derived* from a small number of basic rules. You should provide the interpretations for these rules as long-run averages of values generated by independent repetitions of random experiments.

### Rules for (conditional) expectations

Let  $X$  and  $Y$  be random variables,  $c$  and  $d$  be constants, and  $F_1, F_2, \dots$  be events. Then:

(E1)  $\mathbb{E}(cX + dY \mid \text{info}) = c\mathbb{E}(X \mid \text{info}) + d\mathbb{E}(Y \mid \text{info});$

(E2) if  $X$  can only take the constant value  $c$  under the given “info” then  $\mathbb{E}(X \mid \text{info}) = c;$

(E3) if the given “info” forces  $X \leq Y$  then  $\mathbb{E}(X \mid \text{info}) \leq \mathbb{E}(Y \mid \text{info});$

(E4) if the events  $F_1, F_2, \dots$  are disjoint and have union equal to the whole sample space then

$$\mathbb{E}(X \mid \text{info}) = \sum_i \mathbb{E}(X \mid F_i, \text{info}) \mathbb{P}(F_i \mid \text{info}).$$

Only rule E4 should require much work to interpret. It combines the power of both rules P4 and P5 for conditional probabilities. Here is the frequency interpretation for the case of two disjoint events  $F_1$  and  $F_2$  with union  $S$ .

Repeat the experiment a very large number ( $n$ ) of times, noting for each repetition the value taken by  $X$  and which of  $F_1$  or  $F_2$  occurs.

	1	2	3	4	...					$n-1$	$n$	total
$F_1$ occurs	✓	✓		✓	...					✓	✓	$n_1$
$F_2$ occurs			✓		...	✓	✓	✓				$n_2$
$X$	$x_1$	$x_2$	$x_3$	$x_4$	...					$x_{n-1}$	$x_n$	

Those trials where  $F_1$  occurs correspond to conditioning on  $F_1$ :

$$\mathbb{E}(X \mid F_1, \text{info}) \approx \frac{1}{n_1} \sum_{F_1 \text{ occurs}} x_i.$$

Similarly,

$$\mathbb{E}(X \mid F_2, \text{info}) \approx \frac{1}{n_2} \sum_{F_2 \text{ occurs}} x_i$$

and

$$\begin{aligned}\mathbb{P}(F_1 \mid \text{info}) &\approx n_1/n \\ \mathbb{P}(F_2 \mid \text{info}) &\approx n_2/n.\end{aligned}$$

Thus

$$\begin{aligned}&\mathbb{E}(X \mid F_1, \text{info})\mathbb{P}(F_1 \mid \text{info}) + \mathbb{E}(X \mid F_2, \text{info})\mathbb{P}(F_2 \mid \text{info}) \\&\approx \left(\frac{1}{n_1} \sum_{F_1 \text{ occurs}} x_i\right) \left(\frac{n_1}{n}\right) + \left(\frac{1}{n_2} \sum_{F_2 \text{ occurs}} x_i\right) \left(\frac{n_2}{n}\right) \\&= \frac{1}{n} \sum_{i=1}^n x_i \\&\approx \mathbb{E}(X \mid \text{info}).\end{aligned}$$

As  $n$  gets larger and larger all approximations are supposed to get better and better, and so on.

---

**Example 7: Interpretation of expectations as a fair prices for an uncertain returns.**  
(only for those who don't find the frequency interpretation helpful—not essential reading)

---

Rules E2 and E4 imply immediately a result that is used to calculate expectations from probabilities. Consider the case of a random variable  $Y$  expressible as a function  $g(X)$  of another random variable,  $X$ , which takes on only a discrete set of values  $c_1, c_2, \dots$  (I will return later to the case of so-called continuous random variables.) Let  $F_i$  be the subset of  $S$  on which  $X = c_i$ , that is,

$$F_i = \{X = c_i\}.$$

Then by E2,

$$\mathbb{E}(Y \mid F_i, \text{info}) = g(c_i),$$

and by E5,

$$\mathbb{E}(Y \mid \text{info}) = \sum_i g(c_i)\mathbb{P}(F_i \mid \text{info}).$$

More succinctly,

$$(E5) \quad \mathbb{E}(g(X) \mid \text{info}) = \sum_i g(c_i)\mathbb{P}(X = c_i \mid \text{info}).$$

In particular,

$$(E5)' \quad \mathbb{E}(X \mid \text{info}) = \sum_i c_i\mathbb{P}(X = c_i \mid \text{info}).$$

I will refer to these results as new rules for expectations, even though they are consequences of the other rules. They apply to random variables that take values in the “discrete set”  $\{c_1, c_2, \dots\}$ . If the range of values includes an interval of real numbers, an approximation argument (see Chapter 6) replaces sums by integrals.

REMARK. If we extend E1 to sums of more than two random variables, we get a collection of rules that includes the probability rules P1 through P5 as special cases. The derivation makes use of the **indicator function of an event**, defined by

$$\mathbb{I}_A = \begin{cases} 1 & \text{if the event } A \text{ occurs,} \\ 0 & \text{if the event } A^c \text{ occurs.} \end{cases}$$

Rule E4 with  $F_1 = A$  and  $F_2 = A^c$  gives

$$\begin{aligned} \mathbb{E}(\mathbb{I}_A \mid \text{info}) &= \mathbb{E}(\mathbb{I}_A \mid A, \text{info}) \mathbb{P}(A \mid \text{info}) + \mathbb{E}(\mathbb{I}_A \mid A^c, \text{info}) \mathbb{P}(A^c \mid \text{info}) \\ &= 1 \times \mathbb{P}(A \mid \text{info}) + 0 \times \mathbb{P}(A^c \mid \text{info}) \quad \text{by E2.} \end{aligned}$$

That is,  $\mathbb{E}(\mathbb{I}_A \mid \text{info}) = \mathbb{P}(A \mid \text{info})$ .

If an event  $A$  is a disjoint union of events  $A_1, A_2, \dots$  then  $\mathbb{I}_A = \mathbb{I}_{A_1} + \mathbb{I}_{A_2} + \dots$  (Why?) Taking expectations then invoking the extended E1, we get rule P4.

As an exercise, you might try to derive the other probability rules, but don't spend much time on the task or worry about it too much. Just keep buried somewhere in the back of your mind the idea that you can do more with expectations than with probabilities alone.

You will find it useful to remember that  $\mathbb{E}(\mathbb{I}_A \mid \text{info}) = \mathbb{P}(A \mid \text{info})$ , a result that is easy to reconstruct from the fact that the long-run frequency of occurrence of an event, over many repetitions, is just the long-run average of its indicator function.

#### Example 8: Expected number of tosses to get TTHH.

The calculation of an expectation is often a good way to get a rough feel for the behaviour of a random process. It is helpful to remember expectations for a few standard mechanisms, such as coin tossing, rather than have to rederive them repeatedly.

#### Example 9: Expected value for the geometric distribution.

Probabilists study standard mechanisms, and establish basic results for them, partly in the hope that they will recognize those same mechanisms buried in other problems. In that way, unnecessary calculation can be avoided, making it easier to solve more complex problems. It can, however, take some work to find the hidden mechanism.

**Example 10: [Coupon collector problem]** In order to encourage consumers to buy many packets of cereal, a manufacturer includes a Famous Probabilist card in each packet. There are 10 different types of card: Chung, Feller, Lévy, Kolmogorov, ..., Doob. Suppose that I am seized by the desire to own at least one card of each type. What is the expected number of packets that I need to buy in order to achieve my goal?

For the coupon collectors problem I assumed large numbers of cards of each type, in order to justify the analogy with coin tossing. Without that assumption the depletion of cards from the population would have a noticeable effect on the proportions of each type remaining after each purchase. The next example illustrates the effects of sampling from a finite population without replacement, when the population size is not assumed very large.

The example also provides an illustration of the **method of indicators**, whereby a random variable is expressed as a sum of indicator variables  $\mathbb{I}_{A_1} + \mathbb{I}_{A_2} + \dots$ , in order to reduce

calculation of an expected value to separate calculation of probabilities  $\mathbb{P}A_1, \mathbb{P}A_2, \dots$ . Remember the formula

$$\begin{aligned}\mathbb{E}(\mathbb{I}_{A_1} + \mathbb{I}_{A_2} + \dots \mid \text{info}) &= \mathbb{E}(\mathbb{I}_{A_1} \mid \text{info}) + \mathbb{E}(\mathbb{I}_{A_2} \mid \text{info}) + \dots \\ &= \mathbb{P}(A_1 \mid \text{info}) + \mathbb{P}(A_2 \mid \text{info}) + \dots\end{aligned}$$

---

**Example 11:** Suppose an urn contains  $r$  red balls and  $b$  black balls, all balls identical except for color. Suppose balls are removed from the urn one at a time, without replacement. Assume that the person removing the balls selects them at random from the urn: if  $k$  balls remain then each has probability  $1/k$  of being chosen. Question: What is the expected number of red balls removed before the first black ball?

---

The classical gambler's ruin problem was solved by Abraham de Moivre over two hundred years ago, using a method that has grown into one of the main technical tools of modern probability. The solution makes an elegant application of conditional expectations.

---

**Example 12:** Suppose two players, Alf and Betamax, bet on the tosses of a fair coin: for a head, Alf pays Betamax one dollar; for a tail, Betamax pays Alf one dollar. The stop playing when one player runs out of money. If Alf starts with  $\alpha$  dollar bills, and Betamax starts with  $\beta$  dollars bills (both  $\alpha$  and  $\beta$  whole numbers), what is the probability that Alf ends up with all the money?

---

De Moivre's method also works with biased coins, if we count profits in a different way—an even more elegant application of conditional expectations.

---

**Example 13:** Same problem as in Example 12, except that the coin they toss has probability  $p \neq 1/2$  of landing heads. (Could be skipped.)

---

You could safely skip the final Example. It contains a discussion of a tricky little problem, that can be solved by conditioning or by an elegant symmetry argument.

---

**Example 14:** Big pills, little pills. (Tricky. Should be skipped.)

---

### Things to remember

- Expectations (and conditional expectations) are linear (E1), increasing (E3) functions of random variables, which can be calculated as weighted averages of conditional expectations,

$$\mathbb{E}(X \mid \text{info}) = \sum_i \mathbb{E}(X \mid F_i, \text{info}) \mathbb{P}(F_i \mid \text{info}),$$

where the disjoint events  $F_1, F_2, \dots$  cover all possibilities (the weights sum to one).

- The indicator function of an event  $A$  is the random variable defined by

$$\mathbb{I}_A = \begin{cases} 1 & \text{if the event } A \text{ occurs,} \\ 0 & \text{if the event } A^c \text{ occurs.} \end{cases}$$

The expected value of an indicator variable,  $\mathbb{E}(\mathbb{I}_A \mid \text{info})$ , is the same as the probability of the corresponding event,  $\mathbb{P}(A \mid \text{info})$ .

- As a consequence of the rules,

$$\mathbb{E}(g(X) \mid \text{info}) = \sum_i g(c_i) \mathbb{P}(X = c_i \mid \text{info}),$$

if  $X$  can take only values  $c_1, c_2, \dots$

# EXAMPLE 7: FAIR PRICE INTERPRETATION OF EXPECTATIONS

Consider a situation—a bet if you will—where you stand to receive an uncertain return  $X$ . You could think of  $X$  as a random variable, a real-valued function on a sample space  $S$ . For the moment forget about any probabilities on the sample space  $S$ . Suppose you consider  $p(X)$  the fair price to pay in order to receive  $X$ . What properties must  $p(\cdot)$  have?

Your net return will be the random quantity  $X - p(X)$ , which you should consider to be a **fair return**. Unless you start worrying about the utility of money you should find the following properties reasonable.

- (i) **fair + fair = fair**. That is, if you consider  $p(X)$  fair for  $X$  and  $p(Y)$  fair for  $Y$  then you should be prepared to make both bets, paying  $p(X) + p(Y)$  to receive  $X + Y$ .
- (ii) **constant  $\times$  fair = fair**. That is, you shouldn't object if I suggest you pay  $2p(X)$  to receive  $2X$  (actually, that particular example is a special case of (i)) or  $3.76p(X)$  to receive  $3.76X$ , or  $-p(X)$  to receive  $-X$ . The last example corresponds to willingness to take either side of a fair bet. In general, to receive  $cX$  you should pay  $cp(X)$ , for constant  $c$ .
- (iii) There is no fair bet whose return  $X - p(X)$  is always  $\geq 0$  (except for the trivial situation where  $X - p(X)$  is certain to be zero).

If you were to declare a bet with return  $X - p(X) \geq 0$  under all circumstances to be fair, I would be delighted to offer you the opportunity to receive the “fair” return  $-C(X - p(X))$ , for an arbitrarily large positive constant  $C$ . I couldn't lose.

**Fact 1:** *Properties (i), (ii), and (iii) imply that  $p(\alpha X + \beta Y) = \alpha p(X) + \beta p(Y)$  for all random variables  $X$  and  $Y$ , and all constants  $\alpha$  and  $\beta$ .*

Consider the combined effect of the following fair bets:

you pay me  $\alpha p(X)$  to receive  $\alpha X$

you pay me  $\beta p(Y)$  to receive  $\beta Y$

I pay you  $p(\alpha X + \beta Y)$  to receive  $(\alpha X + \beta Y)$ .

Your net return is a constant,

$$c = p(\alpha X + \beta Y) - \alpha p(X) - \beta p(Y).$$

If  $c > 0$  you violate (iii); if  $c < 0$  take the other side of the bet to violate (iii). The asserted equality follows.

**Fact 2:** *Properties (i), (ii), and (iii) imply that  $p(Y) \leq p(X)$  if the random variable  $Y$  is always  $\leq$  the random variable  $X$ .*

If you claim that  $p(X) < p(Y)$  then I would be happy for you to accept the bet that delivers

$$(Y - p(Y)) - (X - p(X)) = -(X - Y) - (p(Y) - p(X)),$$

which is always  $< 0$ .

The two Facts are analogous to rules E1 and E3 for expectations. You should be able to deduce the analog of E2 from (iii).

As a special case, consider the bet that returns 1 if an event  $F$  occurs, and 0 otherwise. If you identify the event  $F$  with the random variable taking the value 1 on  $F$  and 0 on  $F^c$  (that is, the indicator of the event  $F$ ), then it follows directly from Fact 1 that  $p(\cdot)$  is additive:  $p(F_1 \cup F_2) = p(F_1) + p(F_2)$  for disjoint events  $F_1$  and  $F_2$ , an analog of rule P4 for probabilities.

## Contingent bets

Things become much more interesting if you are prepared to make a bet to receive an amount  $X$ , but only when some event  $F$  occurs. That is, the bet is made **contingent** on the occurrence of  $F$ . Typically, knowledge of the occurrence of  $F$  should change the fair price, which we could denote by  $p(X | F)$ . Let me write  $Z$  for the indicator function of the event  $F$ , that is,

$$Z = \begin{cases} 1 & \text{if event } F \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Then the net return from the contingent bet is  $(X - p(X | F)) Z$ . The indicator function  $Z$  ensures that money changes hands only when  $F$  occurs.

By combining various bets and contingent bets, we can deduce that an analog of rule E4 for expectations: if  $S$  is partitioned into disjoint events  $F_1, \dots, F_k$ , then

$$p(X) = \sum_{i=1}^k p(F_i) p(X | F_i).$$

Make the following bets. Write  $c_i$  for  $p(X | F_i)$ .

- (a) For each  $i$ , pay  $c_i p(F_i)$  in order to receive  $c_i$  if  $F_i$  occurs.
- (b) Pay  $-p(X)$  in order to receive  $-X$ .
- (c) For each  $i$ , make a bet contingent on  $F_i$ : pay  $c_i$  (if  $F_i$  occurs) to receive  $X$ .

If event  $F_k$  occurs, your net profit will be

$$-\sum_i c_i p(F_i) + c_k + p(X) - X - c_k + X = p(X) - \sum_i c_i p(F_i),$$

which does not depend on  $k$ . Your profit is always the same constant value. If the constant were nonzero, requirement (iii) for fair bets would be violated.

If you rewrite  $p(X)$  as the expected value  $\mathbb{E}X$ , and  $p(F)$  as  $\mathbb{P}F$  for an event  $F$ , you will see that the properties of fair prices are completely analogous to the rules for probabilities and expectations. Some authors take the bold step of interpreting probability theory as a calculus of fair prices. The interpretation has the virtue that it makes sense in some situations where there is no reasonable way to imagine an unlimited sequence of repetitions from which to calculate a long-run frequency or average.

□

# EXAMPLE 8: EXPECTED VALUES FOR COIN TOSSING GAME

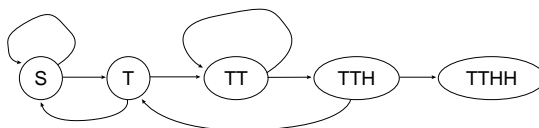
The “HHH versus TTHH” Example in Chapter 1 solved the following problem:

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, and R waits for tthh. The one whose pattern appears first is the winner. What is the probability that M wins?

The answer—that M has probability 5/12 of winning—is slightly surprising, because, at first sight, a pattern of four appears harder to achieve than a pattern of three.

A calculation of expected values will add to the puzzlement. As you will see, if the game is continued until each player sees his pattern, it takes tthh longer (on average) to appear than it takes hhh to appear. However, when the two patterns are competing, the tthh pattern is more likely to appear first. How can that be?

For the moment forget about the competing hhh pattern: calculate the expected number of tosses needed before the pattern tthh is obtained with four successive tosses. That is, if we let  $X$  denote the number of tosses required then the problem asks for the expected value  $\mathbb{E}X$ .



The Markov chain diagram keeps track of the progress from the starting state (labelled S) to the state TTHH where the pattern is achieved. Each arrow in the diagram corresponds to a transition between states with probability 1/2.

Once again it is easier to solve not just the original problem, but a set of problems, one for each starting state. Let

$$\mathcal{E}_S = \mathbb{E}(X \mid \text{start at S})$$

$$\mathcal{E}_T = \mathbb{E}(X \mid \text{start at T})$$

$$\vdots$$

Then the original problem is asking for the value of  $\mathcal{E}_S$ .

Condition on the outcome of the first toss, writing  $\mathcal{H}$  for the event {first toss lands heads} and  $\mathcal{T}$  for the event {first toss lands tails}. From rule E4 for expectations,

$$\mathcal{E}_S = \mathbb{E}(X \mid \text{start at S}, \mathcal{T})\mathbb{P}(\mathcal{T} \mid \text{start at S}) + \mathbb{E}(X \mid \text{start at S}, \mathcal{H})\mathbb{P}(\mathcal{H} \mid \text{start at S})$$

Both the conditional probabilities equal 1/2 (“fair coin”; probability does not depend on the state). For the first of the conditional expectations, count 1 for the first toss, then recognize that the remaining tosses are just those needed to reach TTHH starting from the state  $T$ :

$$\mathbb{E}(X \mid \text{start at S}, \mathcal{T}) = 1 + \mathbb{E}(X \mid \text{start at T})$$

Don’t forget to count the first toss. An analogous argument leads to an analogous expression for the second conditional expectation. Substitution into the expression for  $\mathcal{E}_S$  then gives

$$\mathcal{E}_S = \frac{1}{2}(1 + \mathcal{E}_T) + \frac{1}{2}(1 + \mathcal{E}_S)$$

Similarly,

$$\mathcal{E}_T = \frac{1}{2}(1 + \mathcal{E}_{TT}) + \frac{1}{2}(1 + \mathcal{E}_S)$$

$$\mathcal{E}_{TT} = \frac{1}{2}(1 + \mathcal{E}_{TT}) + \frac{1}{2}(1 + \mathcal{E}_{TTH})$$

$$\mathcal{E}_{TTH} = \frac{1}{2}(1 + 0) + \frac{1}{2}(1 + \mathcal{E}_T)$$

What does the zero in the last equation represent?

The four linear equations in four unknowns have the solution  $\mathcal{E}_S = 16$ ,  $\mathcal{E}_T = 14$ ,  $\mathcal{E}_{TT} = 10$ ,  $\mathcal{E}_{TTH} = 8$ . Thus, the solution to the original problem is that the expected number of tosses to achieve the tthh pattern is 16.

By similar arguments, you can show that the expected number of tosses needed to get hhh, without competition, is 14. The expected number of tosses for the completion of the game with competition between hhh and tthh is  $9\frac{1}{3}$  (see Matlab m-file solve\_hhh\_tthh.m). Notice that the expected value for the game with competition is smaller than the minimum of the expected values for the two games. Why must it be smaller?  $\square$



# EXAMPLE 9: EXPECTED VALUE FOR GEOMETRIC DISTRIBUTION

For independent coin tossing, what is the expected number of tosses to get the first head?

Suppose the coin has probability  $p > 0$  of landing heads. (So we are actually calculating the expected value for the geometric( $p$ ) distribution.) I will present two methods.

## Method A.

Condition on whether the first toss lands heads (H) or tails (T). With  $X$  defined as the number of tosses until the first head,

$$\begin{aligned}\mathbb{E}X &= \mathbb{E}(X | H)\mathbb{P}H + \mathbb{E}(X | T)\mathbb{P}T \\ &= (1)p + (1 + \mathbb{E}X)(1 - p).\end{aligned}$$

The reasoning behind the equality

$$\mathbb{E}(X | T) = 1 + \mathbb{E}X$$

is: After a tail we are back where we started, still counting the number of tosses until a head, except that the first tail must be included in that count.

Solving the equation for  $\mathbb{E}X$  we get

$$\mathbb{E}X = 1/p.$$

Does this answer seem reasonable? (Is it always at least 1? Does it increase as  $p$  increases? What happens as  $p$  tends to zero or one?)

## Method B.

By the formula E5,

$$\mathbb{E}X = \sum_{k=1}^{\infty} k(1-p)^{k-1}p.$$

There are several cute ways to sum this series. Here is my favorite. Write  $q$  for  $1-p$ . Write the  $k$ th summand as a column of  $k$  terms  $pq^{k-1}$ , then sum by rows:

$$\begin{aligned}\mathbb{E}X &= p + pq + pq^2 + pq^3 + \dots \\ &\quad + pq + pq^2 + pq^3 + \dots \\ &\quad \quad + pq^2 + pq^3 + \dots \\ &\quad \quad \quad + pq^3 + \dots \\ &\quad \quad \quad \quad \vdots\end{aligned}$$

Each row is a geometric series.

$$\begin{aligned}\mathbb{E}X &= p/(1-q) + pq/(1-q) + pq^2/(1-q) + \dots \\ &= 1 + q + q^2 + \dots \\ &= 1/(1-q) \\ &= 1/p,\end{aligned}$$

same as before. □

# EXAMPLE 10: COUPON COLLECTOR PROBLEM

In order to encourage consumers to buy many packets of cereal, a manufacturer includes a Famous Probabilist card in each packet. There are 10 different types of card: Chung, Feller, Lévy, Kolmogorov, . . . , Doob. Suppose that I am seized by the desire to own at least one card of each type. What is the expected number of packets that I need to buy in order to achieve my goal?

Assume that the manufacturer has produced enormous numbers of cards, the same number for each type. (If you have ever tried to collect objects of this type, you might doubt the assumption about equal numbers. But, without it, the problem becomes exceedingly difficult.) The assumption ensures, to a good approximation, that the cards in different packets are independent, with probability 1/10 for a Chung, probability 1/10 for a Feller, and so on.

The high points in my life occur at random “times”  $T_1, T_1 + T_2, \dots, T_1 + T_2 + \dots + T_{10}$ , when I add a new type of card to my collection: After one card (that is,  $T_1 = 1$ ) I have my first type; after another  $T_2$  cards I will get something different from the first card; after another  $T_3$  cards I will get a third type; and so on.

The question asks for  $\mathbb{E}(T_1 + T_2 + \dots + T_{10})$ , which rule E1 (applied repeatedly) reexpresses as  $\mathbb{E}T_1 + \mathbb{E}T_2 + \dots + \mathbb{E}T_{10}$ .

The calculation for  $\mathbb{E}T_1$  is trivial because  $T_1$  must equal 1: we get  $\mathbb{E}T_1 = 1$  by rule E2. Consider the mechanism controlling  $T_2$ . For concreteness suppose the first card was a Doob. Each packet after the first is like a coin toss with probability 9/10 of getting a head (= a nonDoob), with  $T_2$  like the number of tosses needed to get the first head. Thus

$T_2$  has a geometric(9/10) distribution.

Deduce from Example 9 that  $\mathbb{E}T_2 = 10/9$ , which is slightly larger than 1.

Now consider the mechanism controlling  $T_3$ . Condition on everything that was observed up to time  $T_1 + T_2$ . Under the assumption of equal abundance and enormous numbers of cards, most of this conditioning information is actually irrelevant; the mechanism controlling  $T_3$  is independent of the past information. (Hard question: Why would the  $T_2$  and  $T_3$  mechanisms not be independent if the cards were not equally abundant?) So what is that  $T_3$  mechanism? I am waiting for any one of the 8 types I have not yet collected. It is like coin tossing with probability 8/10 of heads:

$T_3$  has geometric (8/10) distribution,

and thus  $\mathbb{E}T_3 = 10/8$ . And so on, leading to

$$\mathbb{E}T_1 + \mathbb{E}T_2 + \dots + \mathbb{E}T_{10} = 1 + 10/9 + 10/8 + \dots + 10/1 \approx 29.3.$$

I should expect to buy about 29.3 packets to collect all ten cards. □

Note: The independence between packets was **not** needed to justify the appeal to rule E1, to break the expected value of the sum into a sum of expected values. It did allow us to recognize the various geometric distributions without having to sort through possible effects of large  $T_2$  on the behavior of  $T_3$ , and so on.

You might appreciate better the role of independence if you try to solve a similar (but much harder) problem with just two sorts of card, not in equal proportions.

# EXAMPLE 11: URN EXPERIMENT

Suppose an urn contains  $r$  red balls and  $b$  black balls, all balls identical except for color. Suppose balls are removed from the urn one at a time, without replacement. Assume that the person removing the balls selects them at random from the urn: if  $k$  balls remain then each has probability  $1/k$  of being chosen.

Question: What is the expected number of red balls removed before the first black ball?

The problem might at first appear to require nothing more than a simple application of rule E5' for expectations. We shall see. Let  $T$  be the number of reds removed before the first black. Find the distribution of  $T$ , then appeal to E5' to get

$$\mathbb{E}T = \sum_k k\mathbb{P}\{T = k\}.$$

Sounds easy enough. We have only to calculate the probabilities  $\mathbb{P}\{T = k\}$ .

Define  $R_i = \{i\text{th ball red}\}$  and  $B_i = \{i\text{th ball black}\}$ . The possible values for  $T$  are  $0, 1, \dots, r$ . For  $k$  in this range,

$$\begin{aligned}\mathbb{P}\{T = k\} &= \mathbb{P}\{\text{first } k \text{ balls red, } (k+1)\text{st ball is black}\} \\ &= \mathbb{P}(R_1 R_2 \dots R_k B_{k+1}) \\ &= (\mathbb{P}R_1)\mathbb{P}(R_2 | R_1)\mathbb{P}(R_3 | R_1 R_2) \dots \mathbb{P}(B_{k+1} | R_1 \dots R_k) \\ &= \frac{r}{r+b} \cdot \frac{r-1}{r+b-1} \dots \frac{b}{r+b-k}.\end{aligned}$$

The dependence on  $k$  is fearsome. I wouldn't like to try multiplying by  $k$  and summing. If you are into pain you might try to continue this line of argument. Good luck.

There is a much easier way to calculate the expectation, by breaking  $T$  into a sum of much simpler random variables for which E5' is trivial to apply. This approach is sometimes called the **method of indicators**.

Suppose the red balls are labelled  $1, \dots, r$ . Let  $T_i$  equal 1 if red ball number  $i$  is sampled before the first black ball. (Be careful here. The black balls are not thought of as numbered. The first black ball is not a ball bearing the number 1; it might be any of the  $b$  black balls in the urn.) Then  $T = T_1 + \dots + T_r$ . By symmetry—it is assumed that the numbers have no influence on the order in which red balls are selected—each  $T_i$  has the same expectation. Thus

$$\mathbb{E}T = \mathbb{E}T_1 + \dots + \mathbb{E}T_r = r\mathbb{E}T_1.$$

For the calculation of  $\mathbb{E}T_1$  we can ignore most of the red balls. The event  $\{T_1 = 1\}$  occurs if and only if red ball number 1 is drawn before all  $b$  of the black balls. By symmetry, the event has probability  $1/(b+1)$ . (If  $b+1$  objects are arranged in random order, each object has probability  $1/(1+b)$  of appearing first in the order.)

REMARK. If you are not convinced by the appeal to symmetry, you might find it helpful to consider a thought experiment where all  $r+b$  balls are numbered and they are removed at random from the urn. That is, treat all the balls as distinguishable and sample until the urn is empty. (You might find it easier to follow the argument in a particular case, such as all  $120 = 5!$  orderings for five distinguishable balls, 2 red and 3 black.) The sample space consists of all permutations of the numbers  $1$  to  $r+b$ . Each permutation is equally likely. For each permutation in which red 1 precedes all the black balls there is another equally likely permutation, obtained by interchanging the red ball with the first of the black balls chosen; and there is an equally likely permutation in which it appears after two black balls, obtained by interchanging the red ball with the second of the black balls chosen; and so on. Formally, we are partitioning the whole sample space into equally likely events, each determined by a relative ordering of red 1 and all the black balls. There are  $b+1$  such equally likely events, and their probabilities sum to one.

Now it is easy to calculate the expected value for red 1.

$$\mathbb{E}T_1 = 0 \mathbb{P}\{T_1 = 0\} + 1 \mathbb{P}\{T_1 = 1\} = 1/(b + 1)$$

The expected number of red balls removed before the first black ball is equal to  $r/(b + 1)$ .  
 $\square$

Compare the solution  $r/(b + 1)$  with the result for sampling with replacement, where the number of draws required to get the first black would have a  $\text{geometric}(b/(r + b))$  distribution. With replacement, the expected number of reds removed before the first black would be

$$(b/(r + b))^{-1} - 1 = r/b.$$

Replacement of balls after each draw increases the expected value slightly. Does that make sense?

## EXAMPLE 12: GAMBLER'S RUIN: FAIR COIN

Suppose two players, Alf (A for short) and Betamax (B for short), bet on the tosses of a fair coin: for a head, Alf pays Betamax one dollar; for a tail, Betamax pays Alf one dollar. They stop playing when one player runs out of money. If Alf starts with  $\alpha$  dollar bills, and Betamax starts with  $\beta$  dollars bills (both  $\alpha$  and  $\beta$  whole numbers), what is the probability that Alf ends up with all the money?

Write  $X_n$  for the number of dollars held by A after  $n$  tosses. (Of course, once the game ends the value of  $X_n$  stays fixed from then on, at either  $\alpha + \beta$  or 0, depending on whether A won or not.) It is a random variable taking values in the range  $\{0, 1, 2, \dots, \alpha + \beta\}$ . We start with  $X_0 = \alpha$ . To solve the problem, calculate  $\mathbb{E}X_n$ , for very large  $n$  in two ways, then equate the answers. We need to solve for the unknown  $\theta = \mathbb{P}\{A \text{ wins}\}$ .

### First calculation

Invoke rule E4 with the sample space broken into three pieces,

$$A_n = \{\text{A wins at, or before, the } n\text{th toss}\},$$

$$B_n = \{\text{B wins at, or before, the } n\text{th toss}\},$$

$$C_n = \{\text{game still going after the } n\text{th toss}\}.$$

For very large  $n$  the game is almost sure to be finished, with  $\mathbb{P}A_n \approx \theta$ ,  $\mathbb{P}B_n \approx 1 - \theta$ , and  $\mathbb{P}C_n \approx 0$ . Thus

$$\begin{aligned}\mathbb{E}X_n &= \mathbb{E}(X_n | A_n)\mathbb{P}A_n + \mathbb{E}(X_n | B_n)\mathbb{P}B_n + \mathbb{E}(X_n | C_n)\mathbb{P}C_n \\ &\approx ((\alpha + \beta) \times \theta) + (0 \times (1 - \theta)) + (\text{something} \times 0).\end{aligned}$$

The error in the approximation goes to zero as  $n$  goes to infinity.

### Second calculation

Calculate conditionally on the value of  $X_{n-1}$ . That is, split the sample space into disjoint events  $F_k = \{X_{n-1} = k\}$ , for  $k = 0, 1, \dots, \alpha + \beta$ , then works towards another appeal to rule E4. For  $k = 0$  or  $k = \alpha + \beta$ , the game will be over, and  $X_n$  must take the same value as  $X_{n-1}$ . That is,

$$\mathbb{E}(X_n | F_0) = 0 \quad \text{and} \quad \mathbb{E}(X_n | F_{\alpha+\beta}) = \alpha + \beta.$$

For values of  $k$  between the extremes, the game is still in progress. With the next toss, A's fortune will either increase by one dollar (with probability 1/2) or decrease by one dollar (with probability 1/2). That is, for  $k = 1, 2, \dots, \alpha + \beta - 1$ ,

$$\mathbb{E}(X_n | F_k) = \frac{1}{2}(k + 1) + \frac{1}{2}(k - 1) = k.$$

Now invoke E4.

$$E(X_n) = 0 \times \mathbb{P}F_0 + 1 \times \mathbb{P}F_1 + \dots + (\alpha + \beta) \times \mathbb{P}F_{\alpha+\beta}.$$

Compare with the direct application of E5' to the calculation of  $\mathbb{E}X_{n-1}$ :

$\mathbb{E}(X_{n-1}) = (0 \times \mathbb{P}\{X_{n-1} = 0\}) + (1 \times \mathbb{P}\{X_{n-1} = 1\}) + \dots + ((\alpha + \beta) \times \mathbb{P}\{X_{n-1} = \alpha + \beta\})$ , which is just another way of writing the sum for  $\mathbb{E}X_n$  derived above. Thus we have

$$\mathbb{E}X_n = \mathbb{E}X_{n-1}$$

The expected value doesn't change from one toss to the next.

Follow this fact back through all the previous tosses to get

$$\mathbb{E}X_n = \mathbb{E}X_{n-1} = \mathbb{E}X_{n-2} = \dots = \mathbb{E}X_2 = \mathbb{E}X_1 = \mathbb{E}X_0.$$

But  $X_0$  is equal to  $\alpha$ , for certain, which forces  $\mathbb{E}X_0 = \alpha$ .

### Putting the two answers together

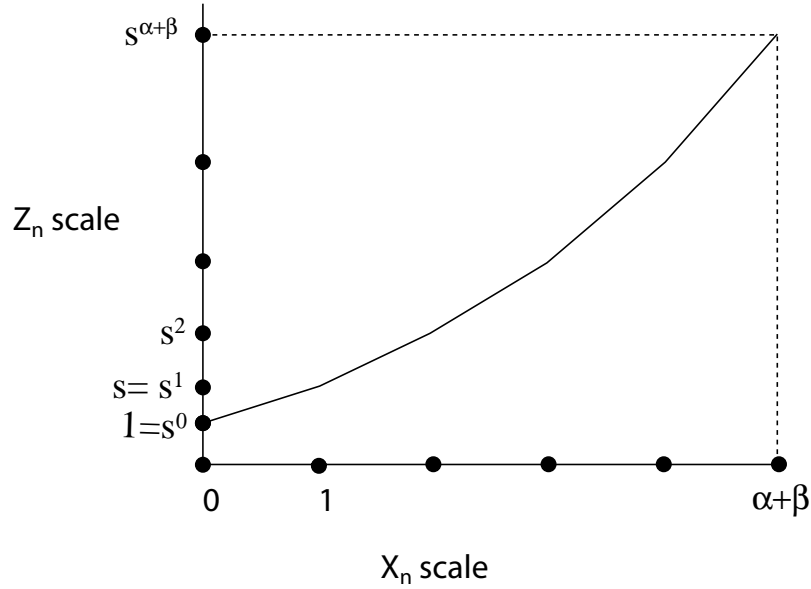
We have two results:  $\mathbb{E}X_n = \alpha$ , no matter how large  $n$  is; and  $\mathbb{E}X_n$  gets arbitrarily close to  $\theta(\alpha + \beta)$  as  $n$  gets larger. We must have  $\alpha = \theta(\alpha + \beta)$ . That is, Alf has probability  $\alpha/(\alpha + \beta)$  of eventually winning all the money.  $\square$

REMARK. Twice I referred to the sample space, without actually having to describe it explicitly. It mattered only that several conditional probabilities were determined by the wording of the problem.

EXAMPLE 13: GAMBLER'S RUIN: BIASED COIN

Same problem as in Example 12, except that the coin they toss has probability  $p \neq 1/2$  of landing heads.

The cases  $p = 0$  and  $p = 1$  are trivial. So let us assume that  $0 < p < 1$ . Essentially De Moivre's idea was that we could use almost the same method as in Example 12 if we kept track of A's fortune on a geometrically expanding scaled. For some number  $s$ , to be specified soon, consider a new random variable  $Z_n = s^{X_n}$ .



Once again write  $\theta$  for  $\mathbb{P}\{A \text{ wins}\}$ , and give the events  $A_n$ ,  $B_n$ , and  $C_n$  the same meaning as in Example 12.

As in the first calculation for the other Example, we have

$$\begin{aligned}\mathbb{E}Z_n &= \mathbb{E}(s^{X_n} | A_n)\mathbb{P}A_n + \mathbb{E}(s^{X_n} | B_n)\mathbb{P}B_n + \mathbb{E}(s^{X_n} | C_n)\mathbb{P}C_n \\ &\approx (s^{\alpha+\beta} \times \theta) + (s^0 \times (1 - \theta)) + (\text{something} \times 0)\end{aligned}$$

if  $n$  is very large.

For the analog of the second calculation, in the cases where the game has ended by at or before the  $(n - 1)$ st toss we have

$$\mathbb{E}(Z_n | X_{n-1} = 0) = s^0 \quad \text{and} \quad \mathbb{E}(Z_n | X_{n-1} = \alpha + \beta) = s^{\alpha+\beta}.$$

For  $0 < k < \alpha + \beta$ , the result of the calculation is slightly different.

$$\mathbb{E}(Z_n | X_{n-1} = k) = ps^{k+1} + (1 - p)s^{k-1} = (ps + (1 - p)s^{-1})s^k.$$

If we choose  $s = (1 - p)/p$ , the factor  $(ps + (1 - p)s^{-1})$  becomes 1. Invoking rule E4 we then get

$$\begin{aligned}\mathbb{E}Z_n &= \mathbb{E}(Z_n | X_{n-1} = 0) \times \mathbb{P}(X_{n-1} = 0) + \mathbb{E}(Z_n | X_{n-1} = 1) \times \mathbb{P}(X_{n-1} = 1) \\ &\quad + \dots + \mathbb{E}(Z_n | X_{n-1} = \alpha + \beta) \times \mathbb{P}(X_{n-1} = \alpha + \beta) \\ &= s^0 \times \mathbb{P}(X_{n-1} = 0) + s^1 \times \mathbb{P}(X_{n-1} = 1) + \dots + s^{\alpha+\beta} \times \mathbb{P}(X_{n-1} = \alpha + \beta)\end{aligned}$$

Compare with the calculation of  $\mathbb{E}Z_{n-1}$  via E5.

$$\begin{aligned}\mathbb{E}Z_{n-1} &= \mathbb{E}(s^{X_{n-1}} | X_{n-1} = 0) \times \mathbb{P}(X_{n-1} = 0) + \mathbb{E}(s^{X_{n-1}} | X_{n-1} = 1) \times \mathbb{P}(X_{n-1} = 1) \\ &\quad + \dots + \mathbb{E}(s^{X_{n-1}} | X_{n-1} = \alpha + \beta) \times \mathbb{P}(X_{n-1} = \alpha + \beta) \\ &= s^0 \times \mathbb{P}(X_{n-1} = 0) + s^1 \times \mathbb{P}(X_{n-1} = 1) + \dots + s^{\alpha+\beta} \times \mathbb{P}(X_{n-1} = \alpha + \beta)\end{aligned}$$

Once again we have a situation where  $\mathbb{E}Z_n$  stays fixed at the initial value  $\mathbb{E}Z_0 = s^\alpha$ , but, with very large  $n$ , it can be made arbitrarily close to  $\theta s^{\alpha+\beta} + (1-\theta)s^0$ . Equating the two values, we deduce that

$$\mathbb{P}\{\text{Alf wins}\} = \theta = \frac{1 - s^\alpha}{1 - s^{\alpha+\beta}} \quad \text{where } s = (1-p)/p.$$

□



# EXAMPLE 14: BIG PILLS, LITTLE PILLS

My interest in the calculations in Example 11 was kindled by a problem that appeared in the August-September 1992 issue of the American Mathematical Monthly. My solution to the problem—the one I first came up with by application of a straightforward conditioning argument—reduces the calculation to several applications of the result from the previous Example. The solution offered by two readers of the Monthly was slicker.

**E 3429** [1991, 264]. *Proposed by Donald E. Knuth and John McCarthy, Stanford University, Stanford, CA.*

A certain pill bottle contains  $m$  large pills and  $n$  small pills initially, where each large pill is equivalent to two small ones. Each day the patient chooses a pill at random; if a small pill is selected, (s)he eats it; otherwise (s)he breaks the selected pill and eats one half, replacing the other half, which thenceforth is considered to be a small pill.

- (a) What is the expected number of small pills remaining when the last large pill is selected?
- (b) On which day can we expect the last large pill to be selected?

SOLUTION FROM AMM:

*Composite solution by Walter Stromquist, Daniel H. Wagner, Associates, Paoli, PA and Tim Hesterberg, Franklin & Marshall College, Lancaster, PA.* The answers are (a)  $n/(m+1) + \sum_{k=1}^m (1/k)$ , and (b)  $2m + n - (n/(m+1)) - \sum_{k=1}^m (1/k)$ . The answer to (a) assumes that the small pill created by breaking the last large pill is to be counted. A small pill present initially remains when the last large pill is selected if and only if it is chosen last from among the  $m+1$  element set consisting of itself and the large pills—an event of probability  $1/(m+1)$ . Thus the expected number of survivors from the original small pills is  $n/(m+1)$ . Similarly, when the  $k$ th large pill is selected ( $k = 1, 2, \dots, m$ ), the resulting small pill will outlast the remaining large pills with probability  $1/(m-k+1)$ , so the expected number of created small pills remaining at the end is  $\sum_{k=1}^m (1/k)$ . Hence the answer to (a) is as above. The bottle will last  $2m+n$  days, so the answer to (b) is just  $2m+n$  minus the answer to (a), as above.

I offer two alternative methods of solution for the problem. The first method uses a conditioning argument to set up a recurrence formula for the expected numbers of small pills remaining in the bottle after each return of half a big pill. The equations are easy to solve by repeated substitution. The second method uses indicator functions to spell out the Hesterberg-Stromquist method in more detail. Apparently the slicker method was not as obvious to most readers of the Monthly (and me):

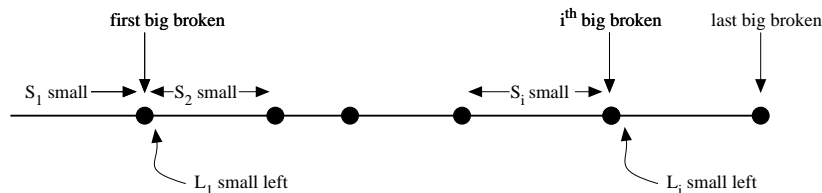
*Editorial comment.* Most solvers derived a recurrence relation, guessed the answer, and verified it by induction. Several commented on the origins of the problem. Robert High saw a version of it in the MIT Technology Review of April, 1990. Helmut Prodinger reports that he proposed it in the Canary Islands in 1982. Daniel Moran attributes the problem to Charles MacCluer of Michigan State University, where it has been known for some time.

Solved by 38 readers (including those cited) and the proposer. One incorrect solution was received.

## Conditioning method.

Invent random variables to describe the depletion of the pills. Initially there are  $L_0 = n$  small pills in the bottle. Let  $S_1$  small pills be consumed before the first large pill is broken. After the small half is returned to the bottle let there be  $L_1$  small pills left. Then let

$S_2$  small pills be consumed before the next big pill is split, leaving  $L_2$  small pills in the bottle. And so on.



With this notation, part (a) is asking for  $\mathbb{E}L_m$ . Part (b) is asking for  $2m + n - \mathbb{E}L_m$ : If the last big pill is selected on day  $X$  then it takes  $X + L_m$  days to consume the  $2m + n$  small pill equivalents, so  $\mathbb{E}X + \mathbb{E}L_m = 2m + n$ .

The random variables are connected by the equation

$$L_i = L_{i-1} - S_i + 1,$$

the  $-S_i$  representing the small pills consumed between the breaking of the  $(i - 1)$ st and  $i$ th big pill, and the  $+1$  representing the half of the big pill that is returned to the bottle. Taking expectations we get

$$\mathbb{E}L_i = \mathbb{E}L_{i-1} - \mathbb{E}S_i + 1.$$

The result from Example 11 will let us calculate  $\mathbb{E}S_i$  in terms of  $\mathbb{E}L_{i-1}$ , thereby producing the recurrence formula for  $\mathbb{E}L_i$ .

Condition on the pill history up to the  $(i - 1)$ st breaking of big pill (and the return of the unconsumed half to the bottle). At that point there are  $L_{i-1}$  small pills and  $m - (i - 1)$  big pills in the bottle. The mechanism controlling  $S_i$  is just like the urn problem of Example 11, with

$$\begin{aligned} r &= L_{i-1} \text{ red balls (= small pills)} \\ b &= m - (i - 1) \text{ black balls (= big pills).} \end{aligned}$$

From that Example,

$$\mathbb{E}(S_i \mid \text{history to } (i - 1)\text{st breaking of a big pill}) = \frac{L_{i-1}}{1 + m - (i - 1)}.$$

To calculate  $\mathbb{E}S_i$  we would need to average out using weights equal to the probability of each particular history:

$$\mathbb{E}S_i = \frac{1}{1 + m - (i - 1)} \sum_{\text{histories}} \mathbb{P}\{\text{history}\}(\text{value of } L_{i-1} \text{ for that history}).$$

The sum on the right-hand side is exactly the sum we would get if we calculated  $\mathbb{E}L_{i-1}$  using rule E4, partitioning the sample space according to possible histories up to the  $(i - 1)$ st breaking of a big pill. Thus

$$\mathbb{E}S_i = \frac{1}{2 + m - i} \mathbb{E}L_{i-1}.$$

Now we can eliminate  $\mathbb{E}S_i$  from equality 15 to get the recurrence formula for the  $\mathbb{E}L_i$  values:

$$\mathbb{E}L_i = \left(1 - \frac{1}{2 + m - i}\right) \mathbb{E}L_{i-1} + 1.$$

If we define  $\theta_i = \mathbb{E}L_i / (1 + m - i)$  the equation becomes

$$\theta_i = \theta_{i-1} + \frac{1}{1 + m - i} \quad \text{for } i = 1, 2, \dots, m,$$

with initial condition  $\theta_0 = \mathbb{E}L_0/(1+m) = n/(1+m)$ . Repeated substitution gives

$$\begin{aligned}\theta_1 &= \theta_0 + \frac{1}{m} \\ \theta_2 &= \theta_1 + \frac{1}{m-1} = \theta_0 + \frac{1}{m} + \frac{1}{m-1} \\ \theta_3 &= \theta_2 + \frac{1}{m-2} = \theta_0 + \frac{1}{m} + \frac{1}{m-1} + \frac{1}{m-2} \\ &\vdots \\ \theta_m &= \dots = \theta_0 + \frac{1}{m} + \frac{1}{m-1} + \dots + \frac{1}{2} + \frac{1}{1}.\end{aligned}$$

That is, the expected number of small pills left after the last big pill is broken equals

$$\begin{aligned}\mathbb{E}L_m &= (1+m-m)\theta_m \\ &= \frac{n}{1+m} + 1 + \frac{1}{2} + \dots + \frac{1}{m}.\end{aligned}$$

### Rewrite of the Stromquist-Hesterberg solution.

Think in terms of half pills, some originally part of big pills. Number the original half pills  $1, \dots, n$ . Define

$$H_i = \begin{cases} +1 & \text{if original half pill } i \text{ survives beyond last big pill} \\ 0 & \text{otherwise.} \end{cases}$$

Number the big pills  $1, \dots, m$ . Use the same numbers to refer to the half pills that are created when a big pill is broken. Define

$$B_j = \begin{cases} +1 & \text{if created half pill } j \text{ survives beyond last big pill} \\ 0 & \text{otherwise.} \end{cases}$$

The number of small pills surviving beyond the last big pill equals

$$H_1 + \dots + H_n + B_1 + \dots + B_m.$$

By symmetry, each  $H_i$  has the same expected value, as does each  $B_j$ . The expected value asked for by part (a) equals

$$<.15> \quad n\mathbb{E}H_1 + m\mathbb{E}B_1 = n\mathbb{P}\{H_1 = 1\} + m\mathbb{P}\{B_1 = 1\}.$$

For the calculation of  $\mathbb{P}\{H_1 = +1\}$  we can ignore all except the relative ordering of the  $m$  big pills and the half pill described by  $H_1$ . By symmetry, the half pill has probability  $1/(m+1)$  of appearing in each of the  $m+1$  possible positions in the relative ordering. In particular,

$$\mathbb{P}\{H_1 = +1\} = \frac{1}{m+1}.$$

For the created half pills the argument is slightly more complicated. If we are given that big pill number 1 the  $k$ th amongst the big pills to be broken, the created half then has to survive beyond the remaining  $m-k$  big pills. Arguing again by symmetry amongst the  $(m-k+1)$  orderings we get

$$\mathbb{P}(B_1 = +1 \mid \text{big number 1 chosen as } k\text{th big}) = \frac{1}{m-k+1}.$$

Also by symmetry,

$$\mathbb{P}\{\text{big 1 chosen as } k\text{th big}\} = \frac{1}{m}.$$

Average out using the conditioning rule E4 to deduce

$$\mathbb{P}\{B_1 = +1\} = \frac{1}{m} \sum_{k=1}^m \frac{1}{m-k+1}.$$

Notice that the summands run through the values  $1/1$  to  $1/m$  in reversed order.

When the values for  $\mathbb{P}\{H_1 = +1\}$  and  $\mathbb{P}\{B_1 = +1\}$  are substituted into 16, the asserted answer to part (a) results.  $\square$

## 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 name comes from the **binomial coefficient**,  $\binom{n}{k}$ , which is defined as the number of subsets of size  $k$  for a set of size  $n$ . (Read the symbol as “ $n$  choose  $k$ ”.) By convention,  $\binom{n}{0} = 1$ .

There is a quick probabilistic way to determine  $\binom{n}{k}$ , for integers  $1 \leq k \leq n$ . Suppose  $k$  balls are sampled at random, without replacement, from an urn containing  $k$  red balls and  $n - k$  black balls. Each of the  $\binom{n}{k}$  different subsets of size  $k$  has probability  $1/\binom{n}{k}$  of being selected. In particular, there is probability  $1/\binom{n}{k}$  that the sample consists of the red balls. We can also calculate that probability, via a conditioning argument, to be

$$\frac{k}{n} \cdot \frac{k-1}{n-1} \cdot \frac{k-2}{n-2} \cdots \frac{1}{n-k+1} :$$

given that the first  $i$  balls are red, the probability that the  $(i+1)$ st is red is  $(k-i)/(n-i)$ . Equating the two values for  $\mathbb{P}\{\text{sample consists of all red balls}\}$ , we get

$$\binom{n}{k} = \frac{n(n-1) \cdots (n-k+1)}{k!} = \frac{n!}{k!(n-k)!}$$

The formula also holds for  $k = 0$  if we interpret  $0!$  as 1.

REMARK. 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  $\alpha$  define

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

Then

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

**Definition.** A random variable is said to have a  $\text{Bin}(n, p)$  distribution, for a parameter  $p$  in the range  $0 \leq p \leq 1$ , if it can take values  $0, 1, \dots, n-1, n$  with probabilities

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

Compare with the binomial expansion,

$$1 = (p + q)^n = \sum_{k=0}^n \binom{n}{k} p^k q^{n-k} \quad \text{where } q = 1 - p.$$

---

**Example 15:** For  $n$  independent tosses of a coin that lands heads with probability  $p$ , show that the total number of heads has a  $\text{Bin}(n, p)$  distribution.

---

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. The distribution appears often in probabilistic modelling; it is worthwhile recording its properties.

---

**Example 16:** Show that the  $\text{Bin}(n, p)$  distribution has expected value  $np$ .

---

The Binomial distribution is often buried within complicated problems.

---

**Example 17:** 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  $1/2$ , or north, with probability  $1/2$ . What is the probability that he is mugged outside the railroad station?

---

The following problem is an example of **Bayesian inference**, based on the probabilistic result known as **Bayes's rule**. You need not memorize the rule, because it is just an application of the chopping/peeling method you already know.

---

**Example 18:** 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 questions correct. What is the probability that he will also get the 7th question correct?

---

To retain a neutral position, I should also give an example of a different approach to statistical inference. The example just happens to involve the Binomial distribution again.

---

**Example 19:** 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?

---

We will meet the Binomial again.

# EXAMPLE 15: BINOMIALS FROM COIN TOSSING

*For  $n$  independent tosses of a coin that lands heads with probability  $p$ , show that the total number of heads has a  $\text{Bin}(n, p)$  distribution.*

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

$$F_1 = \{\text{first } k \text{ gives heads, next } n-k \text{ give tails}\}$$

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

$\vdots$

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  $\{j\text{th toss is a head}\}$ . Then

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

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. 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

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

which is the asserted Binomial distribution. □

EXAMPLE 16: EXPECTED VALUE OF THE BINOMIAL DISTRIBUTION

*Show that the  $\text{Bin}(n, p)$  distribution has expected value  $np$ .*

**Hard way:** By rule E5' 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

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

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 E5',

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

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

The calculation made no use of the independence. If each  $X_i$  has **marginal** distribution  $\text{Bin}(1, p)$ , that is, if

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

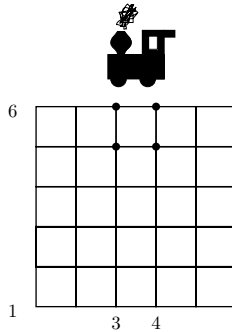
then  $\mathbb{E}(X_1 + \dots + X_n) = np$ , regardless of possible dependence between the tosses. The expectation of a sum is the sum of the expectations, no matter how dependent the summands might be.



# EXAMPLE 17: WHERE TO GET MUGGED

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  $1/2$ , or 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.

$$\begin{aligned} \mathbb{P}\{\text{mugged at railroad}\} &= \mathbb{P}\{\text{reach (3,5), move north}\} + \mathbb{P}\{\text{reach (4,5), move north}\} \\ &= \frac{1}{2}\mathbb{P}\{\text{reach (3,5)}\} + \frac{1}{2}\mathbb{P}\{\text{reach (4,5)}\} \\ &= \frac{1}{2}\mathbb{P}\{\text{move east twice during first 6 blocks}\} \\ &\quad + \frac{1}{2}\mathbb{P}\{\text{move east 3 times during first 7 blocks}\}. \end{aligned}$$

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  $\text{Bin}(7, 1/2)$  lurks behind the calculation. The other calculation involves the  $\text{Bin}(6, 1/2)$ .

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

□

Notice that the events  $\{\text{reach (3,5)}\}$  and  $\{\text{reach (4,5)}\}$  are not disjoint. We need to include the part about moving north to get a clean break.

# EXAMPLE 18: BAYESIAN 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. Let  $C$  denote the event {question 7 answered correctly},  $G$  denote the event {the candidate is a guesser}, and  $S$  denote the event {the candidate is skilled}. Then

- (i) for a guesser,  $X$  has (conditional) distribution  $\text{Bin}(6, 1/3)$
- (ii) for a skilled candidate,  $X$  has (conditional) distribution  $\text{Bin}(6, 1/2)$ .
- (iii)  $\mathbb{P}G = 0.3$  and  $\mathbb{P}S = 0.7$ .

The question asks for  $\mathbb{P}(C \mid X = 4)$ .

Split according to the type of candidate, then condition.

$$\begin{aligned}\mathbb{P}(C \mid X = 4) &= \mathbb{P}\{CS \mid X = 4\} + \mathbb{P}\{CG \mid X = 4\} \\ &= \mathbb{P}(S \mid X = 4)\mathbb{P}(C \mid X = 4, S) + \mathbb{P}(G \mid X = 4)\mathbb{P}(C \mid X = 4, G).\end{aligned}$$

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

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

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:

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

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

Apply the usual splitting/conditioning argument.

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

REMARK. The preceding calculation is an instance of **Bayes's rule**.

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

$$\frac{1}{2}(.869) + \frac{1}{3}(.131) \approx .478$$

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 odds ratio,  $\mathbb{P}S/\mathbb{P}G$ , is multiplied by a factor that reflects the relative support of the data for the two competing explanations “skilled” and “guessing”.

The Example is an instance of **Bayesian inference**, a method of statistical inference followed devoutly by some Statisticians, and derided by others. There is no disagreement regarding the validity of Bayes’s rule; it is the assignment of prior probabilities—such as the  $\mathbb{P}S$  and  $\mathbb{P}G$  of the problem—that is controversial in a general setting.

# EXAMPLE 19: BAD CHECKS

*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 size of the legislature, and if the reporter samples at random from its members, the number of guilty in the sample should be distributed  $\text{Bin}(20, p)$ . 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  $\text{Bin}(20, p)$  distribution. Before the sample is taken we could assert

$$\begin{aligned}\mathbb{P}_p\{X \geq 4\} &= \binom{20}{4}p^4(1-p)^{16} + \binom{20}{5}p^5(1-p)^{15} + \dots + \binom{20}{20}p^{20}(1-p)^0 \\ &= 1 - \left( \binom{20}{0}p^0(1-p)^{20} + \binom{20}{1}p^1(1-p)^{19} + \binom{20}{2}p^2(1-p)^{18} + \binom{20}{3}p^3(1-p)^{17} \right).\end{aligned}$$

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.

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 (the so-called null hypothesis), but more likely under an alternative model. If the event occurs, it casts doubt on the validity of the null hypothesis. For the present example the event  $\{X \geq 4\}$  would have been much more likely under alternative explanations involving larger proportions of bad-check writers amongst the members of the legislature.

## Chapter 4

---

# Symmetry

You should always look for symmetry properties before slogging your way through calculations with what might seem the obvious method. Symmetry, like a fairy godmother, can turn up in unexpected places.

---

Example 20: Suppose an urn initially contains  $r$  red balls and  $b$  black balls. Suppose balls are sampled from the urn one at a time, but after each draw  $k + 1$  balls of the same color are returned to the urn (with thorough mixing between draws, blindfolds, and so on).

- (a) What is the distribution of the number of red balls in the first  $n$  draws?
- (b) What is the probability that the  $i$ th ball drawn is red?
- (c) What is the expected number of red balls in the first  $n$  draws?

---

The return of multiple balls to the urn gives a crude model for contagion, whereby the occurrence of an event (such as selection of a red ball) makes the future occurrence of similar events more likely. The model is known as the **Polya urn** scheme.

The next Example illustrates a slightly different type of argument, where the symmetry enters conditionally.

---

Example 21: A pack of cards consists of 26 reds and 26 blacks. I shuffle the cards, then deal them out one at a time, face up. You are given the chance to win a big prize by correctly predicting when the next card to be dealt will be red. You are allowed to make the prediction for only one card, and you must predict red, not black. What strategy should you adopt to maximize your probability of winning the prize?

---

For the Bet Red problem, one obvious strategy is: wait until there are more red cards than black cards left in the deck. It might seem that such a strategy must ensure a probability greater than  $1/2$  of correctly predicting a red card. The flaw in the method lies in the possibility that we might wait until it is too late; it might happen that the number of red cards revealed is always  $\geq$  the number of black cards revealed. With a deck of 52 cards, the probability might seem to be so small that it can be ignored—but it is not. The calculation of the probability is closely related to the so-called ballot problem, described in the next Example.

---

Example 22: (Can be skipped.) Suppose an urn contains  $r$  red balls and  $b$  black balls, with  $r > b$ . As balls are sampled without replacement from the urn, keep track of the total number of red balls removed and the total number of black balls removed after each draw. Show that the probability of the event {number of reds removed is strictly greater than the number of blacks removed, after every draw} is equal to  $(r - b)/(r + b)$ .

---

# EXAMPLE 20: POLYA URN MODEL

Suppose an urn initially contains  $r$  red balls and  $b$  black balls. Suppose balls are sampled from the urn one at a time, but after each draw  $k + 1$  balls of the same color are returned to the urn (with thorough mixing between draws, blindfolds, and so on).

QUESTIONS (FOR GENERAL  $k$ ):

- (a) What is the distribution of the number of red balls in the first  $n$  draws?
- (b) What is the probability that the  $i$ th ball drawn is red?
- (c) What is the expected number of red balls in the first  $n$  draws?

REMARK. If  $k = 0$ , the procedure is just sampling with replacement. The number of red balls in the first  $n$  draws would then have a  $\text{Bin}(n, r/(r+b))$  distribution. If  $k = -1$ , the procedure is sampling without replacement. If  $k \geq 1$  we will need a very big urn if we intend to sample for a long time: there will be  $r + b + ki$  balls in the urn after the  $i$ th draw and replacement.

To answer these questions we do not need to keep track of exactly which ball is selected at each draw; only its color matters. The questions involve only the events

$$R_i = \{\text{\textit{i}th ball drawn from urn is red}\}$$

and their complements  $B_i$ , for  $i = 1, 2, \dots$ . Clearly  $\mathbb{P}R_1 = r/(r + b)$ .

To get a feel for what is going on, start with some simple calculations for the first few draws, conditioning on the outcomes of the preceeding draws.

$$\begin{aligned}\mathbb{P}R_2 &= \mathbb{P}R_1R_2 + \mathbb{P}B_1R_2 \\ &= \mathbb{P}R_1\mathbb{P}(R_2 \mid R_1) + \mathbb{P}B_1\mathbb{P}(R_2 \mid B_1) \\ &= \left(\frac{r}{r+b} \times \frac{r+k}{r+b+k}\right) + \left(\frac{b}{r+b} \times \frac{r}{r+b+k}\right) \\ &= \frac{r(r+k) + rb}{(r+b)(r+k+b)} \\ &= \frac{r}{r+b}.\end{aligned}$$

Slightly harder:

$$\begin{aligned}\mathbb{P}R_3 &= \mathbb{P}(R_1R_2R_3) + \mathbb{P}(R_1B_2R_3) + \mathbb{P}(B_1R_2R_3) + \mathbb{P}(B_1B_2R_3) \\ &= \frac{r}{r+b} \times \frac{r+k}{r+b+k} \times \frac{r+2k}{r+b+2k} \\ &\quad + \frac{r}{r+b} \times \frac{b}{r+b+k} \times \frac{r+k}{r+b+2k} \\ &\quad + \frac{b}{r+b} \times \frac{r}{r+b+k} \times \frac{r+k}{r+b+2k} \\ &\quad + \frac{b}{r+b} \times \frac{b+k}{r+b+k} \times \frac{r}{r+b+2k}\end{aligned}$$

Each summand has the same denominator:

$$(r+b)(r+b+k)(r+b+2k),$$

corresponding to the fact that the number of balls in the urn increases by  $k$  after each draw.

The sum of the numerators rearranges to

$$\begin{aligned}&(r(r+k)(r+2k) + r(r+k)b) + (rb(r+k) + rb(b+k)) \\ &= r(r+k)(r+2k+b) + rb(r+2k+b) \\ &= r(r+k+b)(r+2k+b)\end{aligned}$$

The last two factors,  $r+k+b$  and  $r+2k+b$ , cancel with the same factors in the denominator, leaving  $\mathbb{P}R_3 = r/(r+b)$ .

REMARK. There is something wrong with the calculation of  $\mathbb{P}R_3$  in the case  $r = 1$  and  $k = -1$  if we interpret each of the factors in a product like

$$\frac{r}{r+b} \times \frac{r+k}{r+b+k} \times \frac{r+2k}{r+b+2k}$$

as a conditional probability. The third factor would become  $(-1)/(b-1)$ , which is negative: the urn had run out of balls after the previous draw. Fortunately the second factor reduces to zero. The product of these factors is zero, which is the correct value for  $\mathbb{P}(R_1 R_2 R_3)$  when  $r = 1$  and  $k = -1$ . The oversight did not invalidate the final answer. *Moral: The value of a conditional probability needn't make sense if it is to be multiplied by zero.*

By now you probably suspect that the answer to question (b) is  $r/(r+b)$ , no matter what the value of  $k$ . A symmetry argument will prove your suspicions correct. Look for the pattern in probabilities like  $\mathbb{P}(R_1 R_2 B_3 \dots)$  when expressed as a ratio of two products. The successive factors in the denominator correspond to the numbers of balls in the urn before each draw. The same factors will appear no matter what string of  $R_i$ 's and  $B_i$ 's is involved. In the numerator, the first appearance of an  $R_i$  contributes an  $r$ , the second appearance contributes an  $r+k$ , and so on. The  $B_i$ 's contribute  $b$ , then  $b+k$ , then  $b+2k$ , and so on. For example,

$$\mathbb{P}(R_1 R_2 B_3 B_4 R_5 B_6 R_7) = \frac{r(r+k)(r+2k)(r+3k)b(b+k)(b+2k)}{(r+b)(r+b+k)(r+b+2k)(r+b+3k)\dots(r+b+6k)}$$

You might like to rearrange the order of the factors in the numerator to make the representation as a product of conditional probabilities clearer.

In short, the probability of a particular string of  $R_i$ 's and  $B_i$ 's, corresponding to a particular sequence of draws from the urn, depends only on the number of  $R_i$  and  $B_i$  terms, and not on their ordering.

### Answer to question (a)

For  $i = 0, 1, \dots, n$ , we need to calculate the probability of getting exactly  $i$  red balls amongst the first  $n$  draws. There are  $\binom{n}{i}$  different orderings for the first  $n$  draws that would give exactly  $i$  reds. (Think of the number of ways to choose the  $i$  positions for the red from the  $n$  available). The event  $\{i \text{ reds in first } n \text{ draws}\}$  is a disjoint union of  $\binom{n}{i}$  equally likely events, whence

$$\begin{aligned} & \mathbb{P}\{i \text{ reds in first } n \text{ draws}\} \\ &= \binom{n}{i} \mathbb{P}R_1 R_2 \dots R_i B_{i+1} B_{i+2} \dots B_n \\ &= \binom{n}{i} \frac{r(r+k)\dots(r+k(i-1))b(b+k)\dots(b+k(n-i-1))}{(r+b)(r+b+k)\dots(r+b+k(n-1))} \end{aligned}$$

As a quick check, notice that when  $k=0$ , the probability reduces to

$$\binom{n}{i} \left(\frac{r}{r+b}\right)^i \left(\frac{b}{r+b}\right)^{n-i},$$

as it should be for a  $\text{Bin}(n, r/(r+b))$  distribution.

For the special case of sampling without replacement ( $k = -1$ ), the probability becomes

$$\begin{aligned}
 & \binom{n}{i} \frac{r(r-1)\dots(r-i+1)b(b-1)\dots(b-n+i+1)}{(r+b)(r+b-1)\dots(r+b-n+1)} \\
 &= \frac{n!}{i!(n-i)!} \frac{r!}{(r-i)!} \frac{b!}{(b-n+i)!} \frac{(r+b-n)!}{(r+b)!} \\
 &= \frac{r!}{i!(r-i)!} \frac{b!}{(n-i)!(b-n+i)!} \frac{n!(r+b-n)!}{(r+b)!} \\
 &= \binom{r}{i} \binom{b}{n-i} / \binom{r+b}{n}.
 \end{aligned}$$

Notice that

$$\begin{aligned}
 \binom{r}{i} &= \text{number of ways to choose } i \text{ from } r \text{ reds} \\
 \binom{b}{n-i} &= \text{number of ways to choose } n-i \text{ from } b \text{ blacks} \\
 \binom{r+b}{n} &= \text{number of ways to choose } n \text{ from } r+b \text{ in urn}
 \end{aligned}$$

Compare the last probability with the direct calculation based on a sample space where all possible subsets from the urn are given equal probability.

Unless you subscribe to tricky conventions about factorials or binomial coefficients, you might want to restrict the last calculation to values of  $i$  and  $n$  for which

$$\begin{aligned}
 0 &\leq i \leq r \\
 0 &\leq n-i \leq b \\
 1 &\leq n \leq r+b
 \end{aligned}$$

**Definition.** A random variable  $X$  is said to have a **hypergeometric distribution** if it takes only integer values in the range  $\max(0, n-b) \leq i \leq r$ , with probabilities

$$\mathbb{P}\{X = i\} = \binom{r}{i} \binom{b}{n-i} / \binom{r+b}{n},$$

where the “parameters” are fixed positive integers for which  $n \leq r+b$ .

REMARK. How would you explain the form of the lower limit for the range of possible values?

I advise against trying to memorize the form of the hypergeometric distribution. Think of it as a slightly “compressed” analog of the  $\text{Bin}(n, r/(r+b))$  distribution. Rederive when absolutely necessary.

### Answer to question (b)

The symmetry property that lets us ignore the ordering when calculating probabilities for particular sequences of draws also lets us eliminate much of the algebra we first used to find  $\mathbb{P}R_3$ . Reconsider that case. We broke the event  $R_3$  into four disjoint pieces:

$$(R_1 R_2 R_3) \cup (R_1 B_2 R_3) \cup (B_1 R_2 R_3) \cup (B_1 B_2 R_3).$$

Each triple ends with an  $R_3$ , with the first two positions giving all possible  $R$  and  $B$  combinations. The probability for each triple is unchanged if we permute the subscripts, because ordering does not matter. Thus

$$\mathbb{P}R_3 = \mathbb{P}(R_3 R_2 R_1) + \mathbb{P}(R_3 B_2 R_1) + \mathbb{P}(B_3 R_2 R_1) + \mathbb{P}(B_3 B_2 R_1).$$



Notice how the triple for each term now ends in an  $R_1$  instead of an  $R_3$ . The last sum is just a decomposition for  $\mathbb{P}R_1$  obtaining by splitting according to the outcome of the second and third draws. It follows that  $\mathbb{P}R_3 = \mathbb{P}R_1$ . Similarly,

$$\mathbb{P}\{i\text{th ball is red}\} = \mathbb{P}R_1 = r/(r+b) \quad \text{for each } i.$$

### Answer to question (c)

You should resist the urge to use the answer to question (a) in a direct attack on question (c). Instead, write the number of reds in  $n$  draws as  $X_1 + \dots + X_n$ , where  $X_i$  denotes the indicator of the event  $R_i$ , that is,

$$X_i = \begin{cases} 1 & \text{if } i\text{th ball red} \\ 0 & \text{otherwise} \end{cases}$$

From the answer to question (b),

$$\mathbb{E}X_i = 1\mathbb{P}\{X_i = 1\} + 0\mathbb{P}\{X_i = 0\} = \mathbb{P}R_i = r/(r+b).$$

It follows that the expected number of reds in the sample of  $n$  is  $nr/(r+b)$ . This expected number does not depend on  $k$ ; it is the same for  $k = 0$  (sampling with replacement, draws independent) and  $k \neq 0$  (draws are dependent), provided we exclude cases where the urn gets emptied out before the  $n$ th draw.  $\square$

# EXAMPLE 21: THE GAME OF BET RED

*A pack of cards consists of 26 reds and 26 blacks. I shuffle the cards, then deal them out one at a time, face up. You are given the chance to win a big prize by correctly predicting when the next card to be dealt will be red. You are allowed to make the prediction for only one card, and you must predict red, not black. What strategy should you adopt to maximize your probability of winning the prize?*

First let us be clear on the rules. Your strategy will predict that card  $\tau + 1$  is red, where  $\tau$  is one of the values  $0, 1, \dots, 51$ . That is, you observe the first  $\tau$  cards then predict that the next one will be red. The value of  $\tau$  is allowed to depend on the cards you observe. For example, a decision to choose  $\tau = 3$  can be based on the observed colors of cards  $0, 1, 2$ , and  $3$ ; but it cannot use information about cards  $4, 5, \dots, 52$ .

REMARK. In the probability jargon,  $\tau$  is called a stopping rule, or stopping time, or several other terms that make sense in other contexts.

Here are some simple-minded strategies: always choose the first card (probability  $1/2$  of winning); always choose the last card (probability  $1/2$  of winning). A more complicated strategy: if the first card is black choose card 2, otherwise choose card 52, which gives

$$\begin{aligned}\mathbb{P}\{\text{win}\} &= \mathbb{P}\{\text{first red, last red}\} + \mathbb{P}\{\text{first black, second red}\} \\ &= \frac{1}{2} \cdot \frac{25}{51} + \frac{1}{2} \cdot \frac{26}{51} \\ &= \frac{1}{2}.\end{aligned}$$

Notice the hidden appeal to (conditional) symmetry to calculate

$$\mathbb{P}\{\text{last red} \mid \text{first red}\} = \mathbb{P}\{\text{second red} \mid \text{first red}\} = \frac{25}{51}.$$

All three strategies give the same probability of a win.

We have to be a bit more cunning. How about: wait until the proportion of reds in the remaining cards is high enough and then go for the next card. As you will soon see, the extra cunning gets us nowhere, because all strategies have the same probability,  $1/2$ , of winning. Amazing!

Consider first an analogous problem for a pack of 3 red and 3 black cards. Why doesn't the following strategy improve one's chances of winning?

WAIT UNTIL  
NUMBER OF REDS OBSERVED IS  $<$  NUMBER OF BLACKS OBSERVED,  
THEN CHOOSE THE NEXT CARD.

With such a small deck we are able to list all possible ways that the cards might appear, calculate  $\tau$  for each outcome, then calculate the probability of a win. There are  $\binom{6}{3} = 20$  possible orderings of 3 reds and 3 blacks, each equally likely. (Here I am treating all red cards as equivalent. You could construct a more detailed sample space, with  $6!$  orderings for the 6 cards, but the calculations would end up with the same conclusion.) With  $r$  denoting a red card, and  $b$  a black card, the outcomes are:

pattern	value of $\tau$	win?
b <b><u>b</u></b> rrr	1	
b <b><u>b</u></b> rbr	1	
b <b><u>b</u></b> rrbr	1	
b <b><u>b</u></b> rrrb	1	
b <b><u>r</u></b> bbrr	1	✓
b <b><u>r</u></b> brbr	1	✓
b <b><u>r</u></b> brrb	1	✓
b <b><u>r</u></b> rbbr	1	✓
b <b><u>r</u></b> rbbr	1	✓
b <b><u>r</u></b> rrbb	1	✓
rbb <b><u>b</u></b> rr	3	
rbb <b><u>b</u></b> rbr	3	✓
rbb <b><u>b</u></b> rb	3	✓
rbrbb <b><u>r</u></b>	5	✓
rbrbrb	?	
rbrrbb	?	
rbbbb <b><u>r</u></b>	5	✓
rbbrbr	?	
rrbrbb	?	
rrrbbb	?	

Where possible I have underlined the card that the strategy would predict to be red. Even though the game ends after the card is predicted, I have written out the whole string, to make calculation of probabilities a mere matter of counting up equally probable events. Notice that in 5 cases (rbrbrb, ..., rrrbbb) the strategy fails to predict. We could modify the strategy by adding

..., BUT IF ONLY ONE CARD REMAINS, CHOOSE IT.

Notice that the addendum has no effect on the probability of a win. There are still only 10 of the 20 equally likely cases that lead to win. The strategy again has probability 1/2 of winning.

The enumeration of outcomes gives a clue to why we keep coming back to 1/2. Look, for example, at the block of ten outcomes beginning  $b????$ . Each of them gives  $\tau = 1$ . There are only ten possible continuations, each having conditional probability 1/10. The strategy  $\tau$  has conditional probability 6/10 of leading to a win; six of the ten possible continuations have an r where  $\tau$  wants it. By symmetry, six of the ten possible continuations have an r in the last position. Thus

$$\mathbb{P}\{\tau \text{ wins} \mid b????\} = \mathbb{P}\{br???? \mid b????\} = \mathbb{P}\{b????r \mid b????\}.$$

It follows that  $\tau$  has the same conditional probability for a win as the strategy for which  $\tau \equiv 5$ .

Now try the same idea on the original problem. Consider a string  $x_1, x_2, \dots, x_{52}$  of 26 reds and 26 blacks in some order such that a strategy  $\tau$  would choose card  $i$ . The strategy must be using information from only the first  $i$  cards. Consider sequences

$$x_1, x_2, \dots, x_i, ? \dots ?$$

that give  $\tau = i$ . These sequences must have the same  $i$  cards at the start. (The particular  $x_1 \dots x_i$  depend on the strategy.) Conditioning on this starting fragment, which triggered the choice  $\tau = i$ , we get

$$\begin{aligned} \mathbb{P}\{\tau \text{ wins} \mid x_1, x_2, \dots, x_i, ? \dots ?\} &= \mathbb{P}\{x_1, x_2, \dots, x_i, r, ? \dots ? \mid x_1, x_2, \dots, x_i, ? \dots ?\} \\ &= \mathbb{P}\{x_1, x_2, \dots, x_i, ? \dots ?r \mid x_1, x_2, \dots, x_i, ? \dots ?\}. \end{aligned}$$

If we write LAST for the strategy of always choosing the 52nd card, the equality becomes

$$\mathbb{P}\{\tau \text{ wins} \mid x_1, x_2, \dots, x_i, ? \dots ?\} = \mathbb{P}\{\text{LAST wins} \mid x_1, x_2, \dots, x_i, ? \dots ?\}$$

Multiply both sides by  $\mathbb{P}\{x_1, x_2, \dots, x_i, ? \dots ?\}$  then sum over all possible starting fragments that trigger a choice for  $\tau$  to deduce that

$$\mathbb{P}\{\tau \text{ wins}\} = \mathbb{P}\{\text{LAST wins}\} = 1/2.$$

Maybe the LAST strategy is not so simple-minded after all.

□

## EXAMPLE 22: THE BALLOT THEOREM

Suppose an urn contains  $r$  red balls and  $b$  black balls, with  $r > b$ . As balls are sampled without replacement from the urn, keep track of the total number of red balls removed and the total number of black balls removed after each draw. Show that the probability of the event {number of reds removed is strictly greater than the number of blacks removed, after every draw} is equal to  $(r - b)/(r + b)$ .

For simplicity, I will refer to the event whose probability we seek as “red always leads”.

The sampling scheme should be understood to imply that all  $(r + b)!$  orderings of the balls (treating balls of the same color as distinguishable for the moment) are equally likely. There is a sneaky way to generate a random permutation, which will lead to an elegant solution to the problem.

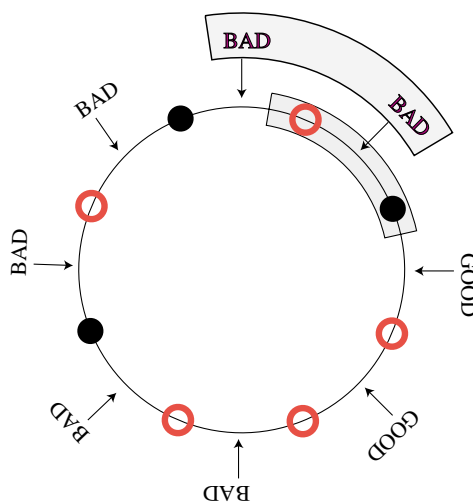
Imagine that the balls are placed into a circular track as they are removed, without any special marker to indicate the position of the first ball. After every ball is placed in the track, choose a starting position at random, with each of the  $r + b$  possible choices equally likely, then select the balls in order moving clockwise from the starting position.

To calculate  $\mathbb{P}\{\text{red always leads}\}$ , condition on the “circle”, the ordering of the balls around the circular track. I will show that

$$(*) \quad \mathbb{P}\{\text{red always leads} \mid \text{circle}\} = \frac{r - b}{r + b},$$

for every circle configuration. Regardless of the probabilities of the various circle configuration, the weighted average of these conditional probabilities must give the asserted result.

The calculation of the conditional probability in (\*) reduces to a simple matter of counting: How many of the  $r + b$  possible starting positions generate a “GOOD” permutation where red always leads?



Imagine the  $r + b$  positions labelled as GOOD or BAD, as in the picture. Somewhere around the circle there must exist a pair red-black, with the black ball immediately following the red ball in the clockwise ordering.

Two of the positions—the one between the red-black pair, and the one just before the initial red—are obviously bad. (Look at the first few balls in the resulting permutation.)

Consider the effect on the total number (not probability) of GOOD starting positions if the red-black pair is removed from the track. Two BAD starting positions are eliminated immediately. It is less obvious, but true, that removal of the pair has no effect on any of the

other starting positions: a GOOD starting position stays GOOD, and a BAD starting position stays BAD. (Consider the effect on the successive red and black counts.) The total number of GOOD starting positions is unchanged.

Repeat the argument with the new circle configuration of  $r + b - 2$  balls, eliminating one more red-black pair but leaving the total GOOD count unchanged. And so on.

After removal of  $b$  red-black pairs all  $r - b$  remaining balls are red, and all  $r - b$  starting positions are GOOD. Initially, therefore, there must also have been  $r - b$  of the GOOD positions out of the  $r + b$  available. The assertion (\*), and thence the main assertion, follow.  $\square$

Reconsider the Bet Red problem. The strategy of waiting for the proportion of red cards left in the deck to exceed  $1/2$ , then betting on the next red, works except when the proportion of reds never gets above  $1/2$ . How likely is that? The answer can be deduced from Example 22.

If a deck contains  $n + 1$  red cards and  $n$  black cards then

$$\mathbb{P}\{\text{\#reds sampled} > \text{\#blacks sampled, always}\} = \frac{1}{2n + 1}.$$

If we condition on the first card being red, then we get

$$\frac{1}{2n + 1} = \frac{n + 1}{2n + 1} \mathbb{P}\{\text{subsequent \#reds} \geq \text{\#blacks} \mid \text{first card red}\}.$$

The conditional probability is the same as the probability, for a deck of  $n$  red cards and  $n$  black cards, that the number of black cards dealt never strictly exceeds the number of red cards dealt. Solving for that probability, we find that the strategy of waiting for a higher proportion of reds in the deck will fail with probability  $1/(n + 1)$  for a deck of  $n$  red and  $n$  black cards. The probability might not seem very large, but apparently it is just large enough to offset the slight advantage gained when the strategy works.

## Chapter 5

# Variances and covariances

The expected value of a random variable gives a crude measure for the “center of location” of the distribution of that random variable. For instance, if the distribution is symmetric about a value  $\mu$  then the expected value equals  $\mu$ . To refine the picture of a distribution distributed about its “center of location” we need some measure of spread (or concentration) around that value. The simplest measure to calculate for many distributions is the **variance** (or, more precisely, the square root of the variance).

**Definition.** The **variance** of a random variable  $X$  with expected value  $\mathbb{E}X = \mu_X$  is defined as  $\text{var}(X) = \mathbb{E}((X - \mu_X)^2)$ . The **covariance** between random variables  $Y$  and  $Z$ , with expected values  $\mu_Y$  and  $\mu_Z$ , is defined as  $\text{cov}(Y, Z) = \mathbb{E}((Y - \mu_Y)(Z - \mu_Z))$ . The **correlation** between  $Y$  and  $Z$  is defined as

$$\text{corr}(Y, Z) = \frac{\text{cov}(Y, Z)}{\sqrt{\text{var}(Y)\text{var}(Z)}}$$

The square root of the variance of a random variable is called its **standard deviation**, sometimes denoted by  $\text{sd}(X)$ .

REMARK. Notice that  $\text{cov}(X, X) = \text{var}(X)$ . Results about covariances contain results about variances as special cases.

Sometimes it is easier to subtract off the expected values at the end of the calculation, by means of the formulae  $\text{cov}(Y, Z) = \mathbb{E}(YZ) - (\mathbb{E}Y)(\mathbb{E}Z)$  and, as a particular case,  $\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$ . Both formulae follow via an expansion of the product:

$$\begin{aligned}\text{cov}(Y, Z) &= \mathbb{E}(YZ - \mu_Y Z - \mu_Z Y + \mu_Y \mu_Z) \\ &= \mathbb{E}(YZ) - \mu_Y \mathbb{E}Z - \mu_Z \mathbb{E}Y + \mu_Y \mu_Z \\ &= \mathbb{E}(YZ) - \mu_Y \mu_Z.\end{aligned}$$

A pair of random variables  $X$  and  $Y$  is said to be **uncorrelated** if  $\text{cov}(X, Y) = 0$ , and **independent** if “every event determined by  $X$  is independent of every event determined by  $Y$ ”. For example, independence implies that events such as  $\{X \leq 5\}$  and  $\{7 \leq Y \leq 18\}$  are independent, and so on. Independence of the random variables also implies independence of functions of those random variables. For example,  $\sin(X)$  would be independent of  $e^Y$ , and so on. Independent random variables are uncorrelated, but uncorrelated random variables need not be independent.

---

**Example 23:** An argument showing that independent random variables are uncorrelated, with an example of uncorrelated random variables that are dependent.

---

The variance of a random variable  $X$  is unchanged by an added constant:  $\text{var}(X + C) = \text{var}(X)$  for every constant  $C$ , because  $(X + C) - \mathbb{E}(X + C) = X - \mathbb{E}X$ , the  $C$ ’s cancelling. It is

a desirable property that the spread should not be affected by a change in location. However, it is also desirable that multiplication by a constant should change the spread:  $\text{var}(CX) = C^2\text{var}(X)$  and  $\text{sd}(CX) = |C|\text{sd}(X)$ , because  $(CX - \mathbb{E}(CX))^2 = C^2(X - \mathbb{E}X)^2$ . In summary:

$$\text{var}(a + bX) = b^2\text{var}(X) \text{ and } \text{sd}(a + bX) = |b|\text{sd}(X) \text{ for constants } a \text{ and } b.$$

REMARK. Try not to confuse properties of expected values with properties of variances: for constants  $a$  and  $b$  we have  $\text{var}(a + bX) = b^2\text{var}(X)$  but  $\mathbb{E}(a + bX) = a + b\mathbb{E}X$ . Measures of location (expected value) and spread (standard deviation) should react differently to linear transformations of the variable. As another example: if a given piece of “information” implies that a random variable  $X$  must take the constant value  $C$  then  $\mathbb{E}(X \mid \text{information}) = C$ , but  $\text{var}(X \mid \text{information}) = 0$ .

It is also a common blunder to confuse the formula for the variance of a difference with the formula  $\mathbb{E}(Y - Z) = \mathbb{E}Y - \mathbb{E}Z$ . If you ever find yourself wanting to assert that  $\text{var}(Y - Z)$  is equal to  $\text{var}(Y) - \text{var}(Z)$ , think again. What would happen if  $\text{var}(Z)$  were larger than  $\text{var}(Y)$ ? Variances can't be negative.

Covariances enter the picture when we consider variances of sums of random variables. For example,  $\text{var}(Y + Z) = \text{var}(Y) + 2\text{cov}(Y, Z) + \text{var}(Z)$ , a result that follows by taking expectations of both sides of the expansion

$$(X + Y - \mu_X - \mu_Y)^2 = (X - \mu_X)^2 + 2(X - \mu_X)(Y - \mu_Y) + (Y - \mu_Y)^2.$$

More generally, For constants  $a, b, c, d$ , and random variables  $U, V, Y, Z$ ,

$$\begin{aligned} \text{cov}(aU + bV, cY + dZ) \\ = ac \text{cov}(U, Y) + bc \text{cov}(V, Y) + ad \text{cov}(U, Z) + bd \text{cov}(V, Z). \end{aligned}$$

It is easier to see the pattern if we work with the centered random variables  $U' = U - \mu_U, \dots, Z' = Z - \mu_Z$ . For then the left-hand side becomes

$$\begin{aligned} \mathbb{E}((aU' + bV')(cY' + dZ')) &= \mathbb{E}(ac U'Y' + bc V'Y' + ad U'Z' + bd V'Z') \\ &= ac \mathbb{E}(U'Y') + bc \mathbb{E}(V'Y') + ad \mathbb{E}(U'Z') + bd \mathbb{E}(V'Z'). \end{aligned}$$

The expected values in the last line correspond to the covariances.

If  $Y$  and  $Z$  are uncorrelated, the covariance term drops out from the expression for the variance of their sum, leaving  $\text{var}(Y + Z) = \text{var}(Y) + \text{var}(Z)$ . Similarly, if  $X_1, \dots, X_n$  are random variables for which  $\text{cov}(X_i, X_j) = 0$  for each  $i \neq j$  then

$$\text{var}(X_1 + \dots + X_n) = \text{var}(X_1) + \dots + \text{var}(X_n) \quad \text{for “pairwise uncorrelated” rv's.}$$

You should check the last assertion by expanding out the quadratic in the variables  $X_i - \mathbb{E}X_i$ , observing how all the cross-product terms disappear because of the zero covariances.

There is an enormous body of probability literature that deals with approximations to distributions, and bounds for probabilities, expressible in terms of expected values and variances. One of the oldest and simplest examples, the Tchebychev inequality, is still useful, even though it is rather crude by modern standards.

---

#### Example 24: The Tchebychev inequality

---

The Tchebychev bound explains an important property of sample means: their distributions concentrate increasingly around their expectations as the sample size increases.

---

#### Example 25: Concentration of sample mean about expected value

---

The concentration phenomenon will also hold for averages of dependent random variables, if the variance is small.

---

#### Example 26: Comparison of spread in sample averages for sampling with and without replacement.

---



As with expectations, variances and covariances can also be calculated conditionally on various pieces of information. The conditioning formula in the final Example has the interpretation of a decomposition of “variability” into distinct sources, a precursor to the statistical technique known as the “analysis of variance”.

---

Example 27: An example to show how variances can sometimes be decomposed into components attributable to different sources. (Can be skipped.)

---

### Things to remember

- the initial definitions of variance and covariance, and their expanded forms  $\text{cov}(Y, Z) = \mathbb{E}(YZ) - (\mathbb{E}Y)(\mathbb{E}Z)$  and  $\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$
- $\text{var}(a + bX) = b^2\text{var}(X)$  and  $\text{sd}(a + bX) = |b|\text{sd}(X)$  for constants  $a$  and  $b$ .
- Sampling without replacement gives smaller variances than sampling with replacement.

### EXAMPLE 23: INDEPENDENT VERSUS UNCORRELATED

Suppose a random variable  $X$  can take values  $x_1, x_2, \dots$ . The expected value  $\mathbb{E}(XY)$  can then be rewritten as a weighted sum of conditional expectations,

$$\begin{aligned}\mathbb{E}(XY) &= \sum_i \mathbb{P}\{X = x_i\} \mathbb{E}(XY \mid X = x_i) && \text{by rule E4 for expectations} \\ &= \sum_i \mathbb{P}\{X = x_i\} x_i \mathbb{E}(Y \mid X = x_i).\end{aligned}$$

If  $Y$  is independent of  $X$ , the information “ $X = x_i$ ” does not help with the calculation of the conditional expectation,  $\mathbb{E}(Y \mid X = x_i) = \mathbb{E}(Y)$ . The last calculation then simplifies to

$$\mathbb{E}(XY) = (\mathbb{E}Y) \sum_i x_i \mathbb{P}\{X = x_i\} = (\mathbb{E}Y)(\mathbb{E}X).$$

It follows that  $\text{cov}(X, Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y) = 0$  if  $Y$  and  $X$  are independent.

Uncorrelated random variables need not be independent. Consider the following example. For two independent rolls of a fair die, let  $X$  denote the value rolled the first time and  $Y$  denote the value rolled the second time. The random variables  $X$  and  $Y$  are independent, and they have the same distribution. Consequently  $\text{cov}(X, Y) = 0$ , and  $\text{var}(X) = \text{var}(Y)$ .

The two random variables  $X + Y$  and  $X - Y$  are uncorrelated,

$$\begin{aligned}\text{cov}(X + Y, X - Y) &= \text{cov}(X, X) + \text{cov}(X, -Y) + \text{cov}(Y, X) + \text{cov}(Y, -Y) \\ &= \text{var}(X) - \text{cov}(X, Y) + \text{cov}(Y, X) - \text{var}(Y) \\ &= 0.\end{aligned}$$

Nevertheless, the sum and difference are not independent. For example,

$$\mathbb{P}\{X + Y = 12\} = \mathbb{P}\{X = 6\} \mathbb{P}\{Y = 6\} = \frac{1}{36}$$

but

$$\mathbb{P}\{X + Y = 12 \mid X - Y = 5\} = \mathbb{P}\{X + Y = 12 \mid X = 6, Y = 1\} = 0.$$

□

#### EXAMPLE 24: THE TCHEBYCHEV INEQUALITY

The inequality asserts: for a random variable  $X$  with expected value  $\mu$ ,

$$\mathbb{P}\{|X - \mu| > \epsilon\} \leq \text{var}(X)/\epsilon^2 \quad \text{for each } \epsilon > 0.$$

The inequality becomes obvious if we write  $F$  for the event  $\{|X - \mu| > \epsilon\}$ . First note that  $\mathbb{I}_F \leq |X - \mu|^2/\epsilon^2$ : when  $\mathbb{I}_F = 0$  the inequality holds for trivial reasons; and when  $\mathbb{I}_F$  takes the value one, the random variable  $|X - \mu|^2$  must be larger than  $\epsilon^2$ . It follows that

$$\mathbb{P}\{|X - \mu| > \epsilon\} = \mathbb{E}\mathbb{I}_F \leq \mathbb{E}|X - \mu|^2/\epsilon^2.$$

In the Chapter on the normal distribution you will find more refined probability approximations involving the variance. □

# EXAMPLE 25: CONCENTRATION OF SAMPLE MEANS

Suppose  $X_1, \dots, X_n$  are uncorrelated random variables, each with expected value  $\mu$  and variance  $\sigma^2$ . By repeated application of the formulae for the variance of a sum of variables with zero covariances,

$$\text{var}(X_1 + \dots + X_n) = \text{var}(X_1) + \dots + \text{var}(X_n) = n\sigma^2.$$

Typically the  $X_i$  would come from repeated independent measurements of some unknown quantity. The random variable  $\bar{X} = (X_1 + \dots + X_n)/n$  is then called the **sample mean**.

The variance of the sample mean decreases like  $1/n$ ,

$$\text{var}(\bar{X}) = (1/n)^2 \text{var}(X_1 + \dots + X_n) = \sigma^2/n.$$

From the Tchebychev inequality,

$$\mathbb{P}\{|\bar{X} - \mu| > \epsilon\} \leq (\sigma^2/n)/\epsilon^2 \quad \text{for each } \epsilon > 0.$$

In particular, for each positive  $C$ ,

$$\mathbb{P}\{|\bar{X} - \mu| > C\sigma/\sqrt{n}\} \leq 1/C^2.$$

For example, there is at most a 1% chance that  $\bar{X}$  lies more than  $10\sigma/\sqrt{n}$  away from  $\mu$ . (A normal approximation will give a much tighter bound.) Note well the dependence on  $n$ .  $\square$

# EXAMPLE 26: VARIABILITY OF A SAMPLE AVERAGE

In the decennial census of housing and population, the Census Bureau obtain some information from a more extensive list of questions sent to only a random sample of housing units. For an area like New Haven, about 1 in 6 units receive the so-called “long form”.

For example, one question on the long form asks for the number of rooms in the housing unit. We could imagine the population of all units numbered  $1, 2, \dots, N$ , with the  $i$ th unit containing  $y_i$  rooms. Complete enumeration would reveal the value of the **population average**,

$$\bar{y} = \frac{1}{N} (y_1 + y_2 + \dots + y_N).$$

A sample can provide a good estimate of  $\bar{y}$  with less work.

Suppose a sample of  $n$  housing units are selected from the population without replacement. (For the decennial census,  $n \approx N/6$ .) The answer from each unit is a random variable that could take each of the values  $y_1, y_2, \dots, y_N$ , each with probability  $1/N$ .

REMARK. It might be better to think of a random variable that takes each of the values  $1, 2, \dots, N$  with probability  $1/N$ , then take the corresponding number of rooms as the value of the random variable that is recorded. Otherwise we can fall into verbal ambiguities when many of the units have the same number of rooms.

That is, the sample consists of random variables  $Y_1, Y_2, \dots, Y_n$ , for each of which

$$\mathbb{P}\{Y_i = y_j\} = \frac{1}{N} \quad \text{for } j = 1, 2, \dots, N.$$

Notice that

$$\mathbb{E}Y_i = \frac{1}{N} \sum_{j=1}^N y_j = \bar{y},$$

and consequently, the sample average  $\bar{Y} = (Y_1 + \dots + Y_n)/n$  also has expected value  $\bar{y}$ . Notice also that each  $Y_i$  has the same variance,

$$\text{var}(Y_i) = \frac{1}{N} \sum_{j=1}^N (y_j - \bar{y})^2,$$

a quantity that I will denote by  $\sigma^2$ .

If the sample is taken without replacement—which, of course, the Census Bureau must do, if only to avoid media ridicule—the random variables are dependent. For example, in the extreme case where  $n = N$ , we would necessarily have

$$Y_1 + Y_2 + \dots + Y_N = y_1 + y_2 + \dots + y_N,$$

in which case  $Y_N$  would be a function of the other  $Y_i$ 's, a most extreme form of dependence. Even if  $n < N$ , there is still some dependence, as you will soon see.

Sampling with replacement would be mathematically simpler, for then the random variables  $Y_i$  would be independent, and, as in Example 25, we would have  $\text{var}(\bar{Y}) = \sigma^2/n$ . With replacement, it is possible that the same unit might be sampled more than once, especially if the sample size is an appreciable fraction of the population size. There is also some

inefficiency in sampling with replacement, as shown by a calculation of variance for sampling without replacement.

$$\begin{aligned}
 \text{var}(\bar{Y}) &= \mathbb{E}(\bar{Y} - \bar{y})^2 \\
 &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n (Y_i - \bar{y})\right)^2 \\
 &= \frac{1}{n^2} \mathbb{E}\left(\sum_{i=1}^n (Y_i - \bar{y})^2 + 2 \sum_{1 \leq i < j \leq n} (Y_i - \bar{y})(Y_j - \bar{y})\right) \\
 &= \frac{1}{n^2} \left(\sum_{i=1}^n \mathbb{E}(Y_i - \bar{y})^2 + 2 \sum_{1 \leq i < j \leq n} \mathbb{E}((Y_i - \bar{y})(Y_j - \bar{y}))\right) \\
 &= \frac{1}{n^2} \left(\sum_{i=1}^n \text{var}(Y_i) + \sum_{1 \leq i \neq j \leq n} \text{cov}(Y_i, Y_j)\right) \quad \text{What formula did I just rederive?}
 \end{aligned}$$

There are  $n$  variance terms and  $n(n-1)$  covariance terms. We know that each  $Y_i$  has variance  $\sigma^2$ , regardless of the dependence between the variables. The effect of the dependence shows up in the covariance terms. By symmetry,  $\text{cov}(Y_i, Y_j)$  is the same for each pair  $i < j$ , a value that I will denote by  $c$ . Thus, for sampling without replacement,

$$(*) \quad \text{var}(\bar{Y}) = \frac{1}{n^2} (n\sigma^2 + n(n-1)c) = \frac{\sigma^2}{n} + \frac{(n-1)c}{n}.$$

We can calculate  $c$  directly, from the fact that the pair  $(Y_1, Y_2)$  takes each of  $N(N-1)$  pairs of values  $(y_i, y_j)$  with equal probability. Thus

$$c = \text{cov}(Y_1, Y_2) = \frac{1}{N(N-1)} \sum_{i \neq j} (y_i - \bar{y})(y_j - \bar{y}).$$

If we added the “diagonal” terms  $(y_i - \bar{y})^2$  to the sum we would have the expansion for the product

$$\sum_{i=1}^N (y_i - \bar{y}) \sum_{j=1}^N (y_j - \bar{y}),$$

which equals zero because  $N\bar{y} = \sum_{i=1}^N y_i$ . The expression for the covariance simplifies to

$$c = \text{cov}(Y_1, Y_2) = \frac{1}{N(N-1)} \left(0^2 - \sum_{i=1}^N (y_i - \bar{y})^2\right) = -\frac{\sigma^2}{N-1}.$$

Substitution in formula (\*) then gives

$$\text{var}(\bar{Y}) = \frac{\sigma^2}{n} \left(1 - \frac{n-1}{N-1}\right) = \frac{\sigma^2}{n} \frac{N-n}{N-1}.$$

Compare with the  $\sigma^2/n$  for  $\text{var}(\bar{Y})$  under sampling with replacement. The **correction factor**  $(N-n)/(N-1)$  is close to 1 if the sample size  $n$  is small compared with the population size  $N$ , but it can decrease the variance of  $\bar{Y}$  appreciably if  $n/N$  is not small. For example, if  $n \approx N/6$  (as with the Census long form) the correction factor is approximately 5/6.

If  $n = N$ , the correction factor is zero. That is,  $\text{var}(\bar{Y}) = 0$  if the whole population is sampled. Indeed, when  $n = N$  we know that  $\bar{Y}$  equals the population mean,  $\bar{y}$ , a constant. A random variable that always takes the same constant value has zero variance. Thus the right-hand side of (\*) must reduce to zero when we put  $n = N$ , which gives a quick method for establishing the equality  $c = -\sigma^2/(N-1)$ , without all the messing around with sums of products and products of sums.  $\square$

# EXAMPLE 27: DECOMPOSITION OF VARIANCE

Consider a two stage method for generating a random variable. Suppose we have  $k$  different random variables  $Y_1, \dots, Y_k$ , with  $\mathbb{E}Y_i = \mu_i$  and  $\text{var}(Y_i) = \sigma_i^2$ . Suppose also that we have a random method for selecting which variable to choose: a random variable  $X$  that is independent of all the  $Y_i$ 's, with  $\mathbb{P}\{X = i\} = p_i$  for  $i = 1, 2, \dots, k$ , where  $p_1 + p_2 + \dots + p_k = 1$ . If  $X$  takes the value  $i$ , define  $Z$  to equal  $Y_i$ .

The variability in  $Z$  is due to two effects: the variability of each  $Y_i$ ; and the variability of  $X$ . Conditional on  $X = i$ , we have  $Z$  equal to  $Y_i$ , and

$$\begin{aligned}\mathbb{E}(Z | X = i) &= \mathbb{E}(Y_i) = \mu_i \\ \text{var}(Z | X = i) &= \mathbb{E}((Z - \mu_i)^2 | X = i) = \text{var}(Y_i) = \sigma_i^2.\end{aligned}$$

From the first formula we get

$$\mathbb{E}Z = \sum_i \mathbb{P}\{X = i\} \mathbb{E}(Z | X = i) = \sum_i p_i \mu_i,$$

a weighted average of the  $\mu_i$ 's that I will denote by  $\bar{\mu}$ . A similar conditioning exercise gives

$$\text{var}(Z) = \mathbb{E}(Z - \bar{\mu})^2 = \sum_i p_i \mathbb{E}((Z - \bar{\mu})^2 | X = i).$$

If we could replace the  $\bar{\mu}$  in the  $i$ th summand by  $\mu_i$ , the sum would become a weighted average of conditional variances. To achieve such an effect, rewrite  $(Z - \bar{\mu})^2$  as

$$(Z - \mu_i + \mu_i - \bar{\mu})^2 = (Z - \mu_i)^2 + 2(\mu_i - \bar{\mu})(Z - \mu_i) + (\mu_i - \bar{\mu})^2.$$

Taking conditional expectations, we then get

$$\mathbb{E}((Z - \bar{\mu})^2 | X = i) = \mathbb{E}((Z - \mu_i)^2 | X = i) + 2(\mu_i - \bar{\mu})\mathbb{E}(Z - \mu_i | X = i) + (\mu_i - \bar{\mu})^2.$$

On the right-hand side, the first term equals  $\sigma_i^2$ , and the middle term disappears because  $\mathbb{E}(Z | X = i) = \mu_i$ . With those simplifications, the expression for the variance becomes

$$\text{var}(Z) = \sum_i p_i \sigma_i^2 + \sum_i p_i (\mu_i - \bar{\mu})^2.$$

If we think of each  $Y_i$  as coming from a separate “population”, the first sum represents the component of variability within the populations, and the second sum represents the variability between the populations.

The formula is sometimes written symbolically as

$$\text{var}(Z) = \mathbb{E}(\text{var}(Z | X)) + \text{var}(\mathbb{E}(Z | X)),$$

where  $\mathbb{E}(Z | X)$  denotes the random variable that takes the value  $\mu_i$  when  $X$  takes the value  $i$ , and  $\text{var}(Z | X)$  denotes the random variable that takes the value  $\sigma_i^2$  when  $X$  takes the value  $i$ . □

## Chapter 6

# Continuous Distributions

All the distributions we have met so far have been **discrete**: the possible values that the random variable could take were a finite set, as in  $0, 1, \dots, n$  for the  $\text{Bin}(n, p)$ , or a sequence, as in  $1, 2, 3, \dots$  for the  $\text{geometric}(p)$ . We shall also encounter random variables with **continuous distributions**, that is, random variables that take values in a continuous range.

The simplest example of a continuous distribution is the  $\text{Uniform}[0, 1]$ , the distribution of a random variable  $U$  that takes values in the interval  $[0, 1]$ , with

$$\mathbb{P}\{a \leq U \leq b\} = b - a \quad \text{for all } 0 \leq a \leq b \leq 1.$$

We have to specify the distribution by describing the probability it puts in intervals, because, for each  $x$  in  $(0, 1)$ ,

$$\mathbb{P}\{U = x\} = \mathbb{P}\{x \leq U \leq x\} = x - x = 0.$$

The probability is smeared out so smoothly that none of it can pile up exactly at the point  $x$ . The next best thing would be to specify how much probability is given to small intervals around  $x$ ,

$$\mathbb{P}\{x \leq U \leq x + \delta\} = \delta \quad \text{for small enough } \delta > 0.$$

Notice that the amount of probability in the interval is exactly proportion to the length, provided that  $\delta$  is small enough that  $[x, x + \delta]$  does not poke outside  $[0, 1]$ .

**REMARK.** Of course, to actually simulate a  $\text{Uniform}[0, 1]$  distribution on a computer one would work with a discrete approximation. For example, if numbers were specified to only 7 decimal places, one would be approximating  $\text{Uniform}[0, 1]$  by a discrete distribution placing probabilities of about  $10^{-7}$  on a fine grid of about  $10^7$  equispaced points in the interval. You might think of the  $\text{Uniform}[0, 1]$  as a convenient idealization of the discrete approximation.

For a general continuous distribution, the probability in small intervals is again (approximately) proportional to the length of the small interval, but now the constant of proportionality need not be the same at every point.

**Definition.** A random variable  $Y$  is said to have a **continuous distribution with density function**  $g(\cdot)$  if  $\mathbb{P}\{t \leq Y \leq t + \delta\} = g(t)\delta + \text{terms of order } \delta \text{ or smaller, for each small interval } [t, t + \delta]$ .

Probabilities of larger intervals are given areas under the curve defined by the density function,

$$\mathbb{P}\{a \leq Y \leq b\} = \int_a^b g(t) dt \quad \text{for all intervals } [a, b].$$

The formula is obtained by splitting  $[a, b]$  into smaller intervals, to each of which the defining property of the density applies, then passing to a limit. More precisely, for a large



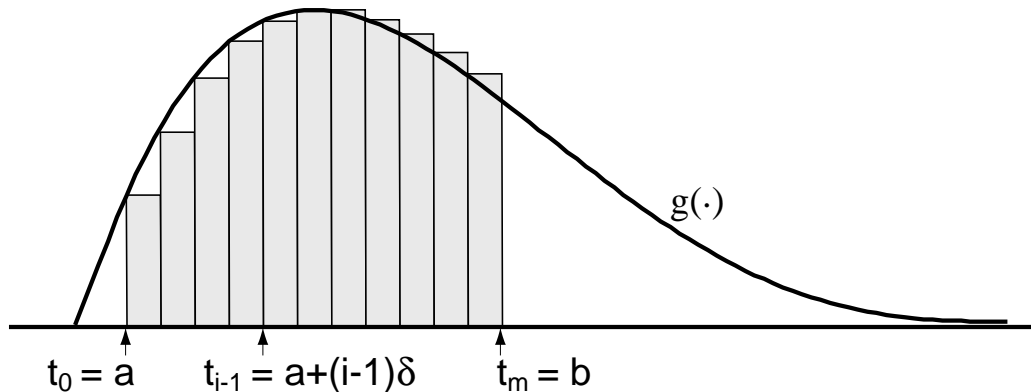
$m$  break  $[a, b]$  into a union of disjoint subintervals let  $I_1, \dots, I_m$  with lengths  $\delta = (b - a)/m$  and left end points  $t_{i-1} = a + (i - 1)\delta$ . When  $\delta$  is small enough,

$$P\{Y \in I_i\} = g(t_i)\delta + \text{terms of order } \delta^2 \text{ or smaller}$$

Sum over the subintervals.

$$\mathbb{P}\{Y \in [a, b]\} = \delta \sum_{i=1}^m g(t_i) + \text{remainder of order } \delta \text{ or smaller.}$$

Notice how  $m$  contributions of order  $\delta^2$  (or smaller) can amount to a remainder of order at worst  $\delta$  (or smaller), because  $m$  increases like  $1/\delta$ . (Can you make this argument rigorous?)



The sum  $\delta \sum_{i=1}^m g(t_i)$  corresponds to the shaded area in the picture. It is an approximation to the integral of  $g$  over  $[a, b]$ . As  $\delta$  tends to zero, the sum converges to that integral. The remainder terms tend to zero with  $\delta$ . The left-hand side just sits there. In the limit we get the asserted integral formula for  $\mathbb{P}\{Y \in [a, b]\}$ .

REMARK. Densities are usually defined via the integral property, rather than as constants of proportionality. The integral definition has the advantage that the probabilities are not affected by changes in the definition of  $g$  at isolated points. We don't really need to worry about the precise definition at end points of a range, for example. However, I find the interpretation as a constant of proportionality the more useful when calculating densities, or when deriving facts about continuous distributions.

Note well: the density  $g(t)$  is the constant of proportionality, and not a probability; it is not the same as  $\mathbb{P}\{Y = t\}$ , which is zero for every  $t$ . The density function,  $g$ , must be non-negative, for otherwise some tiny interval would receive a negative probability. Also it must integrate to one over the whole line, because  $1 = \mathbb{P}\{-\infty < Y < \infty\} = \int_{-\infty}^{\infty} g(t) dt$ .

REMARK. I prefer to think of densities as being defined on the whole real line, with values outside the range of the random variable being handled by setting the density function equal to zero appropriately. That way my integrals always run from  $-\infty$  to  $\infty$ , with the zero density killing off unwanted contributions. This convention will be useful when we consider densities that vanish outside a range depending on a parameter of the distribution; it will also help us avoid some amusing calculus blunders.

Calculations with continuous distributions typically involve calculations of integrals or derivatives, where discrete distribution involve sums or probabilities attached to individual points.

---

#### Example 28: Functions of a random variable with a continuous distribution

---

Here is a nontrivial example showing one method for finding a density. The trick is to work with very small intervals, so that higher order terms in the lengths of the intervals can be ignored. (More formally, the errors in approximation tend to zero as the intervals shrink.)

---

**Example 29: The distribution of the order statistics from the uniform distribution**

---

The distribution from the previous Example is a member of a family whose name is derived from the **beta function**, defined by

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1} (1-t)^{\beta-1} dt \quad \text{for } \alpha > 0, \beta > 0.$$

The equality

$$\int_0^1 t^{k-1} (1-t)^{n-k} dt = \frac{(k-1)!(n-k)!}{n!},$$

noted at the end of the Example, gives the value for  $B(k, n-k+1)$ .

In general, if we divide  $t^{\alpha-1}(1-t)^{\beta-1}$  by  $B(\alpha, \beta)$  we get a candidate for a density function: non-negative and integrating to 1.

**Definition.** For  $\alpha > 0$  and  $\beta > 0$  the *Beta*( $\alpha, \beta$ ) distribution is defined by the density function

$$\frac{t^{\alpha-1}(1-t)^{\beta-1}}{B(\alpha, \beta)} \quad \text{for } 0 < t < 1.$$

The density is zero outside  $(0, 1)$ .

For example, the  $k$ th order statistic from a sample of  $n$  independently generated random variables with Uniform[0, 1] distributions, from Example 29, is Beta( $k, n-k+1$ ) distributed.

---

**Example 30: The beta distribution: pictures and computing**

---

There is an interesting exact relationship between the tails of the beta and Binomial distributions.

---

**Example 31: Binomial tail probabilities from beta distributions**

---

**REMARK.** For many purposes it suffices to have a good approximation to the Binomial tail probabilities. The best known method—the normal approximation, due to de Moivre (1733)—will be described in Chapter 7. The relationship between Binomial and beta distributions can be used as the starting point for a particularly precise version of the normal approximation.

The formulae developed in previous chapters for expectations and variances of random variables have analogs for continuous distributions.

---

**Example 32: Expectation of a random variable with a continuous distribution**

---

**Things to remember**

- The density  $g(t)$  is the constant of proportionality, and not a probability; it is not the same as  $\mathbb{P}\{Y = t\}$ , which is zero for every  $t$ . A density function,  $g$ , must be non-negative, for otherwise some tiny interval would receive a negative probability. Also it must integrate to one over the whole line,  $1 = \mathbb{P}\{-\infty < Y < \infty\} = \int_{-\infty}^{\infty} g(t) dt$ .
- Expected value of a function of a random variable with a continuous distribution:

$$\mathbb{E}H(X) = \int_{-\infty}^{\infty} H(x)f(x) dx \quad \text{where } X \text{ has density } f.$$

- Be very careful not to confuse the formulae for expectations in the discrete and continuous cases. Think again if you find yourself integrating probabilities or summing expressions involving probability densities.

EXAMPLE 28: FUNCTIONS OF RANDOM VARIABLES WITH CONTINUOUS DISTRIBUTIONS

Suppose  $X$  has a uniform distribution on the interval  $(-\pi/2, \pi/2)$ . That is, it has a continuous distribution given by the density function

$$f(x) = \begin{cases} 1/\pi & \text{for } -\pi/2 < x < \pi/2 \\ 0 & \text{elsewhere} \end{cases}$$

Let a new random variable be defined by  $Y = \tan(X)$ . It takes values over the whole real line. For a fixed real  $y$ , and a small positive  $\delta$ , we have

$$(*) \quad y \leq Y \leq y + \delta \quad \text{if and only if} \quad x \leq X \leq x + \epsilon,$$

where  $x$  and  $\epsilon$  are related to  $y$  and  $\delta$  by the equalities

$$y = \tan(x) \quad \text{and} \quad y + \delta = \tan(x + \epsilon).$$

By Calculus,

$$\delta = \epsilon \times \frac{\tan(x + \epsilon) - \tan(x)}{\epsilon} \approx \frac{\epsilon}{\cos^2 x}.$$

Compare with the definition of the derivative:

$$\lim_{\epsilon \rightarrow 0} \frac{\tan(x + \epsilon) - \tan(x)}{\epsilon} = \frac{d \tan(x)}{dx} = \frac{1}{\cos^2 x}.$$

Thus

$$\begin{aligned} \mathbb{P}\{y \leq Y \leq y + \delta\} &= \mathbb{P}\{x \leq X \leq x + \epsilon\} \\ &\approx \epsilon f(x) \quad \text{definition of density for } X \\ &\approx \frac{\delta \cos^2 x}{\pi}. \end{aligned}$$

We need to express  $\cos^2 x$  as a function of  $y$ . Note that

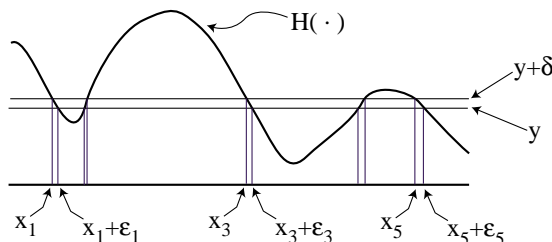
$$1 + y^2 = 1 + \frac{\sin^2 x}{\cos^2 x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x} = \frac{1}{\cos^2 x}.$$

Thus  $Y$  has a continuous distribution with density

$$g(y) = \frac{1}{\pi(1 + y^2)} \quad \text{for } -\infty < y < \infty.$$

REMARK. This distribution is called the Cauchy.

For functions that are not one-to-one, the analog of  $(*)$  can require a little more work. In general, we can have a random variable  $Y$  defined as  $H(X)$ , a function of another random variable. If  $X$  has a continuous distribution with density  $f$ , and if  $H$  is a smooth function with derivative  $H'$ , then we can calculate a density for  $Y$  by an extension of the method above.



A small interval  $[y, y + \delta]$  in the range of values taken by  $Y$  can correspond to a more complicated range of values for  $X$ . For instance, it might consist of a union of several intervals  $[x_1, x_1 + \epsilon_1]$ ,  $[x_2, x_2 + \epsilon_2]$ ,  $\dots$ . The number of pieces in the  $X$  range might be different for different values of  $y$ .

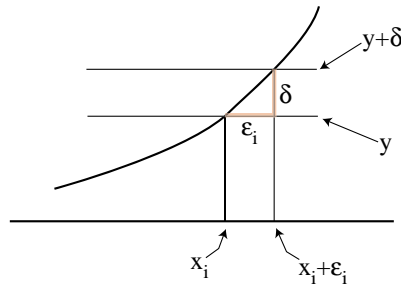
From the representation of  $\{y \leq Y \leq y + \delta\}$  as a disjoint union of events

$$\{x_1 \leq X \leq x_1 + \epsilon_1\} \cup \{x_2 \leq X \leq x_2 + \epsilon_2\} \cup \dots,$$

we get, via the defining property of the density  $f$  for  $X$ ,

$$\begin{aligned} \mathbb{P}\{y \leq Y \leq y + \delta\} &= \mathbb{P}\{x_1 \leq X \leq x_1 + \epsilon_1\} + \mathbb{P}\{x_2 \leq X \leq x_2 + \epsilon_2\} + \dots \\ &\approx \epsilon_1 f(x_1) + \epsilon_2 f(x_2) + \dots \end{aligned}$$

For each small interval, the ratio of  $\delta$  to  $\epsilon_i$  is close to the derivative of the function  $H$  at the corresponding  $x_i$ . That is,  $\epsilon_i \approx \delta / H'(x_i)$ .



Adding the contributions from each such interval, we then have an approximation that tells us the density for  $Y$ ,

$$\mathbb{P}\{y \leq Y \leq y + \delta\} \approx \delta \left( \frac{f(x_1)}{H'(x_1)} + \frac{f(x_2)}{H'(x_2)} + \dots \right)$$

That is, the density for  $Y$  at the particular point  $y$  in its range equals

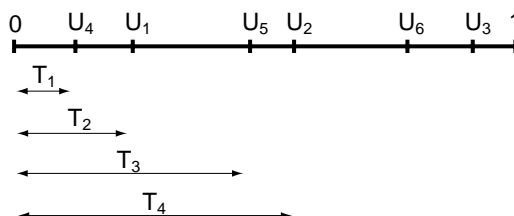
$$\frac{f(x_1)}{H'(x_1)} + \frac{f(x_2)}{H'(x_2)} + \dots$$

Of course we should reexpress each  $x_i$  as a function of  $y$ , to get the density in a usable form. □

I recommend that you remember the method used in the Example, rather than trying to memorize the result for various special cases. In each particular application, rederive. That way, you will be less likely to miss multiple contributions to a density.

EXAMPLE 29: ORDER STATISTICS FROM THE UNIFORM DISTRIBUTION

Suppose  $U_1, U_2, \dots, U_n$  are independent random variables, each with a Uniform[0, 1] distribution. They define  $n$  points in the unit interval. If we measure the distance of each point from 0 we obtain random variables  $0 \leq T_1 < T_2 < \dots < T_n$ , the values  $U_1, \dots, U_n$  rearranged into increasing order. For  $n = 6$ , the picture (with  $T_5$  and  $T_6$  not shown) looks like:



If we repeated the process by generating a new sample of  $U_i$ 's, we would probably not have  $U_4$  as the smallest,  $U_1$  as the second smallest, and so on. That is,  $T_1$  might correspond to a different  $U_i$ .

**Problem:** Find the distribution of  $T_k$ , the  $k$ th smallest of the ordered values (also known as the  $k$ th **order statistic**). For a very short interval  $[t, t + \delta]$ , with  $0 < t < t + \delta < 1$  and  $\delta$  very small, we need to show that  $\mathbb{P}\{t \leq T_k \leq t + \delta\}$  is roughly proportional to  $\delta$ , then determine  $f(t)$ , the constant of proportionality.

Write  $N$  for the number of  $U_i$  points that land in  $[t, t + \delta]$ . To get  $t \leq T_k \leq t + \delta$  we must have  $N \geq 1$ . If  $N = 1$  then we must have exactly  $k - 1$  points in  $[0, t)$  to get  $t \leq T_k \leq t + \delta$ . If  $N \geq 2$  then it becomes more complicated to describe all the ways that we would get  $t \leq T_k \leq t + \delta$ . Luckily for us, the contributions from all those complicated expressions will turn out to be small enough to ignore if  $\delta$  is small. Calculate.

$$\begin{aligned} \mathbb{P}\{t \leq T_k \leq t + \delta\} &= \mathbb{P}\{N = 1 \text{ and exactly } k - 1 \text{ points in } [0, t)\} \\ &\quad + \mathbb{P}\{N \geq 2 \text{ and } t \leq T_k \leq t + \delta\}. \end{aligned}$$

Let me first dispose of the second contribution, where  $N \geq 2$ . The indicator function of the event

$$F_2 = \{N \geq 2\} \cap \{t \leq T_k \leq t + \delta\}$$

is less than the sum of indicator functions

$$\sum_{1 \leq i < j \leq n} \mathbb{I}\{U_i, U_j \text{ both in } [t, t + \delta]\}$$

You should check this assertion by verifying that the sum of indicators is nonnegative and that it takes a value  $\geq 1$  if the event  $F_2$  occurs. Take expectations, remembering that the probability of an event is equal to the expectation of its indicator function, to deduce that

$$\mathbb{P}F_2 \leq \sum_{1 \leq i < j \leq n} \mathbb{P}\{U_i, U_j \text{ both in } [t, t + \delta]\}.$$

By symmetry, all  $\binom{n}{2}$  terms in the sum are equal to

$$\begin{aligned} &\mathbb{P}\{U_1, U_2 \text{ both in } [t, t + \delta]\} \\ &= \mathbb{P}\{t \leq U_1 \leq t + \delta\} \mathbb{P}\{t \leq U_2 \leq t + \delta\} \quad \text{by independence} \\ &= \delta^2. \end{aligned}$$

Thus  $\mathbb{P}F_2 \leq \binom{n}{2} \delta^2$ , which tends to zero much faster than  $\delta$  as  $\delta \rightarrow 0$ . (The value of  $n$  stays fixed throughout the calculation.)

Next consider the contribution from the event

$$F_1 = \{N = 1\} \cap \{\text{exactly } k - 1 \text{ points in } [0, t)\}.$$

Break into disjoint pieces like

$$\{U_1, \dots, U_{k-1} \text{ in } [0, t), U_k \text{ in } [t, t + \delta], U_{k+1}, \dots, U_n \text{ in } (t + \delta, 1]\}.$$

Again by virtue of the independence between the  $\{U_i\}$ , this piece has probability

$$\mathbb{P}\{U_1 < t\} \mathbb{P}\{U_2 < t\} \dots \mathbb{P}\{U_{k-1} < t\} \mathbb{P}\{U_k \text{ in } [t, t + \delta]\} \mathbb{P}\{U_{k+1} > t + \delta\} \dots \mathbb{P}\{U_n > t + \delta\},$$

Invoke the defining property of the uniform distribution to factorize the probability as

$$t^{k-1} \delta (1 - t - \delta)^{n-k} = t^{k-1} (1 - t)^{n-k} \delta + \text{terms of order } \delta^2 \text{ or smaller.}$$

How many such pieces are there? There are  $\binom{n}{k-1}$  ways to choose the  $k-1$  of the  $U_i$ 's to land in  $[0, t)$ , and for each of these ways there are  $n - k + 1$  ways to choose the single observation to land in  $[t, t + \delta]$ . The remaining observations must go in  $(t + \delta, 1]$ . We must add up

$$\binom{n}{k-1} \times (n - k + 1) = \frac{n!}{(k-1)!(n-k)!}$$

pieces with the same probability to calculate  $\mathbb{P}F_1$ .

Consolidating all the small contributions from  $\mathbb{P}F_1$  and  $\mathbb{P}F_2$  we then get

$$\mathbb{P}\{t \leq T_k \leq t + \delta\} = \frac{n!}{(k-1)!(n-k)!} t^{k-1} (1 - t)^{n-k} \delta + \text{terms of order } \delta^2 \text{ or smaller.}$$

That is, the distribution of  $T_k$  is continuous with density function

$$f(t) = \frac{n!}{(k-1)!(n-k)!} t^{k-1} (1 - t)^{n-k} \quad \text{for } 0 < t < 1.$$

Outside  $(0, 1)$  the density is zero. □

REMARK. Remember that it makes no difference how we define  $f(t)$  at  $t = 0$  and  $t = 1$ , because it can have no effect on integrals  $\int_a^b f(t) dt$ .

From the fact that the density must integrate to 1, we get

$$1 = \int_{-\infty}^0 0 dt + \frac{n!}{(k-1)!(n-k)!} \int_0^1 t^{k-1} (1 - t)^{n-k} dt + \int_1^{\infty} 0 dt$$

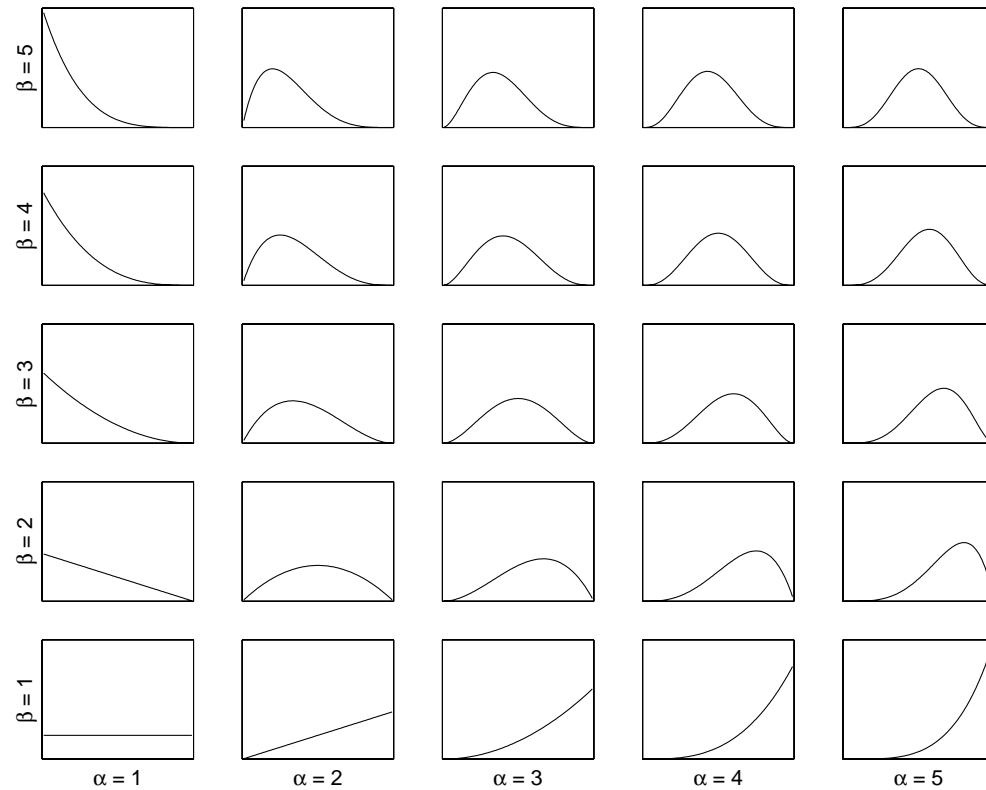
That is,

$$\int_0^1 t^{k-1} (1 - t)^{n-k} dt = \frac{(k-1)!(n-k)!}{n!},$$

a fact that you might try to prove by direct calculation.

# EXAMPLE 30: THE BETA DISTRIBUTION

Beta densities:  $t^{\alpha-1} (1-t)^{\beta-1} / B(\alpha, \beta)$  for  $0 < t < 1$  and vertical range (0,5)



The function *beta* in Matlab calculates the beta function, defined for  $z > 0$  and  $w > 0$  by

$$\text{beta}(z, w) = \int_0^1 t^{z-1} (1-t)^{w-1} dt.$$

The function *betainc* in Matlab calculates the incomplete beta function, defined by

$$\text{betainc}(x, z, w) = \int_0^x \frac{t^{z-1} (1-t)^{w-1}}{\text{beta}(z, w)} dt \quad \text{for } 0 \leq x \leq 1.$$

I used the function *beta* to draw the pictures of the density functions for various values of the parameters. See the Matlab m-file drawbeta.m for the calculations. □

### EXAMPLE 31: BINOMIAL TAIL PROBABILITIES

In principle it is easy to calculate probabilities such as  $\mathbb{P}\{\text{Bin}(30, p) \geq 17\}$  for various values of  $p$ : one has only to sum the series

$$\binom{30}{17} p^{17} (1-p)^{13} + \binom{30}{18} p^{18} (1-p)^{12} + \dots + (1-p)^{30}$$

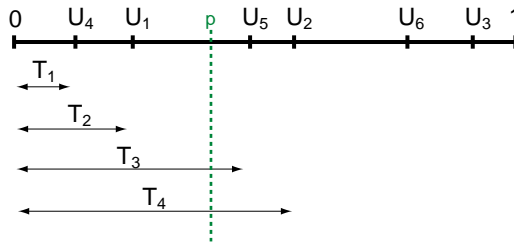
With a computer (compare with the Matlab m-file `BinProbs.m`) such a task would not be as arduous as it used to be back in the days of hand calculation. However, there is a neater method, based on the facts about the order statistics, which relates the Binomial and beta tail probabilities.

The relationship becomes clear from a special method for simulating coin tosses. For a fixed  $n$  (such as  $n = 30$ ), generate independently  $n$  random variables  $U_1, \dots, U_n$ , each distributed uniformly on  $[0, 1]$ . Fix a  $p$  in  $[0, 1]$ . Then the independent events

$$\{U_1 \leq p\}, \{U_2 \leq p\}, \dots, \{U_n \leq p\}$$

are like  $n$  independent flips of a coin that lands heads with probability  $p$ . The number,  $X_n$ , of such events that occur has a  $\text{Bin}(n, p)$  distribution.

As in Example 29, write  $T_k$  for the  $k$ th smallest value when the  $U_i$ 's are sorted into increasing order.



The random variables  $X_n$  and  $T_k$  are related by an equivalence,

$$X_n \geq k \text{ if and only if } T_k \leq p.$$

That is, there are  $k$  or more of the  $U_i$ 's in  $[0, p]$  if and only if the  $k$ th smallest of them is in  $[0, p]$ . Thus

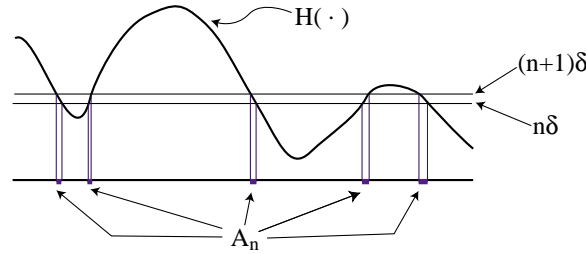
$$\mathbb{P}\{X_n \geq k\} = \mathbb{P}\{T_k \leq p\} = \frac{n!}{(k-1)!(n-k)!} \int_p^1 t^{k-1} (1-t)^{n-k} dt.$$

I know of at least one statistical package that calculates Binomial tail probabilities by means of this representation. □



EXAMPLE 32: EXPECTATION OF A RANDOM VARIABLE WITH A CONTINUOUS DISTRIBUTION

Consider a random variable  $X$  whose distribution has density function  $f(\cdot)$ . Let  $Y = H(X)$  be a new random variable, defined as a function of  $X$ . We calculate  $\mathbb{E}Y$  by an approximation argument similar to the one used in Example 28.



Cut the range of values taken by  $Y$  into disjoint intervals of the form  $n\delta \leq y < (n+1)\delta$ , for a small, positive  $\delta$ . Write  $A_n$  for the set of  $X$  values corresponding to this range, that is,

$$\{n\delta \leq Y < (n+1)\delta\} = \{X \in A_n\}.$$

When  $X$  lies in  $A_n$ , the value of  $Y$  must lie somewhere between  $n\delta$  and  $(n+1)\delta$ . Thus

$$\mathbb{E}(Y \mid X \in A_n) \approx n\delta,$$

with an error of approximation smaller than  $\delta$ . We could also approximate  $\mathbb{P}\{X \in A_n\}$  by a sum, as in Example 28, but it is better just to leave it in the form of an integral of the density  $f$  of the corresponding range.

$$\mathbb{P}\{X \in A_n\} = \int_{A_n} f(x) dx.$$

From rule E4 for expectations,

$$\begin{aligned} \mathbb{E}Y &= \sum_n \mathbb{P}\{X \in A_n\} \mathbb{E}(Y \mid X \in A_n) \\ &\approx \sum_n \int_{A_n} (n\delta) f(x) dx \\ &\approx \sum_n \int_{A_n} H(x) f(x) dx \quad \text{because } H(x) \approx n\delta \text{ when } x \in A_n \\ &= \int_{-\infty}^{\infty} H(x) f(x) dx. \end{aligned}$$

As  $\delta$  is made smaller and smaller, the error in the approximations tends to zero. In the limit we are left with the desired formula,

$$\mathbb{E}Y = \mathbb{E}H(X) = \int_{-\infty}^{\infty} H(x) f(x) dx \quad \text{where } X \text{ has density } f.$$

Compare with the formula for a random variable  $X^*$  taking only a discrete set of values  $x_1, x_2, \dots$ ,

$$\mathbb{E}H(X^*) = \sum_i H(x_i) \mathbb{P}\{X^* = x_i\}$$

In the passage from discrete to continuous distributions, discrete probabilities get replaced by densities and sums get replaced by integrals.  $\square$

You should be very careful not to confuse the formulae for expectations in the discrete and continuous cases. Think again if you find yourself integrating probabilities or summing expressions involving probability densities.

## Chapter 7

---

# Normal distribution

In 1733, Abraham de Moivre presented an approximation to the Binomial distribution. He later appended the derivation of his approximation to the solution of a problem asking for the calculation of an expected value for a particular game. He posed the rhetorical question (see Appendix A7 for a more extensive quotation) of how we might show that experimental proportions should be close to their expected values.

*In answer to this, I'll take the liberty to say, that this is the hardest Problem that can be proposed on the Subject of Chance, for which reason I have reserved it for the last, but I hope to be forgiven if my Solution is not fitted to the capacity of all Readers; however I shall derive from it some Conclusions that may be of use to every body: in order thereto, I shall here translate a Paper of mine which was printed November 12, 1733, and communicated to some Friends, but never yet made public, reserving to myself the right of enlarging my own Thoughts, as occasion shall require.*

*Novemb. 12, 1733*

*A Method of approximating the Sum of the Terms of the Binomial  $\overline{a+b}^n$  expanded into a Series, from whence are deduced some practical Rules to estimate the Degree of Assent which is to be given to Experiments.*

*Altho' the Solution of problems of Chance often requires that several Terms of the Binomial  $\overline{a+b}^n$  be added together, nevertheless in very high Powers the thing appears so laborious, and of so great difficulty, that few people have undertaken that Task; for besides *James* and *Nicolas Bernouilli*, two great Mathematicians, I know of no body that has attempted it; in which, tho' they have shown very great skill, and have the praise that is due to their Industry, yet some things were further required; for what they have done is not so much an Approximation as the determining very wide limits, within which they demonstrated that the Sum of the Terms was contained. Now the method ...*

A. De Moivre, *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play*, 3rd edition (1756). (Photographic reprint of final edition by Chelsea Publishing Company, 1967. The 1733 paper on the normal approximation is included as pages 243–259.)

This Chapter will explain de Moivre's approximation, then describe its modern counterpart, the so-called **central limit theorem**, which is used to justify a huge array of normal approximations.

What does a Binomial look like? Recall that Tchebychev's inequality suggests it should be clustered around the expected value, with a spread determined by the standard deviation.

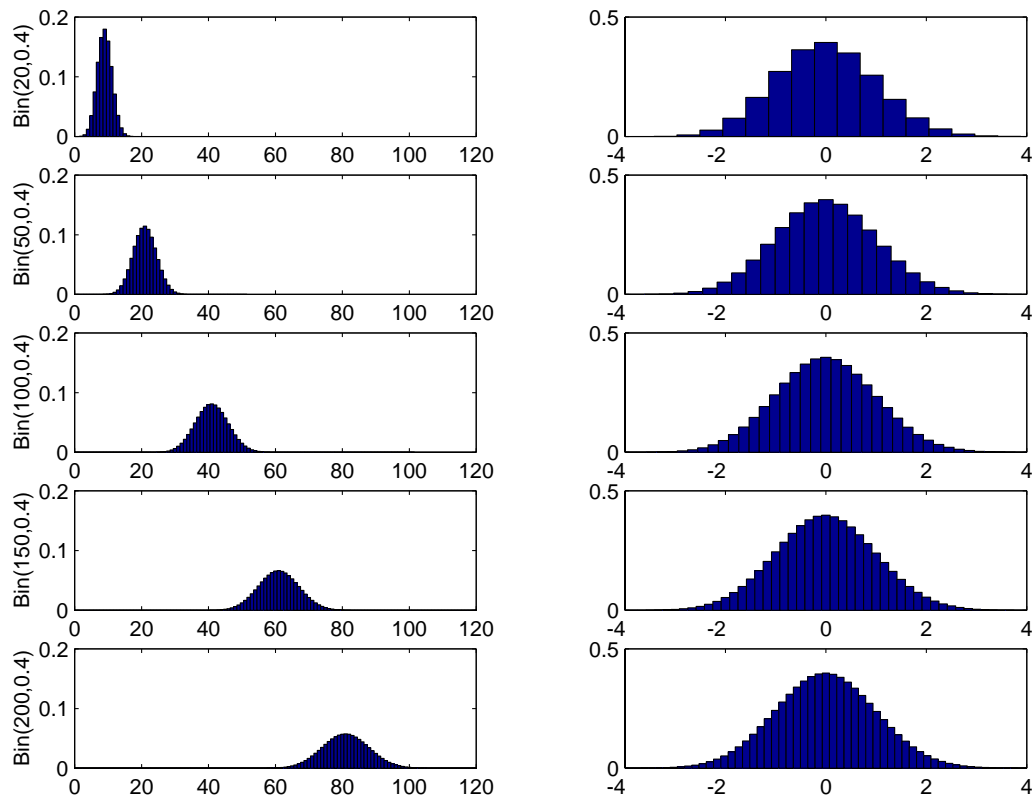
**Example 33:** A random variable  $X$  with a  $\text{Bin}(n, p)$  distribution has  $\mathbb{E}X = np$  and  $\text{var}(X) = np(1 - p)$ .

Also, from Problem sheet 5, we know that if  $X$  has a  $\text{Bin}(n, p)$  distribution, then the probabilities

$$b_n(k) = \mathbb{P}\{X = k\} = \binom{n}{k} p^k q^{n-k} \quad \text{for } k = 0, 1, \dots, n, \text{ where } q = 1 - p,$$

achieve their maximum value at a value  $k_{\max}$  close to  $np$ , the expected value. Moreover, the values of  $b(k)$  are increasing for  $k < k_{\max}$ , and decreasing for  $k > k_{\max}$ . The pictures on the left-hand side of the next display, for the  $\text{Bin}(n, 0.4)$  distribution with  $n = 20, 50, 100, 150, 200$ , exhibit this behavior. Each plot shows bars of height  $b_n(k)$  and width 1, centered at  $k$ . The maxima occur near  $n \times 0.4$  for each picture. As  $n$  increases, the spread also increases, reflecting the increase in the standard deviations  $\sigma_n = \sqrt{npq}$  for  $p = 0.4$ .

The gross effect of the increasing expected value and standard deviation is removed from the pictures on the right-hand side of the plot, where a bar of height  $\sigma_n \times b_n(k)$  now has width  $1/\sigma_n$  and is centered at  $(k - np)/\sigma_n$ , again with  $p = 0.4$ . The pictures now highlight the common shape of the distributions. The shaded region still has area 1.



Notice how the plots on the right settle down to a symmetric ‘bell-shaped’ curve. The shape of the “standardized” Binomial quickly stabilizes as  $n$  increases.

De Moivre expressed this stability by showing that

$$\mathbb{P}\{X = k_{\max} + m\} \approx b(k_{\max}) \exp\left(-\frac{m^2}{2npq}\right).$$

(Here, and subsequently, I translate de Moivre's results into modern notation.)

---

**Example 34: Derivation of de Moivre's approximation.**

---

Using the fact that the probabilities sum to 1, he was also able to show for  $p = 1/2$  that the  $b(k_{\max})$  should decrease like  $2/(B\sqrt{n})$ , for a constant  $B$  that he was initially only able to express as an infinite sum. Referring to his calculation of the ratio of the maximum term in the expansion of  $(1 + 1)^n$  to the sum,  $2^n$ , he wrote (page 244 of the *Doctrine of Chances*):

When I first began that inquiry, I contented myself to determine at large the Value of  $B$ , which was done by the addition of some Terms of the above-written Series; but as I perceived that it converged but slowly, and seeing at the same time that what I had done answered my purpose tolerably well, I desisted from proceeding further till my worthy and learned Friend Mr. James Stirling, who had applied himself after me to that inquiry, found that the Quantity  $B$  did denote the Square-root of the Circumference of a Circle whose Radius is Unity, so that if that Circumference be called  $c$ , the Ratio of the middle Term to the Sum of all the Terms will be expressed by  $\frac{2}{\sqrt{nc}}$ .

With Stirling's refinement, the approximation becomes, for general  $p$ ,

$$\mathbb{P}\{X = k_{\max} + m\} \approx \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{m^2}{2npq}\right),$$

or, substituting  $np$  for  $k_{\max}$  and writing  $k$  for  $k_{\max} + m$ ,

$$\mathbb{P}\{X = k\} \approx \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{(k - np)^2}{2npq}\right).$$

That is,  $\mathbb{P}\{X = k\}$  is approximately equal to the area under the smooth curve

$$f(x) = \frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{(x - np)^2}{2npq}\right),$$

for the interval  $k - 1/2 \leq x \leq k + 1/2$ . (The length of the interval is 1, so it does not appear in the previous display.) Effectively, we have replaced a calculation for the discrete Binomial distribution by a calculation for an approximating continuous distribution.

**Definition.** A random variable is said to have a **normal distribution** with parameters  $\mu$  and  $\sigma$  if it has a continuous distribution with density

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \quad \text{for } -\infty < x < \infty.$$

The normal distribution is denoted by  $N(\mu, \sigma^2)$ . The parameter  $\sigma$  must be positive, otherwise the density would not be positive. The parameter  $\mu$  can be any real value.

The special case where  $\mu = 0$  and  $\sigma = 1$  is called the **standard normal**. The density function for this  $N(0, 1)$  distribution is usually denoted by the special letter  $\phi$ ,

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{for } -\infty < x < \infty.$$

For this function to be a well defined density it must integrate to 1, that is,

$$\int_{-\infty}^{\infty} e^{-x^2/2} dx = \sqrt{2\pi},$$

a far from obvious result. (See the Appendix for a direct way to establish this equality.)

De Moivre's result asserts that the  $\text{Bin}(n, p)$  distribution is well approximated by the  $N(np, npq)$ . The next Example explains why the parameters of the approximating normal are taken as the expected value and variance of the Binomial.

---

**Example 35:** The  $N(\mu, \sigma^2)$  distribution has expected value  $\mu$  and variance  $\sigma^2$ .

---

For many purposes it is easier to think in terms of standardized random variables and their distributions. De Moivre's result tells us that a Binomially distributed  $X$  can be approximated by a  $N(\mu, \sigma^2)$ , where  $\mu = np = \mathbb{E}X$  and  $\sigma^2 = npq = \text{var}(X)$ . Equivalently, the distribution of the random variable  $(X - np)/\sqrt{npq}$  is approximated by the corresponding standardized normal, which turns out to have a  $N(0, 1)$  distribution.

---

**Example 36:** If  $Y$  has  $N(\mu, \sigma^2)$  distribution then the standardized random variable  $(Y - \mu)/\sigma$  has a standard normal distribution.

---

How does one actually perform a normal approximation? Back in the olden days, one would interpolate from tables found in most statistics texts. For example, if  $X$  has a  $\text{Bin}(100, 1/2)$  distribution,

$$\mathbb{P}\{45 \leq X \leq 55\} = \mathbb{P}\left\{\frac{45 - 50}{5} \leq \frac{X - 50}{5} \leq \frac{55 - 50}{5}\right\} \approx \mathbb{P}\{-1 \leq Z \leq +1\}$$

where  $Z$  has a standard normal distribution. From the tables one finds,  $\mathbb{P}\{Z \leq 1\} \approx .8413$ ; and by symmetry (draw a picture) of the  $N(0, 1)$  density,  $\mathbb{P}\{Z \geq -1\} \approx .8413$ , so that  $\mathbb{P}\{Z \leq -1\} \approx 1 - .8413$ . By subtraction,

$$\mathbb{P}\{-1 \leq Z \leq +1\} = \mathbb{P}\{Z \leq 1\} - \mathbb{P}\{Z \leq -1\} \approx .6826$$

That is, by the normal approximation,

$$\mathbb{P}\{45 \leq X \leq 55\} \approx .68$$

More concretely, there is about a 68% chance that 100 tosses of a fair coin will give somewhere between 45 and 55 heads.

It is possible to be more careful about the atoms of probability at 45 and 55 to improve the approximation, but the refinement is usually not vital.

These days, many computer packages will calculate areas under the normal density curve directly. However one must be careful to read the fine print about exactly which curve and which area is used.

### The central limit theorem

The normal approximation to the binomial is just one example of a general phenomenon corresponding to the mathematical result known as the **central limit theorem**. Roughly stated, the theorem asserts:

*If  $X$  can be written as a sum of a large number of relatively small, independent random variables, then it has approximately a  $N(\mu, \sigma^2)$  distribution, where  $\mu = \mathbb{E}X$  and  $\sigma^2 = \text{var}(X)$ . Equivalently, the standardized variable  $(X - \mu)/\sigma$  has approximately a standard normal distribution.*

The normal distribution has many agreeable properties that make it easy to work with. Many statistical procedures have been developed under normality assumptions, with occasional obeisance toward the central limit theorem. Modern theory has been much concerned with possible harmful effects of unwarranted assumptions such as normality. The modern fix

often substitutes huge amounts of computing for neat, closed-form, analytic expressions; but normality still lurks behind some of the modern data analytic tools.

---

**Example 37: A hidden normal approximation—the boxplot**

---

The normal approximation is heavily used to give an estimate of variability for the results from sampling.

---

**Example 38: Normal approximations for sample means**

---

**Things to remember**

- A random variable is said to have a **normal distribution** with parameters  $\mu$  and  $\sigma$  if it has a continuous distribution with density

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for } -\infty < x < \infty$$

The normal distribution is denoted by  $N(\mu, \sigma^2)$ . The parameter  $\sigma$  must be positive, otherwise the density would not be positive. The parameter  $\mu$  can be any real value.

- If  $Y$  has  $N(\mu, \sigma^2)$  distribution then  $\mathbb{E}X = \mu$  and  $\text{var}(X) = \sigma^2$ . The standardized random variable  $(Y - \mu)/\sigma$  has a standard normal distribution,  $N(0, 1)$ .
- If  $X$  can be written as a sum of a large number of relatively small, independent random variables, then it has approximately a  $N(\mu, \sigma^2)$  distribution, where  $\mu = \mathbb{E}X$  and  $\sigma^2 = \text{var}(X)$ . Equivalently, the standardized variable  $(X - \mu)/\sigma$  has approximately a standard normal distribution.

EXAMPLE 33: EXPECTED VALUE AND VARIANCE OF THE BINOMIAL DISTRIBUTION

Remember that the  $\text{Bin}(n, p)$  comes from counting the number of heads in  $n$  independent tosses of a coin that lands heads with probability  $p$ . We can write the total number of heads as  $X = \xi_1 + \xi_2 + \dots + \xi_n$ , where  $\xi_i$  takes the value 1 if the  $i$ th toss lands heads, and 0 otherwise.

It is easy to calculate the expected value and variance for each  $\xi_i$ .

$$\mathbb{E}\xi_i = (0 \times \mathbb{P}\{\xi_i = 0\}) + (1 \times \mathbb{P}\{\xi_i = 1\}) = p,$$

and

$$\begin{aligned} \text{var}(\xi_i) &= \mathbb{E}(\xi_i - p)^2 \\ &= ((0 - p)^2 \times \mathbb{P}\{\xi_i = 0\}) + ((1 - p)^2 \times \mathbb{P}\{\xi_i = 1\}) \\ &= p^2(1 - p) + (1 - p)^2 p \\ &= p(1 - p) \end{aligned}$$

The rule for expectation of a sum gives

$$\mathbb{E}X = \mathbb{E}\xi_1 + \dots + \mathbb{E}\xi_n = np.$$

The rule for calculating the variance of a sum simplifies because all the covariance terms equal zero (by independence of the  $\xi_i$ 's), leaving

$$\text{var}(X) = \text{var}(\xi_1) + \dots + \text{var}(\xi_n) = np(1 - p).$$

□

EXAMPLE 34: NORMAL APPROXIMATION TO THE BINOMIAL

The normal approximation to the binomial is largely explained by two simple facts:

$$\begin{aligned}\log(1+x) &\approx x && \text{for } x \text{ near } 0, \\ 1+2+3+\dots+m &= \frac{1}{2}m(m+1) \approx \frac{1}{2}m^2.\end{aligned}$$

From Problem sheet 5, you know that

$$\frac{b(k)}{b(k-1)} = \frac{(n-k+1)p}{kq},$$

which helped you to show that the value  $k_{\max}$  maximizing  $b(k)$  is close to  $np$ . The ratio takes a simpler form if we replace  $k$  by  $k_{\max} + i$ .

$$\frac{b(k_{\max} + i)}{b(k_{\max} + i - 1)} = \frac{(n - k_{\max} - i + 1)p}{(k_{\max} + i)q} \approx \frac{(nq - i)p}{(np + i)q} = \frac{1 - i/(nq)}{1 + i/(np)}.$$

The logarithm of the last ratio equals

$$\log\left(1 - \frac{i}{nq}\right) - \log\left(1 + \frac{i}{np}\right) \approx -\frac{i}{nq} - \frac{i}{np} = -\frac{i}{npq}.$$

By summing such terms we get an approximation for the logarithm of the ratio of the individual Binomial probabilities to their largest term. For example, if  $m \geq 1$  and  $k_{\max} + m \leq n$ ,

$$\begin{aligned}\log \frac{b(k_{\max} + m)}{b(k_{\max})} &= \log \left( \frac{b(k_{\max} + 1)}{b(k_{\max})} \times \frac{b(k_{\max} + 2)}{b(k_{\max} + 1)} \times \dots \times \frac{b(k_{\max} + m)}{b(k_{\max} + m - 1)} \right) \\ &= \log \frac{b(k_{\max} + 1)}{b(k_{\max})} + \log \frac{b(k_{\max} + 2)}{b(k_{\max} + 1)} + \dots + \log \frac{b(k_{\max} + m)}{b(k_{\max} + m - 1)} \\ &\approx \frac{-1 - 2 - \dots - m}{npq} \\ &\approx -\frac{1}{2} \frac{m^2}{npq}.\end{aligned}$$

Thus

$$\mathbb{P}\{X = k_{\max} + m\} \approx b(k_{\max}) \exp\left(-\frac{m^2}{2npq}\right) \quad \text{for } m \text{ not too large.}$$

An analogous approximation holds for  $0 \leq k_{\max} + m \leq k_{\max}$ . □



EXAMPLE 35: THE  $N(\mu, \sigma^2)$  DISTRIBUTION HAS EXPECTED VALUE  $\mu$  AND VARIANCE  $\sigma^2$ .

The  $N(\mu, \sigma^2)$  is a continuous distribution with density

$$f_{\mu, \sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad \text{for } -\infty < x < \infty.$$

If  $X$  has this distribution then, from Example 32,

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_{\mu, \sigma}(x) dx.$$

Make the change of variable  $y = (x - \mu)/\sigma$  to rewrite the integral as

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma y) \exp(-y^2/2) \sigma dy = \mu \int_{-\infty}^{\infty} \phi(y) dy + \sigma \int_{-\infty}^{\infty} y \phi(y) dy.$$

We know (from the fact that  $\phi$  is a density function) that the coefficient of  $\mu$  equals 1. Antisymmetry of  $y\phi(y)$  makes it integrate to 0. Thus

$$\mathbb{E}X = \mu \quad \text{when } X \text{ has a } N(\mu, \sigma^2) \text{ distribution.}$$

A similar appeal to Example 32, followed by the same change of variable gives

$$\text{var}(X) = \mathbb{E}(X - \mu)^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f_{\mu, \sigma}(x) dx = \sigma^2 \int_{-\infty}^{\infty} y^2 \phi(y) dy.$$

An integration-by-parts, using the fact that  $d\phi(y)/dy = -y\phi(y)$ , simplifies the integral,

$$\int_{-\infty}^{\infty} -y \frac{d\phi(y)}{dy} dy = [-y\phi(y)]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \phi(y) dy = 1.$$

Thus

$$\text{var}(X) = \sigma^2 \quad \text{when } X \text{ has a } N(\mu, \sigma^2) \text{ distribution.}$$

□

EXAMPLE 36: STANDARDIZATION OF THE NORMAL DISTRIBUTION

If  $Y$  has a  $N(\mu, \sigma^2)$  distribution, the standardized random variable  $Z = (Y - \mu)/\sigma$  has

$$\begin{aligned}\mathbb{E}Z &= \frac{1}{\sigma} (\mathbb{E}X - \mu) = 0, \\ \text{var}(Z) &= \frac{1}{\sigma^2} \text{var}(X - \mu) = 1.\end{aligned}$$

In fact, even more is true. Calculate the density for  $Z$ . For small, positive  $\delta$ ,

$$\begin{aligned}\mathbb{P}\{z \leq Z \leq z + \delta\} &= \mathbb{P}\{\mu + \sigma z \leq X \leq \mu + \sigma(z + \delta)\} \\ &\approx (\sigma\delta) f_{\mu, \sigma}(\mu + \sigma z) \quad \text{density for } X \\ &= \sigma\delta \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\mu + \sigma z - \mu)^2}{2\sigma^2}\right) \\ &= \delta\phi(z).\end{aligned}$$

That is,  $Z$  has a continuous distribution with density  $\phi$ ; it has a  $N(0, 1)$  distribution.

Put another way, we can construct an  $Y$  with a  $N(\mu, \sigma^2)$  by defining it as  $\mu + \sigma Z$ , where  $Z$  has a standard normal distribution.  $\square$

### EXAMPLE 37: THE BOX PLOT

The boxplot provides a convenient way of summarizing data (such as grades in Statistics 241). The method is:

(i) arrange the data in increasing order

(ii) find the split points

LQ = lower quartile: 25% of the data smaller than LQ

M = median: 50% of the data smaller than M

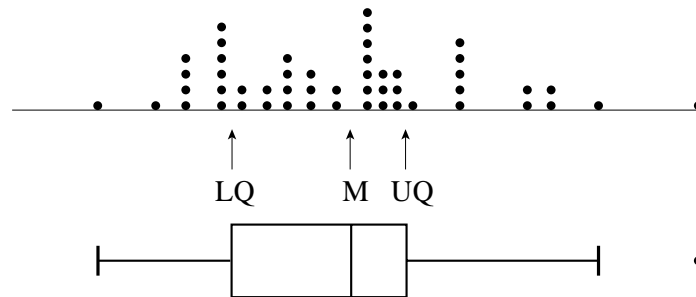
UQ = upper quartile: 75% of the data smaller than UQ

(iii) calculate IQR (= inter-quartile range) = UQ – LQ

(iv) draw a box with ends at LQ and UQ, and a dot or a line at M

(v) draw whiskers out to  $UQ + 1.5 \times IQR$  and  $LQ - 1.5 \times IQR$ , but then trim them back to the most extreme data point in those ranges

(vi) draw dots for each individual data point outside the box and whiskers (There are various ways to deal with cases where the number of observations is not a multiple of four, or where there are ties, or . . .)



Where does the  $1.5 \times IQR$  come from? Consider  $n$  independent observations from a  $N(\mu, \sigma^2)$  distribution. The proportion of observations smaller than any fixed  $x$  should be approximately equal to  $\mathbb{P}\{W \leq x\}$ , where  $W$  has a  $N(\mu, \sigma^2)$  distribution. From normal tables (or a computer),

$$\mathbb{P}\{W \leq \mu + .675\sigma\} \approx .75$$

$$\mathbb{P}\{W \leq \mu - .675\sigma\} \approx .25$$

and, of course,

$$\mathbb{P}\{W \leq \mu\} = .5$$

For the sample we should expect

$$LQ \approx \mu - .675\sigma$$

$$UQ \approx \mu + .675\sigma$$

$$M \approx \mu$$

and consequently,

$$IQR \approx 1.35\sigma$$

Check that  $0.675 + (1.5 \times 1.35) = 2.70$ . Before trimming, the whiskers should approximately reach to the ends of the range  $\mu \pm 2.70\sigma$ . From computer (or tables),

$$\mathbb{P}\{W \leq \mu - 2.70\sigma\} = \mathbb{P}\{W \geq \mu + 2.70\sigma\} = .003$$

Only about 0.6% of the sample should be out beyond the whiskers. □

# EXAMPLE 38: NORMAL APPROXIMATIONS IN SAMPLING

In Example 26 we found the expected value and variance of a sample mean  $\bar{Y}$  for a sample of size  $n$  from a population  $\{y_1, y_2, \dots, y_N\}$ :

$$\mathbb{E}\bar{Y} = \bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$$

and, for sampling with replacement,

$$\text{var}(\bar{Y}) = \frac{\sigma^2}{n} \quad \text{where } \sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \bar{y})^2.$$

If  $Z$  has a  $N(0, 1)$  distribution,

$$\mathbb{P}\{-1.96 \leq Z \leq 1.96\} \approx 0.95.$$

The standardized random variable  $(\bar{Y} - \bar{y})/\sqrt{\sigma^2/n}$  is well approximated by the  $N(0, 1)$ . Thus

$$\mathbb{P}\left\{-\frac{1.96\sigma}{\sqrt{n}} \leq \bar{Y} - \bar{y} \leq \frac{1.96\sigma}{\sqrt{n}}\right\} \approx 0.95.$$

Before we sample, we can assert that we have about a 95% chance of getting a value of  $\bar{Y}$  in the range  $\bar{y} \pm 1.96\sigma/\sqrt{n}$ . (For the post-sampling interpretation of the approximation, you should take Statistics 242/542.)

Of course, we would not know the value  $\sigma$ , so it must be estimated.

For sampling without replacement, the variance of the sample mean is multiplied by the correction factor  $(N - n)/(N - 1)$ . The sample mean is no longer an average of many *independent* summands, but the normal approximation can still be used. (The explanation would take us beyond 241/541.) □

## A PASSAGE FROM DE MOIVRE

## Corollary.

From this it follows, that if after taking a great number of Experiments, it should be perceived that the happenings and failings have been nearly in a certain proportion, such as of 2 to 1, it may safely be concluded that the Probabilities of happening or failing at any one time assigned will be very near in that proportion, and that the greater the number of Experiments has been, so much nearer the Truth will the conjectures be that are derived from them.

But suppose it should be said, that notwithstanding the reasonableness of building Conjectures upon Observations, still considering the great Power of Chance, Events might at long run fall out in a different proportion from the real Bent which they have to happen one way or the other; and that supposing for Instance that an Event might as easily happen as not happen, whether after three thousand Experiments it may not be possible it should have happened two thousand times and failed a thousand; and that therefore the Odds against so great a variation from Equality should be assigned, whereby the Mind would be the better disposed in the Conclusions derived from the Experiments.

In answer to this, I'll take the liberty to say, that this is the hardest Problem that can be proposed on the Subject of Chance, for which reason I have reserved it for the last, but I hope to be forgiven if my Solution is not fitted to the capacity of all Readers; however I shall derive from it some Conclusions that may be of use to every body: in order thereto, I shall here translate a Paper of mine which was printed November 12, 1733, and communicated to some Friends, but never yet made public, reserving to myself the right of enlarging my own Thoughts, as occasion shall require.

Novemb. 12, 1733

A Method of approximating the Sum of the Terms of the Binomial  $a + b^n$  expanded into a Series, from whence are deduced some practical Rules to estimate the Degree of Assent which is to be given to Experiments.

Altho' the Solution of problems of Chance often requires that several Terms of the Binomial  $a + b^n$  be added together, nevertheless in very high Powers the thing appears so laborious, and of so great difficulty, that few people have undertaken that Task; for besides James and Nicolas Bernouilli, two great Mathematicians, I know of no body that has attempted it; in which, tho' they have shown very great skill, and have the praise that is due to their Industry, yet some things were further required; for what they have done is not so much an Approximation as the determining very wide limits, within which they demonstrated that the Sum of the Terms was contained.

A. De Moivre, *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play*, 3rd edition (1756), pages 242–243. (Photographic reprint of final edition by Chelsea Publishing Company, 1967.)

THE MYSTERIOUS  $\sqrt{2\pi}$ 

Why is it that the constant  $C = \int_{-\infty}^{\infty} \exp(-x^2/2) dx$  is equal to  $\sqrt{2\pi}$ ? Equivalently, why is the constant  $C^2 = \iint \exp(-(x^2 + y^2)/2) dx dy$  equal to  $2\pi$ ? (Here, and subsequently, the double integral runs over the whole plane.)

Using the fact that

$$\int_0^{\infty} \mathbb{I}\{r \leq z\} e^{-z} dz = e^{-r} \quad \text{for } r > 0,$$

we may rewrite  $C^2$  as a triple integral: replace  $r$  by  $(x^2 + y^2)/2$ , then substitute into the double integral to get

$$C^2 = \iint \left( \int_0^{\infty} \mathbb{I}\{x^2 + y^2 \leq 2z\} dz \right) dx dy = \int_0^{\infty} \left( \iint \mathbb{I}\{x^2 + y^2 \leq 2z\} dx dy \right) dz.$$

With the change in the order of integration, the double integral is now calculating the area of a circle centered at the origin and with radius  $\sqrt{2z}$ . The triple integral reduces to

$$\int_0^{\infty} \pi \left( \sqrt{2z} \right)^2 e^{-z} dz = \int_0^{\infty} \pi 2z e^{-z} dz = 2\pi.$$

That is,  $C = \sqrt{2\pi}$ , as asserted.

## STIRLING'S FORMULA

For positive integers  $n$ , the formula asserts that

$$<A7.1> \quad n! \approx \sqrt{2\pi} n^{n+1/2} \exp(-n),$$

in the sense that the ratio of both sides tends to 1 as  $n$  goes to infinity.

As the first step towards a proof, write

$$\log n! = \log 1 + \log 2 + \dots + \log n$$

as a sum of integrals of indicator functions:

$$\log n! = \sum_{i=1}^n \int_1^n \mathbb{I}\{1 \leq x < i\} \frac{1}{x} dx = \int_1^n \sum_{i=1}^n \mathbb{I}\{1 \leq x < i\} \frac{1}{x} dx$$

The sum of indicator functions counts the number of integers in the range  $1, 2, \dots, n$  that are greater than  $x$ . It equals  $n - \lfloor x \rfloor$ , where  $\lfloor x \rfloor$  denotes the integer part of  $x$ . The difference  $\psi(x) = x - \lfloor x \rfloor$  lies in the range  $[0, 1)$ ; it gives the fractional part of  $x$ .

The integral representing  $\log(n!)$  is equal

$$\int_1^n \frac{n - \lfloor x \rfloor}{x} dx = \int_1^n \frac{n - x + \psi(x)}{x} dx = n \log n - n + \int_1^n \frac{\psi(x)}{x} dx.$$

The last integral diverges as  $n$  tends to infinity, because the contribution from the interval  $[i, i+1)$  equals

$$\int_i^{i+1} \frac{x-i}{x} dx = \int_0^1 \frac{t}{t+i} dt \approx \frac{1}{2i}.$$

For the approximation I have treated the  $t+i$  in the denominator as approximately equal to  $i$  and then noted that  $\int_0^1 t dt = 1/2$ . The sum of the contributions from the integral involving  $\psi$  increases like  $\frac{1}{2} \log n$ .

It seems we have to subtract off an extra  $\frac{1}{2} \log n = \frac{1}{2} \int_1^n \frac{1}{x} dx$  to keep the remainder term under control. Splitting the integral into contributions from intervals  $[i, i+1)$ , we then get

$$<A7.2> \quad \log(n!) - (n + 1/2) \log n - n = \sum_{i=1}^n \int_0^1 \frac{t - 1/2}{t+i} dt$$

With the subtraction of the  $1/2$  we will get some cancellation between the negative contribution for  $0 \leq t \leq 1/2$  and the positive contribution for  $1/2 < t \leq 1$ .

Make the change of variable  $s = 1/2 - t$  for the integral over  $[0, 1/2]$ , and the change of variable  $s = t - 1/2$  over  $(1/2, 1]$ .

$$\begin{aligned} \int_0^1 \frac{t - 1/2}{t+i} dt &= \int_0^{1/2} \frac{-s}{i + 1/2 - s} ds + \int_0^{1/2} \frac{s}{i + 1/2 + s} ds \\ &= -2 \int_0^{1/2} \frac{s^2}{(i + 1/2)^2 - s^2} ds. \end{aligned}$$

The last expression is bounded in absolute value by  $i^{-2}$ . The sum of the integrals forms a convergent series. That is, for some constant  $c$ ,

$$\int_1^n \frac{\psi(x) - 1/2}{x} dx \rightarrow c \quad \text{as } n \rightarrow \infty.$$

Equivalently, from  $<A7.2>$ ,

$$\frac{n!}{n^{n+1/2} e^{-n}} \rightarrow e^c \quad \text{as } n \rightarrow \infty$$

This result is equivalent to formula  $<A7.1>$ , except for the identification of  $e^c$  as the constant  $\sqrt{2\pi}$ . See the discussion on the next page for a way of deriving the value of the constant.

For an alternative derivation of Stirling's formula, see Feller I, pages 52–53.

## THE MAXIMUM OF THE BINOMIAL PROBABILITIES

We know that the probability  $b(k) = \binom{n}{k} p^k q^{n-k}$  achieves its maximum at a value  $K$  close to  $np$ . Temporarily write  $L$  for  $n - K \approx nq$ . Use Stirling's formula to approximate the value at the maximum,

$$b(K) = \frac{n!}{L!K!} p^K q^L \approx \frac{C n^{n+1/2} e^{-n} p^K q^L}{C L^{L+1/2} e^{-L} C K^{K+1/2} e^{-K}} \approx \frac{n^{n+1/2} p^{np} q^{nq}}{C (np)^{np+1/2} (nq)^{nq+1/2}},$$

where  $C$  denotes the constant  $\sqrt{2\pi}$ . The actual value is unimportant for the present calculation. After some very satisfying cancellations, we are left with

$$b(K) \approx \frac{1}{C \sqrt{npq}}.$$

When combined with the approximation from Example 34, this result gives an approximation for all the Binomial probabilities,

$$b(k) \approx \frac{1}{C \sqrt{npq}} \int_{k-1/2}^{k+1/2} \exp\left(-\frac{(x - np)^2}{2npq}\right) dx.$$

We know that the probabilities must sum to 1. Thus

$$C \approx \frac{1}{\sqrt{npq}} \sum_{k=0}^n \int_{k-1/2}^{k+1/2} \exp\left(-\frac{(x - np)^2}{2npq}\right) dx = \int_{-np-1/2}^{np+1/2} \exp(-y^2/2) dy \rightarrow \sqrt{2\pi}.$$

In the limit we recover  $C = \sqrt{2\pi}$ .



## Chapter 8

# Poisson approximations

The  $\text{Bin}(n, p)$  can be thought of as the distribution of a sum of independent indicator random variables  $X_1 + \dots + X_n$ , with  $\{X_i = 1\}$  denoting a head on the  $i$ th toss of a coin. The normal approximation to the Binomial works best when the variance  $np(1-p)$  is large, for then each of the standardized summands  $(X_i - p)/\sqrt{np(1-p)}$  makes a relatively small contribution to the standardized sum. When  $n$  is large but  $p$  is small, in such a way that  $np$  is not large, a different type of approximation to the Binomial is better.

**Definition.** A random variable  $Y$  is said to have a **Poisson distribution** with parameter  $\lambda$  if it can take values  $0, 1, 2, \dots$  with probabilities

$$\mathbb{P}\{Y = k\} = \frac{e^{-\lambda} \lambda^k}{k!} \quad \text{for } k = 0, 1, 2, \dots$$

The parameter  $\lambda$  must be positive. The distribution is denoted by  $\text{Poisson}(\lambda)$ .

---

### Example 39: $\text{Poisson}(np)$ approximation to the $\text{Binomial}(n, p)$

---

The Poisson inherits several properties from the Binomial. For example, the  $\text{Bin}(n, p)$  has expected value  $np$  and variance  $np(1-p)$ . One might suspect that the  $\text{Poisson}(\lambda)$  should therefore have expected value  $\lambda = n(\lambda/n)$  and variance  $\lambda = \lim_{n \rightarrow \infty} n(\lambda/n)(1 - \lambda/n)$ . Also, the coin-tossing origins of the Binomial show that if  $X$  has a  $\text{Bin}(m, p)$  distribution and  $X'$  has  $\text{Bin}(n, p)$  distribution independent of  $X$ , then  $X + X'$  has a  $\text{Bin}(n + m, p)$  distribution. Putting  $\lambda = mp$  and  $\mu = np$  one would then suspect that the sum of independent  $\text{Poisson}(\lambda)$  and  $\text{Poisson}(\mu)$  distributed random variables is  $\text{Poisson}(\lambda + \mu)$  distributed.

---

**Example 40:** If  $X$  has a  $\text{Poisson}(\lambda)$  distribution, then  $\mathbb{E}X = \text{var}(X) = \lambda$ . If also  $Y$  has a  $\text{Poisson}(\mu)$  distribution, and  $Y$  is independent of  $X$ , then  $X + Y$  has a  $\text{Poisson}(\lambda + \mu)$  distribution.

---

Counts of rare events—such as the number of atoms undergoing radioactive decay during a short period of time, or the number of aphids on a leaf—are often modelled by Poisson distributions, at least as a first approximation. In some situations it makes sense to think of the counts as the number of successes in a large number of independent trials, with the chance of a success on any particular trial being very small (“rare events”). In such a setting, the Poisson arises as an approximation for the Binomial. The Poisson approximation also applies in many settings where the trials are “almost independent” but not quite.

---

### Example 41: Poisson approximation for a matching problem

---

EXAMPLE 39: POISSON( $np$ ) APPROXIMATION TO BINOMIAL( $n, p$ )

The Poisson( $\lambda$ ) appears as an approximation to the Bin( $n, p$ ) when  $n$  is large,  $p$  is small, and  $\lambda = np$ :

$$\begin{aligned} \binom{n}{k} p^k (1-p)^{n-k} &= \frac{n(n-1)\dots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= 1 \times \left(1 - \frac{1}{n}\right) \times \dots \times \left(1 - \frac{k-1}{n}\right) \left(1 - \frac{\lambda}{n}\right)^{-k} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \\ &\approx \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \quad \text{if } k \text{ is small relative to } n \\ &\approx \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{if } n \text{ is large.} \end{aligned}$$

The final  $e^{-\lambda}$  comes from an approximation to the logarithm,

$$\log \left(1 - \frac{\lambda}{n}\right)^n = n \log \left(1 - \frac{\lambda}{n}\right) = n \left(-\frac{\lambda}{n} - \frac{1}{2} \frac{\lambda^2}{n^2} - \dots\right) \approx -\lambda \quad \text{if } \lambda/n \approx 0.$$

By careful consideration of the error terms, one can find explicit bounds for the error of approximation. For example, it can be shown<sup>1</sup> that if  $X$  is distributed Bin( $n, p$ ) and  $Y$  is distributed Poisson( $np$ ) then

$$\sum_{k=0}^{\infty} |\mathbb{P}\{X = k\} - \mathbb{P}\{Y = k\}| \leq 4p.$$

Clearly the Poisson is an excellent approximation when  $p$  is small. □

---

<sup>1</sup> Le Cam, page 187 of “On the distribution of sums of independent random variables”, in *Bernoulli, Bayes, Laplace: Anniversary Volume*, J. Neyman and L Le Cam, eds., Springer-Verlag 1965. The 1992 book by Barbour, Holst, and Janson (“Poisson Approximation,” Oxford University Press) contains an extensive discussion of similar results.

EXAMPLE 40: PROPERTIES OF THE POISSON DISTRIBUTION

Verify the properties of the Poisson distribution suggested by the Binomial analogy: If  $X$  has a  $\text{Poisson}(\lambda)$  distribution, show that

(i)  $\mathbb{E}X = \lambda$

(ii)  $\text{var}(X) = \lambda$

Also, if  $Y$  has a  $\text{Poisson}(\mu)$  distribution independent of  $X$ , show that

(iii)  $X + Y$  has a  $\text{Poisson}(\lambda + \mu)$  distribution

SOLUTION: Assertion (i) comes from a routine application of the formula for the expectation of a random variable with a discrete distribution.

$$\begin{aligned}\mathbb{E}X &= \sum_{k=0}^{\infty} k \mathbb{P}\{X = k\} = \sum_{k=1}^{\infty} k \frac{e^{-\lambda} \lambda^k}{k!} && \text{What happens to } k = 0? \\ &= e^{-\lambda} \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ &= e^{-\lambda} \lambda e^{\lambda} \\ &= \lambda.\end{aligned}$$

Notice how the  $k$  cancelled out one factor from the  $k!$  in the denominator.

If we were to calculate  $\mathbb{E}(X^2)$  in the same way, one factor in the  $k^2$  would cancel the leading  $k$  from the  $k!$ , but would leave an unpleasant  $k/(k-1)!$  in the sum. Too bad the  $k^2$  cannot be replaced by  $k(k-1)$ . Well, why not?

$$\begin{aligned}\mathbb{E}(X^2 - X) &= \sum_{k=0}^{\infty} k(k-1) \mathbb{P}\{X = k\} \\ &= e^{-\lambda} \sum_{k=2}^{\infty} k(k-1) \frac{\lambda^k}{k!} && \text{What happens to } k = 0 \text{ and } k = 1? \\ &= e^{-\lambda} \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} \\ &= \lambda^2.\end{aligned}$$

Now calculate the variance.

$$\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}(X^2 - X) + \mathbb{E}X - (\mathbb{E}X)^2 = \lambda.$$

For assertion (iii), first note that  $X + Y$  can take only values  $0, 1, 2, \dots$ . For a fixed  $k$  in this range, decompose the event  $\{X + Y = k\}$  into disjoint pieces whose probabilities can be simplified by means of the independence between  $X$  and  $Y$ .

$$\begin{aligned}\mathbb{P}\{X + Y = k\} &= \mathbb{P}\{X = 0, Y = k\} + \mathbb{P}\{X = 1, Y = k-1\} + \dots + \mathbb{P}\{X = k, Y = 0\} \\ &= \mathbb{P}\{X = 0\} \mathbb{P}\{Y = k\} + \mathbb{P}\{X = 1\} \mathbb{P}\{Y = k-1\} + \dots + \mathbb{P}\{X = k\} \mathbb{P}\{Y = 0\} \\ &= \frac{e^{-\lambda} \lambda^0}{0!} \frac{e^{-\mu} \mu^k}{k!} + \dots + \frac{e^{-\lambda} \lambda^k}{k!} \frac{e^{-\mu} \mu^0}{0!} \\ &= \frac{e^{-\lambda-\mu}}{k!} \left( \frac{k!}{0!k!} \lambda^0 \mu^k + \frac{k!}{1!(k-1)!} \lambda^1 \mu^{k-1} + \dots + \frac{k!}{k!0!} \lambda^k \mu^0 \right) \\ &= \frac{e^{-\lambda-\mu}}{k!} (\lambda + \mu)^k.\end{aligned}$$

The bracketed sum in the second last line is just the binomial expansion of  $(\lambda + \mu)^k$ . □

Question: How do you interpret the notation in the last calculation when  $k = 0$ ? I always feel slightly awkward about a contribution from  $k-1$  if  $k = 0$ .

EXAMPLE 41: POISSON APPROXIMATION WITH DEPENDENCE

Suppose  $N$  letters are placed at random into  $N$  envelopes, one letter per envelope. The total number of correct matches,  $X$ , can be written as a sum  $X_1 + \dots + X_N$  of indicators,

$$X_i = \begin{cases} 1 & \text{if letter } i \text{ is placed in envelope } i, \\ 0 & \text{otherwise.} \end{cases}$$

The  $X_i$  are dependent on each other. For example, symmetry implies that

$$\mathbb{P}\{X_i = 1\} = 1/N \quad \text{for each } i$$

and

$$\mathbb{P}\{X_i = 1 \mid X_1 = X_2 = \dots = X_{i-1} = 1\} = \frac{1}{N - i + 1}$$

We could eliminate the dependence by relaxing the requirement of only one letter per envelope. The number of letters placed in the correct envelope (possibly together with other, incorrect letters) would then have a  $\text{Bin}(N, 1/N)$  distribution, which approximates  $\text{Poisson}(1)$  if  $N$  is large.

We can get some supporting evidence for  $X$  having something close to a  $\text{Poisson}(1)$  distribution under the original assumption (one letter per envelope) by calculating some moments,

$$\mathbb{E}X = \sum_{i \leq N} \mathbb{E}X_i = N\mathbb{P}\{X_i = 1\} = 1$$

and

$$\begin{aligned} \mathbb{E}X^2 &= \mathbb{E}\left(X_1^2 + \dots + X_N^2 + 2 \sum_{i < j} X_i X_j\right) \\ &= N\mathbb{E}X_1^2 + 2\binom{N}{2}\mathbb{E}X_1 X_2 \quad \text{by symmetry} \\ &= N\mathbb{P}\{X_1 = 1\} + (N^2 - N)\mathbb{P}\{X_1 = 1, X_2 = 1\} \\ &= \left(N \times \frac{1}{N}\right) + (N^2 - N) \times \frac{1}{N(N-1)} \\ &= 2. \end{aligned}$$

Compare with Example 40, which gives  $\mathbb{E}Y = 1$  and  $\mathbb{E}Y^2 = 2$  for a  $Y$  distributed  $\text{Poisson}(1)$ . □

Using the **method of inclusion and exclusion**, it is possible<sup>2</sup> to calculate the exact distribution of the number of correct matches,

$$\mathbb{P}\{X = k\} = \frac{1}{k!} \left(1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \dots \pm \frac{1}{(N-k)!}\right).$$

For fixed  $k$ , as  $N \rightarrow \infty$  the probability converges to

$$\frac{1}{k!} \left(1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \dots\right) = \frac{e^{-1}}{k!},$$

which is the probability that  $Y = k$  if  $Y$  has a  $\text{Poisson}(1)$  distribution.

---

<sup>2</sup> Feller Vol 1, Chapter 4

## Chapter 9

# Poisson processes

The Binomial distribution and the geometric distribution describe the behavior of two random variables derived from the random mechanism that I have called coin tossing. The name *coin tossing* describes the whole mechanism; the names *Binomial* and *geometric* refer to particular aspects of that mechanism. If we increase the tossing rate to  $n$  tosses per second and decrease the probability of heads to a small  $p$ , while keeping the expected number of heads per second fixed at  $\lambda = np$ , the number of heads in a  $t$  second interval will have approximately a  $\text{Bin}(nt, p)$  distribution, which is close to the  $\text{Poisson}(\lambda t)$ . Also, the numbers of heads tossed during disjoint time intervals will still be independent random variables. In the limit, as  $n \rightarrow \infty$ , we get an idealization called a **Poisson process**.

**Definition.** A Poisson process with rate  $\lambda$  on  $[0, \infty)$  is a random mechanism that generates “points” strung out along  $[0, \infty)$  in such a way that

- (i) the number of points landing in any subinterval of length  $t$  is a random variable with a  $\text{Poisson}(\lambda t)$  distribution
- (ii) the numbers of points landing in disjoint (= non-overlapping) intervals are independent random variables.

The double use of the name Poisson is unfortunate. Much confusion would be avoided if we all agreed to refer to the mechanism as “idealized-very-fast-coin-tossing”, or some such. Then the Poisson distribution would have the same relationship to idealized-very-fast-coin-tossing as the Binomial distribution has to coin-tossing. Obversely, we could create more confusion by renaming coin tossing as “the binomial process”. Neither suggestion is likely to be adopted, so you should just get used to having two closely related objects with the name Poisson.

Why bother about Poisson processes? When we pass to the idealized mechanism of points generated in continuous time, several awkward artifacts of discrete-time coin tossing disappear.

---

Example 42: (Gamma distribution from Poisson process) The waiting time  $W_k$  to the  $k$ th point in a Poisson process with rate  $\lambda$  has a continuous distribution, with density

$$g_k(w) = \frac{\lambda^k w^{k-1} e^{-\lambda w}}{(k-1)!} \quad \text{for } w > 0.$$

---

As noted at the end of the Example, the density for  $T_k = \lambda W_k$ , takes a simpler form, namely,

$$f_k(t) = \frac{t^{k-1} e^{-t}}{(k-1)!} \quad \text{for } t > 0,$$

the so-called gamma( $k$ ) density. More generally, for each  $\alpha > 0$ ,

$$f_{\alpha}(t) = \begin{cases} \frac{t^{\alpha-1} e^{-t}}{\Gamma(\alpha)} & \text{for } t > 0, \\ 0 & \text{otherwise,} \end{cases}$$

is called the **gamma( $\alpha$ ) density**. The scaling constant,  $\Gamma(\alpha)$ , which ensures that the density integrates to one, is given by

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx \quad \text{for each } \alpha > 0.$$

The function  $\Gamma(\cdot)$  is called the **gamma function**. Don't confuse the gamma density (or the gamma distribution that it defines) with the gamma function.

---

**Example 43:** Facts about the gamma function:  $\Gamma(k) = (k-1)!$  for  $k = 1, 2, \dots$ , and  $\Gamma(1/2) = \sqrt{\pi}$ . Moments of the gamma( $\alpha$ ) distribution.

---

The special case of the gamma distribution when the parameter equals 1 is called the **(standard) exponential distribution**, with density  $f_1(t) = e^{-t}$  for  $t > 0$ , and zero elsewhere. From Example 43, if  $T_1$  has a standard exponential distribution then  $\mathbb{E}T_1 = 1$ . The waiting time  $W_1$  to the first point in a Poisson process with rate  $\lambda$  has the same distribution as  $T_1/\lambda$ , that is, a continuous distribution with density  $\lambda e^{-\lambda t}$  for  $t > 0$ , an exponential distribution with expected value  $1/\lambda$ . Don't confuse the exponential density (or the exponential distribution that it defines) with the exponential function.

Notice the parallels between the negative binomial distribution (in discrete time) and the gamma distribution (in continuous time). Each distribution corresponds to the waiting time to the  $k$ th occurrence of something, for various values of  $k$ . The negative binomial (see Problem Sheet 4) can be written as a sum of independent random variables, each with a geometric distribution. The gamma( $k$ ) can be written as a sum of  $k$  independent random variables,

$$T_k = T_1 + (T_2 - T_1) + (T_3 - T_2) + \dots + (T_k - T_{k-1}),$$

each with a standard exponential distribution. (For a Poisson process, the independence between the counts in disjoint intervals ensures that the mechanism determining the time  $W_2 - W_1$  between the first and the second points is just another Poisson process started off at time  $W_1$ . And so on.) The times between points in a Poisson process are independent, exponentially distributed random variables. *If you don't feel comfortable with this explanation, wait for the more analytic argument in the next Chapter.*

The gamma distribution is related to the normal distribution.

---

**Example 44:** The connection between gamma(1/2) and the standard normal

---

The final Example will derive probabilities related to waiting times for Poisson processes of arrivals. As part of the calculations we will need to find probabilities by conditioning on the values of a random variable with a continuous distribution. As before, the trick is first to condition on a discretized approximation to the variable, and then pass to a limit.

---

**Example 45:** Conditioning on a random variable with a continuous distribution

---

The Poisson process is often used to model the arrivals of customers in a waiting line, or the arrival of telephone calls at an exchange. The underlying idea is that of a large population of potential customers, each of whom acts independently of all the others.

---

**Example 46:** A queuing problem with a surprising solution (can be skipped)

---

**Things to remember**

- Analogies between coin tossing, as a discrete time mechanism, and the Poisson process, as a continuous time mechanism:

## DISCRETE TIME

## CONTINUOUS TIME

coin tossing, prob  $p$  of headsPoisson process with rate  $\lambda$  $X = \# \text{heads in } n \text{ tosses} \sim \text{Bin}(n, p)$  $X = \# \text{ points in } [a, a + t] \sim \text{Poisson}(\lambda t)$  $\mathbb{P}\{X = i\} = \binom{n}{i} p^i q^{n-i} \text{ for } i = 0, 1, \dots, n$  $\mathbb{P}\{X = i\} = e^{-\lambda t} (\lambda t)^i / i! \text{ for } i = 0, 1, 2, \dots$ 

(geometric)

(exponential)

 $N_1 = \# \text{ tosses to first head;}$  $T_1 / \lambda = \text{time to first point;}$  $\mathbb{P}\{N_1 = 1 + i\} = q^i p \text{ for } i = 0, 1, 2, \dots$  $T_1 \text{ has density } f_1(t) = e^{-t} \text{ for } t > 0$ 

(negative binomial)

(gamma)

 $N_k = \# \text{ tosses to } k\text{th head;}$  $T_k / \lambda = \text{time to } k\text{th point;}$ 

$$\mathbb{P}\{N_k = k + i\} = \binom{k+i-1}{k-1} q^i p^k = \binom{-k}{i} (-q)^i p^k$$

for  $i = 0, 1, 2, \dots$

$$T_k \text{ has density}$$

$$f_k(t) = t^{k-1} e^{-t} / k! \text{ for } t > 0$$

negative binomial as sum of  
independent geometrics

gamma( $k$ ) as sum of  
independent exponentials

EXAMPLE 42: GAMMA DISTRIBUTIONS FROM POISSON PROCESS

Let  $W_k$  denote the waiting time to the  $k$ th point in a Poisson process on  $[0, \infty)$  with rate  $\lambda$ . It has a continuous distribution, whose density  $g_k$  we can find by an argument similar to the one in Example 29.

For a given  $w > 0$  and small  $\delta > 0$ , write  $M$  for the number of points landing in the interval  $[0, t)$ , and  $N$  for the number of points landing in the interval  $[w, w + \delta]$ . From the definition of a Poisson process,  $M$  and  $N$  are independent random variables with

$$M \sim \text{Poisson}(\lambda w) \quad \text{and} \quad N \sim \text{Poisson}(\lambda \delta).$$

To have  $W_k$  lie in the interval  $[w, w + \delta]$  we must have  $N \geq 1$ . When  $N = 1$ , we need exactly  $k - 1$  points to land in  $[0, t)$ .

$$\mathbb{P}\{w \leq W_k \leq w + \delta\} = \mathbb{P}\{M = k - 1, N = 1\} + \mathbb{P}\{w \leq W_k < w + \delta, N \geq 2\}.$$

When  $N \geq 2$ , the exact specification of what we need becomes more complicated, but luckily all such terms make a contribution of order  $\delta^2$  because

$$\mathbb{P}\{N \geq 2\} = \frac{e^{-\lambda \delta} (\lambda \delta)^2}{2!} + \frac{e^{-\lambda \delta} (\lambda \delta)^3}{3!} \dots$$

Independence of  $M$  and  $N$  lets us factorize the contribution from  $N = 1$  into

$$\begin{aligned} \mathbb{P}\{M = k - 1\} \mathbb{P}\{N = 1\} &= \frac{e^{-\lambda w} (\lambda w)^{k-1}}{(k-1)!} \frac{e^{-\lambda \delta} (\lambda \delta)^1}{1!} \\ &= \frac{e^{-\lambda w} \lambda^k w^{k-1}}{(k-1)!} \delta + \text{smaller order terms,} \end{aligned}$$

Thus

$$\mathbb{P}\{w \leq W_k \leq w + \delta\} = \frac{e^{-\lambda w} \lambda^k w^{k-1}}{(k-1)!} \delta + \text{smaller order terms,}$$

which makes

$$g_k(w) = \frac{e^{-\lambda w} \lambda^k w^{k-1}}{(k-1)!} \quad \text{for } w > 0.$$

the density function for  $W_k$ . □

It is easier to remember the distribution if we rescale, defining  $T_k = \lambda W_k$ . Then for each  $t > 0$  and each small  $\delta > 0$ ,

$$\begin{aligned} \mathbb{P}\{t \leq T_k \leq t + \delta\} &= \mathbb{P}\{t/\lambda \leq W_k \leq (t + \delta)/\lambda\} \\ &= g_k(t/\lambda)(\delta/\lambda) + \text{smaller order terms} \\ &= \frac{t^{k-1} e^{-t}}{(k-1)!} \delta + \dots \end{aligned}$$

That is,  $T_k$  has a continuous distribution with a **gamma( $k$ ) density**,

$$f_k(t) = \frac{t^{k-1} e^{-t}}{(k-1)!} \quad \text{for } t > 0.$$

REMARK. Notice that  $g_k = f_k$  when  $\lambda = 1$ . That is,  $T_k$  is the waiting time to the  $k$ th point for a Poisson process with rate 1. Put another way, we can generate a Poisson process with rate  $\lambda$  by taking the points appearing at times  $0 < T_1 < T_2 < T_3 < \dots$  from a Poisson process with rate 1, then rescaling to produce a new process with points at

$$0 < \frac{T_1}{\lambda} < \frac{T_2}{\lambda} < \frac{T_3}{\lambda} < \dots$$

You could verify this assertion by checking the two defining properties for a Poisson process with rate  $\lambda$ . Doesn't it make sense that, as  $\lambda$  gets bigger, the points appear more rapidly?



EXAMPLE 43: FACTS ABOUT THE GAMMA FUNCTION AND GAMMA DISTRIBUTION

The gamma function is defined for  $\alpha > 0$  by

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx.$$

By direct integration,  $\Gamma(1) = \int_0^{\infty} e^{-x} dx = 1$ . Also, a change of variable  $y = \sqrt{2x}$  gives

$$\begin{aligned} \Gamma(1/2) &= \int_0^{\infty} x^{-1/2} e^{-x} dx \\ &= \int_0^{\infty} \sqrt{2} e^{-y^2/2} dy \\ &= \frac{\sqrt{2}}{2} \int_{-\infty}^{\infty} e^{-y^2/2} dy \\ &= \sqrt{\pi} \quad \text{from the Appendix to Chapter 7.} \end{aligned}$$

For each  $\alpha > 0$ , an integration by parts gives

$$\begin{aligned} \Gamma(\alpha + 1) &= \int_0^{\infty} x^{\alpha} e^{-x} dx \\ &= [-x^{\alpha} e^{-x}]_0^{\infty} + \alpha \int_0^{\infty} x^{\alpha-1} e^{-x} dx \\ &= \alpha \Gamma(\alpha). \end{aligned}$$

Repeated appeals of the same formula, for each  $\alpha > 0$  and each positive integer  $m$ , give

$$(*) \quad \Gamma(\alpha + m) = (\alpha + m - 1)(\alpha + m - 2) \dots (\alpha) \Gamma(\alpha).$$

In particular,

$$\Gamma(k) = (k - 1)(k - 2)(k - 3) \dots (2)(1) \Gamma(1) = (k - 1)! \quad \text{for } k = 1, 2, \dots$$

### Gamma distribution

For parameter value  $\alpha > 0$ , the gamma( $\alpha$ ) distribution is defined by its density

$$f_{\alpha}(t) = \begin{cases} t^{\alpha-1} e^{-t} / \Gamma(\alpha) & \text{for } t > 0 \\ 0 & \text{otherwise} \end{cases}$$

If a random variable  $T$  has a gamma( $\alpha$ ) distribution then, for each positive integer  $m$ ,

$$\begin{aligned} \mathbb{E}T^m &= \int_0^{\infty} t^m f_{\alpha}(t) dt \\ &= \int_0^{\infty} \frac{t^m t^{\alpha-1} e^{-t}}{\Gamma(\alpha)} dt \\ &= \frac{\Gamma(\alpha + m)}{\Gamma(\alpha)} \\ &= (\alpha + m - 1)(\alpha + m - 2) \dots (\alpha) \quad \text{by equality } (*). \end{aligned}$$

In particular,  $\mathbb{E}T = \alpha$  and

$$\text{var}(T) = \mathbb{E}(T^2) - (\mathbb{E}T)^2 = (\alpha + 1)\alpha - \alpha^2 = \alpha.$$

□

EXAMPLE 44: GAMMAS FROM NORMALS

Suppose  $Z$  has a standard normal distribution, with density  $\phi(t) = \exp(-t^2/2)/\sqrt{2\pi}$  for  $-\infty < t < \infty$ . The random variable  $Y = Z^2/2$  has a continuous distribution concentrated on the positive half line  $(0, \infty)$ . For  $y > 0$ , and  $\delta > 0$  small,

$$\begin{aligned}\mathbb{P}\{y < Y < y + \delta\} &= \mathbb{P}\{2y < Z^2 < 2y + 2\delta\} \\ &= \mathbb{P}\{\sqrt{2y} < Z < \sqrt{2y + 2\delta} \text{ or } -\sqrt{2y + 2\delta} < Z < -\sqrt{2y}\}.\end{aligned}$$

Notice the two contributions; the square function is not one-to-one.

As in Example 28, Calculus gives a good approximation to the length of the short interval from  $\sqrt{2y}$  to  $\sqrt{2y + 2\delta}$ . Temporarily write  $g(y)$  for  $\sqrt{2y}$ . Then

$$\sqrt{2y + 2\delta} - \sqrt{2y} = g(y + \delta) - g(y) \approx \delta g'(y) = \delta/\sqrt{2y}.$$

The interval from  $-\sqrt{2y + 2\delta}$  to  $-\sqrt{2y}$  has the same length. Thus

$$\begin{aligned}\mathbb{P}\{y < Y < y + \delta\} &\approx \frac{\delta}{\sqrt{2y}}\phi(\sqrt{2y}) + \frac{\delta}{\sqrt{2y}}\phi(-\sqrt{2y}) \\ &= \frac{2\delta}{\sqrt{2y}} \frac{1}{\sqrt{2\pi}} \exp\left(-\left(\sqrt{2y}\right)^2/2\right) \\ &= \frac{\delta}{\sqrt{\pi}} y^{-1/2} e^{-y}.\end{aligned}$$

That is,  $Y$  has the distribution with density

$$\frac{1}{\sqrt{\pi}} y^{-1/2} e^{-y} \quad \text{for } y > 0.$$

Compare with the gamma(1/2) density,

$$\frac{y^{1-1/2} e^{-y}}{\Gamma(1/2)} \quad \text{for } y > 0.$$

The distribution of  $Z^2/2$  is gamma (1/2). □

REMARK. The change of variable in Example 43, used to prove  $\Gamma(1/2) = \sqrt{\pi}$ , is essentially the same idea as the calculation used to prove  $\sqrt{2y + 2\delta} - \sqrt{2y} \approx \delta/\sqrt{2y}$ .

EXAMPLE 45: CONDITIONING ON A RV WITH A CONTINUOUS DISTRIBUTION

Suppose  $T$  has density  $f(\cdot)$ , and let  $X$  be another random variable. We can calculate  $\mathbb{E}X$  as a weighted average of the conditional expectations  $\mathbb{E}(X \mid T = t)$ , by means of an approximation argument.

First break the whole range for  $T$  into small intervals, each of length  $\delta$ . Rule E4 for expectations gives

$$\mathbb{E}X = \sum_{j=-\infty}^{\infty} \mathbb{E}(X \mid j\delta \leq T < (j+1)\delta) \mathbb{P}\{j\delta \leq T < (j+1)\delta\}$$

If  $\delta$  is small, the first factor in the  $j$ th summand is close to  $\mathbb{E}(X \mid T = j\delta)$ , and the second factor is close to  $f(j\delta)\delta$ , allowing us to approximate,

$$\mathbb{E}X \approx \sum_{j=-\infty}^{\infty} g(j\delta) f(j\delta) \delta \quad \text{where } g(t) \text{ denotes } \mathbb{E}(X \mid T = t).$$

The last sum is an approximation to  $\int_{-\infty}^{\infty} g(t) f(t) dt$ . As  $\delta$  tends to zero, the errors of approximation to both the expectation and the integral tend to zero, leaving (in the limit)

$$\mathbb{E}X = \int \mathbb{E}(X \mid T = t) f(t) dt \quad \text{for each random variable } X.$$

As a special case, when  $X$  is replaced by the indicator function of an event, we get

$$\mathbb{P}A = \int \mathbb{P}(A \mid T = t) f(t) dt \quad \text{for each event } A,$$

Rule E4 for expectations strikes again. □

# EXAMPLE 46: A QUEUING PROBLEM

Suppose an office receives two different types of inquiry: persons who walk in off the street, and persons who call by telephone. Suppose the two types of arrival are described by independent Poisson processes, with rate  $\lambda_w$  for the walk-ins, and rate  $\lambda_c$  for the callers. What is the distribution of the number of telephone calls received before the first walk-in customer?

Write  $T$  for the arrival time of the first walk-in, and let  $N$  be the number of calls in  $[0, T)$ . The time  $T$  has a continuous distribution, with the exponential density  $f(t) = \lambda_w e^{-\lambda_w t}$  for  $t > 0$ . We need to calculate  $\mathbb{P}\{N = i\}$  for  $i = 0, 1, 2, \dots$ . From Example 45, with  $A$  equal to  $\{N = i\}$ ,

$$\mathbb{P}\{N = i\} = \int_0^\infty \mathbb{P}\{N = i \mid T = t\} f(t) dt.$$

The conditional distribution of  $N$  is affected by the walk-in process only insofar as that process determines the length of the time interval over which  $N$  counts. Given  $T = t$ , the random variable  $N$  has a  $\text{Poisson}(\lambda_c t)$  conditional distribution. Thus

$$\begin{aligned} \mathbb{P}\{N = i\} &= \int_0^\infty \frac{e^{-\lambda_c t} (\lambda_c t)^i}{i!} \lambda_w e^{-\lambda_w t} dt \\ &= \lambda_w \frac{\lambda_c^i}{i!} \int_0^\infty \left( \frac{x}{\lambda_c + \lambda_w} \right)^i e^{-x} \frac{dx}{\lambda_c + \lambda_w} \quad \text{putting } x = (\lambda_c + \lambda_w)t \\ &= \frac{\lambda_w}{\lambda_c + \lambda_w} \left( \frac{\lambda_c}{\lambda_c + \lambda_w} \right)^i \frac{1}{i!} \int_0^\infty x^i e^{-x} dx \end{aligned}$$

The  $1/i!$  and the last integral cancel. (Compare with  $\Gamma(i+1)$ .) Writing  $p$  for  $\lambda_w/(\lambda_c + \lambda_w)$  we have

$$\mathbb{P}\{N = i\} = p(1-p)^i \quad \text{for } i = 0, 1, 2, \dots$$

Compare with the  $\text{geometric}(p)$  distribution. The random variable  $N$  has the distribution of the number of tails tossed before the first head, for independent tosses of a coin that lands heads with probability  $p$ . □

Such a nice clean result couldn't happen just by accident. Maybe we don't need all the Calculus to arrive at the distribution for  $N$ . In fact, the properties of the Poisson distribution and Problem 7.6 show what is going on, as I will now explain.

Consider the process of all inquiries, both walk-ins and calls. In an interval of length  $t$ , the total number of inquiries is the sum of a  $\text{Poisson}(\lambda_w t)$  distributed random variable and an independent  $\text{Poisson}(\lambda_c t)$  distributed random variable; the total has a  $\text{Poisson}(\lambda_w t + \lambda_c t)$  distribution. Both walk-ins and calls contribute independent counts to disjoint intervals; the total counts for disjoint intervals are independent random variables. It follows that the process of all arrivals is a Poisson process with rate  $\lambda_w + \lambda_c$ .

Now consider an interval of length  $t$  in which there are  $X$  walk-ins and  $Y$  calls. From Problem 7.6, given that  $X + Y = n$ , the conditional distribution of  $X$  is  $\text{Bin}(n, p)$ , where

$$p = \frac{\lambda_w t}{\lambda_w t + \lambda_c t} = \frac{\lambda_w}{\lambda_w + \lambda_c}$$

That is,  $X$  has the conditional distribution that would be generated by the following mechanism:

- (1) Generate inquiries as a Poisson process with rate  $\lambda_w + \lambda_c$ .
- (2) For each inquiry, toss a coin that lands heads with probability  $p = \lambda_w/(\lambda_w + \lambda_c)$ . For a head, declare the arrival to be a walk-in, for a tail declare it to be a call.

A formal proof that this two-step mechanism does generate a pair of independent Poisson processes, with rates  $\lambda_w$  and  $\lambda_c$ , would involve:

(1') Prove independence between disjoint intervals. (Easy)

(2') If step 2 generates  $X$  walk-ins and  $Y$  calls in an interval of length  $t$ , show that

$$\begin{aligned}\mathbb{P}\{X = i, Y = j\} &= \mathbb{P}\{X = i\}\mathbb{P}\{Y = j\} \\ X \sim \text{Poisson}(\lambda_w t) \quad &\text{and} \quad Y \sim \text{Poisson}(\lambda_c t)\end{aligned}$$

You should be able to write out the necessary conditioning argument for (2').

The two-step mechanism explains the appearance of the geometric distribution in the problem posed at the start of the Example. The classification of each inquiry as either a walk-in or a call is effectively carried out by a sequence of independent coin tosses, with probability  $p$  of a head (= a walk-in). The problem asks for the distribution of the number of tails before the first head. The embedding of the inquiries into continuous time is irrelevant.

## Chapter 10

# Joint densities

Consider the general problem of describing probabilities involving two random variables,  $X$  and  $Y$ . If both have discrete distributions, with  $X$  taking values  $x_1, x_2, \dots$  and  $Y$  taking values  $y_1, y_2, \dots$ , then everything about the joint behavior of  $X$  and  $Y$  can be deduced from the set of probabilities

$$\mathbb{P}\{X = x_i, Y = y_j\} \quad \text{for } i = 1, 2, \dots \text{ and } j = 1, 2, \dots$$

We have been working for some time with problems involving such pairs of random variables, but we have not needed to formalize the concept of a joint distribution. When both  $X$  and  $Y$  have continuous distributions, it becomes more important to have a systematic way to describe how one might calculate probabilities of the form  $\mathbb{P}\{(X, Y) \in B\}$  for various subsets  $B$  of the plane. For example, how could one calculate  $\mathbb{P}\{X < Y\}$  or  $\mathbb{P}\{X^2 + Y^2 \leq 9\}$  or  $\mathbb{P}\{X + Y \leq 7\}$ ?

**Definition.** Say that random variables  $X$  and  $Y$  have a jointly continuous distribution with **joint density function**  $f(\cdot, \cdot)$  if, for each subset  $B$  of  $\mathbb{R}^2$ ,

$$\mathbb{P}\{(X, Y) \in B\} = \iint_{\{(x, y) \in B\}} f(x, y) dx dy.$$

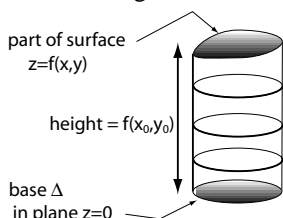
Equivalently, for each  $(x_0, y_0)$  and each small region  $\Delta$  containing  $(x_0, y_0)$ ,

$$\mathbb{P}\{(X, Y) \in \Delta\} = (\text{area of } \Delta)f(x_0, y_0) + \text{smaller order terms.}$$

REMARK. There might be a small set of points  $(x_0, y_0)$  at which the equivalent form of the definition does not work. For example, if  $f$  takes a particular form in some region  $R$ , but is zero outside that region, we might get different answers for  $f(x_0, y_0)$  at points on the boundary of  $R$ , depending on whether  $\Delta$  is chosen to poke into the region  $R$ , or outside  $R$ . In fact, we need not worry about these points because it will always turn out that they make no contribution to any of the double integrals defining probabilities  $\mathbb{P}\{(X, Y) \in B\}$ .

Apart from the replacement of single integrals by double integrals, and the replacement of intervals of small length by regions of small area, the definition of a joint density is essentially the same as the definition for densities on the real line in Chapter 6.

The density function defines a surface, via the equation  $z = f(x, y)$ . A thin column, sitting on the base  $\Delta$  around  $(x_0, y_0)$  in the plane  $z = 0$  and reaching up to that surface, has



volume approximately equal to  $(\text{area of } \Delta) \times f(x_0, y_0)$ . The approximation comes from ignoring variations in the height of the column. To calculate  $\mathbb{P}\{(X, Y) \in B\}$  for a larger region  $B$ , we could partition  $B$  into small regions  $\Delta_1, \Delta_2, \dots$ , then add up the probabilities  $\mathbb{P}\{(X, Y) \in \Delta_1\} + \mathbb{P}\{(X, Y) \in \Delta_2\} + \dots$ . The contribution from each  $\Delta_i$  is approximately equal to the volume of a thin column. The sum of the contributions is approximately equal to the volume of the

entire region bounded by the surface and the plane  $z = 0$ , and lying above the set  $B$ , a volume given precisely by the double integral. As we make the partitions finer, the errors of approximation go to zero. In the limit,  $\mathbb{P}\{(X, Y) \in B\}$  is recovered as the double integral.

To ensure that  $\mathbb{P}\{(X, Y) \in B\}$  is nonnegative and that it equals one when  $B$  is the whole of  $\mathbb{R}^2$ , we must require

$$f \geq 0 \quad \text{and} \quad \iint \{(x, y) \in \mathbb{R}^2\} f(x, y) dx dy = 1.$$

When we wish to calculate a density, the small region  $\Delta$  can be chosen in many ways—small rectangles, small disks, small blobs, and even small shapes that don't have any particular name—whatever suits the needs of a particular calculation.

---

#### Example 47: Joint densities for independent random variables

---

When pairs of random variables are not independent it takes more work to find a joint density. The prototypical case, where new random variables are constructed as linear functions of random variables with a known joint density, illustrates a general method for deriving joint densities.

---

#### Example 48: Joint densities for linear combinations

---

Read through the details of the following important special case, to make sure you understand the notation from Example 48.

---

#### Example 49: Linear combinations of independent normals

---

The method used in Example 48, for linear transformations, gives a good approximation for more general *smooth* transformations when applied to small regions. Densities describe the behaviour of distributions in small regions; in small regions smooth transformations are approximately linear; the density formula for linear transformations gives the density formula for smooth transformations in small regions.

---

**Example 50:** Suppose  $X$  and  $Y$  are independent random variables, with  $X$  having a  $\text{gamma}(\alpha)$  distribution and  $Y$  having a  $\text{gamma}(\beta)$  distribution. Find the joint density for the random variables  $U = X/(X + Y)$  and  $V = X + Y$ .

---

As shown in Example 50, the random variables  $U$  and  $V$  have joint density

$$\psi(u, v) = g(u)h(v) \quad \text{for } 0 < u < 1 \text{ and } 0 < v < \infty,$$

where

$$g(u) = \frac{u^{\alpha-1}(1-u)^{\beta-1}}{B(\alpha, \beta)} \quad \text{for } 0 < u < 1, \text{ the beta}(\alpha, \beta) \text{ density}$$

$$h(v) = \frac{v^{\alpha+\beta-1}e^{-v}}{\Gamma(\alpha + \beta)} \quad \text{for } 0 < v < \infty, \text{ the gamma}(\alpha + \beta) \text{ density.}$$

The factorization of the joint density implies that the random variables  $U$  and  $V$  are independent. To see why, consider any pair of subsets  $A$  and  $B$  of the real line. The defining property of the joint density gives

$$\begin{aligned} \mathbb{P}\{U \in A\} &= \mathbb{P}\{U \in A, 0 < V < \infty\} \\ &= \iint \{u \in A, 0 < v < \infty\} g(u)h(v) du dv \\ &= \int \{u \in A\} g(u) du. \end{aligned}$$

That is,  $U$  has density  $g$ ; it has a  $\text{beta}(\alpha, \beta)$  distribution. Similarly,  $V$  has a continuous distribution with density  $h$ ; it has a  $\text{gamma}(\alpha + \beta)$  distribution. Finally,

$$\begin{aligned}\mathbb{P}\{U \in A, V \in B\} &= \iint \{u \in A, v \in B\} \psi(u, v) du dv \\ &= \int \{u \in A\} g(u) du \int \{v \in B\} h(v) dv = \mathbb{P}\{U \in A\} \mathbb{P}\{V \in B\}.\end{aligned}$$

The events  $\{U \in A\}$  and  $\{V \in B\}$  are independent, for all choices of  $A$  and  $B$ . In summary: *if  $X \sim \text{gamma}(\alpha)$  independently of  $Y \sim \text{gamma}(\beta)$ , then  $X/(X + Y) \sim \text{beta}(\alpha, \beta)$  independently of  $X + Y \sim \text{gamma}(\alpha + \beta)$ .*

In general, if  $X$  and  $Y$  have a joint density function  $f(x, y)$  then

$$\mathbb{P}\{X \in A\} = \iint \{x \in A, -\infty < y < \infty\} f(x, y) dx dy = \int \{x \in A\} f_X(x) dx,$$

where

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

That is,  $X$  has a continuous distribution with (marginal) density function  $f_X$ . Similarly,  $Y$  has a continuous distribution with (marginal) density function  $f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$ . Remember that the word marginal is redundant; it serves merely to stress that a calculation refers only to one of the random variables.

The conclusion about  $X + Y$  from Example 50 extends to sums of more than two independent random variables, each with a gamma distribution. The result has a particularly important special case, involving the sums of squares of independent standard normals.

---

#### Example 51: Sums of independent gamma random variables

---

#### APPENDIX: AREA OF A PARALLELOGRAM

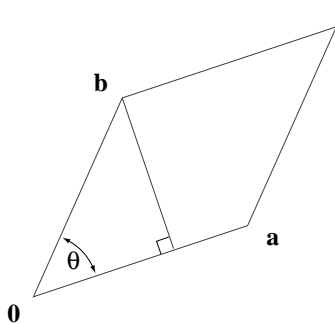
Let  $R$  be a parallelogram in the plane with corners at  $\mathbf{0} = (0, 0)$ ,  $\mathbf{a} = (a_1, a_2)$ ,  $\mathbf{b} = (b_1, b_2)$ , and  $\mathbf{a} + \mathbf{b}$ . The area of  $R$  is equal to the absolute value of the determinant of the matrix

$$J = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \end{pmatrix} = (\mathbf{a}, \mathbf{b}).$$

That is, the area of  $R$  equals  $|a_1 b_2 - a_2 b_1|$ .

*Proof.* Let  $\theta$  denotes the angle between  $\mathbf{a}$  and  $\mathbf{b}$ . Remember that

$$\|\mathbf{a}\| \times \|\mathbf{b}\| \times \cos(\theta) = \mathbf{a} \cdot \mathbf{b}$$



With the side from  $\mathbf{0}$  to  $\mathbf{a}$ , which has length  $\|\mathbf{a}\|$ , as the base, the vertical height is  $\|\mathbf{b}\| \times |\sin \theta|$ . The absolute value of the area equals  $\|\mathbf{a}\| \times \|\mathbf{b}\| \times |\sin \theta|$ . The square of the area equals

$$\begin{aligned}\|\mathbf{a}\|^2 \|\mathbf{b}\|^2 \sin^2(\theta) &= \|\mathbf{a}\|^2 \|\mathbf{b}\|^2 - \|\mathbf{a}\|^2 \|\mathbf{b}\|^2 \cos^2(\theta) \\ &= (\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b}) - (\mathbf{a} \cdot \mathbf{b})^2 \\ &= \det \begin{pmatrix} \mathbf{a} \cdot \mathbf{a} & \mathbf{a} \cdot \mathbf{b} \\ \mathbf{a} \cdot \mathbf{b} & \mathbf{b} \cdot \mathbf{b} \end{pmatrix} \\ &= \det(J'J) \\ &= (\det J)^2.\end{aligned}$$

If you are not sure about the properties of determinants used in the last two lines, you

□ should check directly that  $(a_1^2 + a_2^2)(b_1^2 + b_2^2) = (a_1 b_2 - a_2 b_1)^2$ .



#### EXAMPLE 47: JOINT DENSITIES FOR INDEPENDENT RANDOM VARIABLES

When  $X$  has density  $g(x)$  and  $Y$  has density  $h(y)$ , and  $X$  is independent of  $Y$ , the joint density is particularly easy to calculate. Let  $\Delta$  be a small rectangle with one corner at  $(x_0, y_0)$  and small sides of length  $\delta_x > 0$  and  $\delta_y > 0$ ,

$$\Delta = \{(x, y) \in \mathbb{R}^2 : x_0 \leq x \leq x_0 + \delta_x, y_0 \leq y \leq y_0 + \delta_y\}.$$

By independence,

$$\mathbb{P}\{(X, Y) \in \Delta\} = \mathbb{P}\{x_0 \leq X \leq x_0 + \delta_x\} \mathbb{P}\{y_0 \leq Y \leq y_0 + \delta_y\}$$

Invoke the defining property of the densities  $g$  and  $h$  to approximate the last product by

$$(g(x_0)\delta_x + \text{smaller order terms})(h(y_0)\delta_y + \text{smaller order terms}) \approx \delta_x \delta_y g(x_0)h(y_0).$$

Thus  $f(x_0, y_0) = g(x_0)h(y_0)$ . That is, the joint density  $f$  is the product of the **marginal densities**  $g$  and  $h$ . The word *marginal* is used here to distinguish the joint density for  $(X, Y)$  from the individual densities  $g$  and  $h$ . □

EXAMPLE 48: JOINT DENSITIES FOR LINEAR COMBINATIONS

Suppose  $X$  and  $Y$  have a jointly continuous distribution with joint density  $f(x, y)$ . For constants  $a, b, c, d$ , define

$$U = aX + bY \quad \text{and} \quad V = cX + dY$$

Find the joint density function  $\psi(u, v)$  for  $(U, V)$ , under the assumption that the quantity  $\kappa = ad - bc$  is nonzero.

Think of the pair  $(U, V)$  as defining a new random point in  $\mathbb{R}^2$ . That is  $(U, V) = T(X, Y)$ , where  $T$  maps the point  $(x, y) \in \mathbb{R}^2$  to the point  $(u, v) \in \mathbb{R}^2$  with

$$u = ax + by \quad \text{and} \quad v = cx + dy,$$

or in matrix notation,

$$(u, v) = (x, y)A \quad \text{where} \quad A = \begin{pmatrix} a & c \\ b & d \end{pmatrix}.$$

Notice that  $\det A = ad - bc = \kappa$ . The assumption that  $\kappa \neq 0$  ensures that the transformation is invertible:

$$(u, v)A^{-1} = (x, y) \quad \text{where} \quad A^{-1} = \frac{1}{\kappa} \begin{pmatrix} d & -c \\ -b & a \end{pmatrix}.$$

That is,

$$\frac{du - bv}{\kappa} = x \quad \text{and} \quad \frac{-cu + av}{\kappa} = y.$$

Notice that  $\det(A^{-1}) = 1/\kappa = 1/(\det A)$ .

It helps to distinguish between the two roles for  $\mathbb{R}^2$ , referring to the domain of  $T$  as the  $(X, Y)$ -plane and the range as the  $(U, V)$ -plane.

The joint density function  $\psi(u, v)$  is characterized by the property that

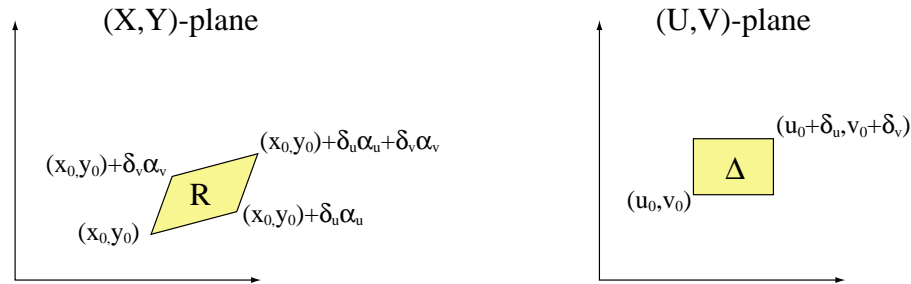
$$\mathbb{P}\{u_0 \leq U \leq u_0 + \delta_u, v_0 \leq V \leq v_0 + \delta_v\} \approx \psi(u_0, v_0)\delta_u\delta_v$$

for each  $(u_0, v_0)$  in the  $(U, V)$ -plane, and small  $(\delta_u, \delta_v)$ . To calculate the probability on the left-hand side we need to find the region  $R$  in the  $(X, Y)$ -plane corresponding to the small rectangle  $\Delta$ , with corners at  $(u_0, v_0)$  and  $(u_0 + \delta_u, v_0 + \delta_v)$ , in the  $(U, V)$ -plane.

The linear transformation  $A^{-1}$  maps parallel straight lines in the  $(U, V)$ -plane into parallel straight lines in the  $(X, Y)$ -plane. The region  $R$  must be a parallelogram, with vertices

$$\begin{aligned} (x_0, y_0 + \delta_y) &= (u_0, v_0 + \delta_v)A^{-1} & \text{and} & & (x_0 + \delta_x, y_0 + \delta_y) &= (u_0 + \delta_u, v_0 + \delta_v)A^{-1} \\ (x_0, y_0) &= (u_0, v_0)A^{-1} & \text{and} & & (x_0 + \delta_x, y_0) &= (u_0 + \delta_u, v_0)A^{-1} \end{aligned}$$

More succinctly,  $(\delta_x, \delta_y) = (\delta_u, \delta_v)A^{-1} = \delta_u\alpha_u + \delta_v\alpha_v$ , where  $\alpha_u = (d, -c)/\kappa$  and  $\alpha_v = (-b, a)/\kappa$  denote the two rows of  $A^{-1}$ .



From the formula in the Appendix to the Chapter, the parallelogram  $R$  has area

$$|\det(\delta_u \alpha'_u, \delta_v \alpha'_v)| = \delta_u \delta_v |\det(A^{-1})'| = \frac{\delta_u \delta_v}{|\det A|}.$$

For small  $\delta_u > 0$  and  $\delta_v > 0$ ,

$$\begin{aligned}\psi(u_0, v_0)\delta_u\delta_v &\approx \mathbb{P}\{(U, V) \in \Delta\} \\ &= \mathbb{P}\{(X, Y) \in R\} \\ &\approx (\text{area of } R)f(x_0, y_0) \\ &\approx \delta_u\delta_v f(x_0, y_0)/|\det(A)|\end{aligned}$$

It follows that  $(U, V)$  have joint density

$$\psi(u, v) = \frac{1}{|\det A|} f(x, y) \quad \text{where } (x, y) = (u, v)A^{-1}.$$

On the right-hand side you should substitute  $(du - bv)/\kappa$  for  $x$  and  $(-cu + av)/\kappa$  for  $y$ , in order to get an expression involving only  $u$  and  $v$ .

In effect, we have calculated a Jacobian by first principles. □

EXAMPLE 49: LINEAR COMBINATIONS OF INDEPENDENT NORMALS

Suppose  $X$  and  $Y$  are independent random variables, each distributed  $N(0, 1)$ . By Example 47, the joint density for  $(X, Y)$  equals

$$f(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right) \quad \text{for all } x, y.$$

By Example 48, the joint distribution of the random variables

$$U = aX + bY \quad \text{and} \quad V = cX + dY$$

has the joint density

$$\begin{aligned} \psi(u, v) &= \frac{1}{2\pi|\kappa|} \exp\left(-\frac{1}{2}\left(\frac{du - bv}{\kappa}\right)^2 - \frac{1}{2}\left(\frac{-cu + av}{\kappa}\right)^2\right) \quad \text{where } \kappa = ad - bc \\ &= \frac{1}{2\pi|\kappa|} \exp\left(-\frac{(c^2 + d^2)u^2 - 2(db + ac)uv + (a^2 + b^2)v^2}{2\kappa^2}\right) \end{aligned}$$

You'll learn more about joint normal distributions in Chapter 12. □

# EXAMPLE 50: BETAS FROM GAMMAS

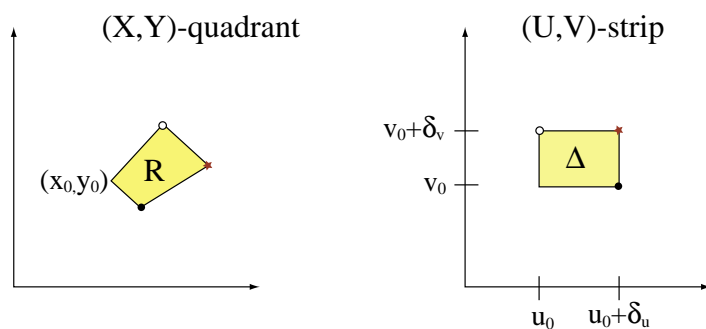
Suppose  $X$  and  $Y$  are independent random variables, with  $X$  having a  $\text{gamma}(\alpha)$  distribution and  $Y$  having a  $\text{gamma}(\beta)$  distribution. Find the joint density for the random variables  $U = X/(X + Y)$  and  $V = X + Y$ .

Write  $U$  for  $X/(X + Y)$  and  $V$  for  $X + Y$ . The pair  $(X, Y)$  takes values ranging over the positive quadrant  $(0, \infty)^2$ , with joint density function

$$f(x, y) = \frac{x^{\alpha-1} e^{-x}}{\Gamma(\alpha)} \times \frac{y^{\beta-1} e^{-y}}{\Gamma(\beta)} \quad \text{for } x > 0, y > 0.$$

The pair  $(U, V)$  takes values in a strip where  $0 < u < 1$  and  $0 < v < \infty$ . The joint density function,  $\psi(u, v)$ , for  $(U, V)$  can be determined by considering corresponding points  $(x_0, y_0)$  in the  $(X, Y)$ -quadrant and  $(u_0, v_0)$  in the  $(U, V)$ -strip, where

$$\begin{aligned} u_0 &= x_0/(x_0 + y_0) & \text{and} & & v_0 &= x_0 + y_0, \\ x_0 &= u_0 v_0 & \text{and} & & y_0 &= (1 - u_0) v_0. \end{aligned}$$



When  $(U, V)$  lies near  $(u_0, v_0)$  then  $(X, Y)$  lies near  $(x_0, y_0)$ . More precisely, for small positive  $\delta_u$  and  $\delta_v$ , there is a small region  $R$  in the  $(X, Y)$ -quadrant corresponding to the small rectangle

$$\Delta = \{(u, v) : u_0 \leq u \leq u_0 + \delta_u, v_0 \leq v \leq v_0 + \delta_v\}$$

in the  $(U, V)$ -strip. First locate the points corresponding to the corners of  $\Delta$ .

$$\begin{aligned} (u_0 + \delta_u, v_0) &\mapsto (x_0, y_0) + (\delta_u v_0, -\delta_u v_0) \\ (u_0, v_0 + \delta_v) &\mapsto (x_0, y_0) + (\delta_v u_0, \delta_v (1 - u_0)) \\ (u_0 + \delta_u, v_0 + \delta_v) &\mapsto (x_0, y_0) + (\delta_u v_0 + \delta_v u_0, -\delta_u v_0 + \delta_v (1 - u_0)) + (\delta_u \delta_v, -\delta_u \delta_v) \end{aligned}$$

In matrix notation,

$$\begin{aligned} (u_0, v_0) + (\delta_u, 0) &\mapsto (x_0, y_0) + (\delta_u, 0)J \\ (u_0, v_0) + (0, \delta_v) &\mapsto (x_0, y_0) + (0, \delta_v)J \\ (u_0, v_0) + (\delta_u, \delta_v) &\mapsto (x_0, y_0) + (\delta_u, \delta_v)J + \text{smaller order terms} \end{aligned} \quad \text{where } J = \begin{pmatrix} v_0 & -v_0 \\ u_0 & 1 - u_0 \end{pmatrix}$$

You might recognize  $J$  as the **Jacobian matrix** of partial derivatives

$$\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{pmatrix}$$

evaluated at  $(u_0, v_0)$ . For small perturbations, the transformation from  $(u, v)$  to  $(x, y)$  is approximately linear.

The region  $R$  is approximately a parallelogram, with the edges oblique to the coordinate axes. To a good approximation, the area of  $R$  is equal to  $\delta_u \delta_v$  times the area of the parallelogram with corners at

$$(0, 0) \quad \text{and} \quad \mathbf{a} = (v_0, -v_0) \quad \text{and} \quad \mathbf{b} = (u_0, 1 - u_0) \quad \text{and} \quad \mathbf{a} + \mathbf{b},$$

which, from the Appendix to the Chapter, equals  $|\det(J)| = v_0$ .

The rest of the calculation of the joint density  $\psi(\cdot, \cdot)$  for  $(U, V)$  is easy:

$$\begin{aligned} \delta_u \delta_v \psi(u_0, v_0) &\approx \mathbb{P}\{(U, V) \in \Delta\} \\ &= \mathbb{P}\{(X, Y) \in R\} \\ &\approx f(x_0, y_0)(\text{area of } R) \\ &\approx \frac{x_0^{\alpha-1} e^{-x_0}}{\Gamma(\alpha)} \frac{y_0^{\beta-1} e^{-y_0}}{\Gamma(\beta)} \delta_u \delta_v v_0 \end{aligned}$$

Substitute  $x_0 = u_0 v_0$  and  $y_0 = (1 - u_0) v_0$  to get the joint density

$$\psi(u_0, v_0) = \frac{u_0^{\alpha-1} v_0^{\alpha-1} e^{-u_0 v_0}}{\Gamma(\alpha)} \frac{(1 - u_0)^{\beta-1} v_0^{\beta-1} e^{-v_0 + u_0 v_0}}{\Gamma(\beta)} v_0$$

If we write

$$\begin{aligned} g(u) &= \frac{u^{\alpha-1} (1 - u)^{\beta-1}}{B(\alpha, \beta)} && \text{the beta}(\alpha, \beta) \text{ density} \\ h(v) &= \frac{v^{\alpha+\beta-1} e^{-v}}{\Gamma(\alpha + \beta)} && \text{the gamma}(\alpha + \beta) \text{ density.} \end{aligned}$$

then

$$\psi(u, v) = g(u)h(v) \frac{B(\alpha, \beta)\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \quad \text{for } 0 < u < 1 \text{ and } 0 < v < \infty.$$

I have dropped the subscripting zeros because I no longer need to keep your attention fixed on a particular  $(u_0, v_0)$  in the  $(U, V)$  strip. The jumble of constants involving beta and gamma functions must reduce to the constant 1, because

$$\begin{aligned} 1 &= \mathbb{P}\{0 < U < 1, 0 < V < \infty\} \\ &= \iint \{0 < u < 1, 0 < v < \infty\} \psi(u, v) du dv \\ &= \int_0^1 g(u) du \int_0^\infty h(v) dv \frac{B(\alpha, \beta)\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \end{aligned}$$

Notice how the double integral has split into a product of two single integrals because the joint density factorized into a product of a function of  $u$  and a function of  $v$ . Both the single integrals equal 1 because both  $g$  and  $h$  are density functions. We have earned a bonus,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad \text{for } \alpha > 0 \text{ and } \beta > 0$$

which is a useful expression relating beta and gamma functions.

REMARK. The fact that  $\Gamma(1/2) = \sqrt{\pi}$  follows from the equality

$$\begin{aligned} \frac{\Gamma(1/2)\Gamma(1/2)}{\Gamma(1)} &= B(1/2, 1/2) = \int_0^1 t^{-1/2} (1 - t)^{-1/2} dt \quad \text{put } t = \sin^2(\theta) \\ &= \int_0^{\pi/2} \frac{1}{\sin(\theta) \cos(\theta)} 2 \sin(\theta) \cos(\theta) d\theta = \pi. \end{aligned}$$

The random variables  $U$  and  $V$  have joint density

$$\psi(u, v) = g(u)h(v) \quad \text{for } 0 < u < 1 \text{ and } 0 < v < \infty,$$

where  $g$  denotes the beta( $\alpha, \beta$ ) density, and  $h$  denotes the gamma( $\alpha + \beta$ ) density.

# EXAMPLE 51: SUMS OF INDEPENDENT GAMMA RANDOM VARIABLES

If  $X_1, X_2, \dots, X_k$  are independent random variables, with  $X_i$  distributed  $\text{gamma}(\alpha_i)$  for  $i = 1, \dots, k$ , then

$$X_1 + X_2 \sim \text{gamma}(\alpha_1 + \alpha_2),$$

$$X_1 + X_2 + X_3 = (X_1 + X_2) + X_3 \sim \text{gamma}(\alpha_1 + \alpha_2 + \alpha_3)$$

$$X_1 + X_2 + X_3 + X_4 = (X_1 + X_2 + X_3) + X_4 \sim \text{gamma}(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)$$

...

$$X_1 + X_2 + \dots + X_k \sim \text{gamma}(\alpha_1 + \alpha_2 + \dots + \alpha_k)$$

A particular case has great significance for Statistics. Suppose  $Z_1, \dots, Z_k$  are independent random variables, each distributed  $N(0,1)$ . From Chapter 9, the random variables  $Z_1^2/2, \dots, Z_k^2/2$  are independent  $\text{gamma}(1/2)$  distributed random variables. The sum

$$(Z_1^2 + \dots + Z_k^2)/2$$

must have a  $\text{gamma}(k/2)$  distribution with density  $t^{k/2-1}e^{-t}/\Gamma(k/2)$  for  $t > 0$ . The sum  $Z_1^2 + \dots + Z_k^2$  has density

$$\frac{(t/2)^{k/2-1}e^{-t/2}}{2\Gamma(k/2)} \quad \text{for } t > 0$$

This distribution is called the **chi-squared** on  $k$  degrees of freedom, usually denoted by  $\chi_k^2$ . The letter  $\chi$  is a lowercase Greek chi. □

EXAMPLE 48: JOINT DENSITIES FOR LINEAR COMBINATIONS

Suppose  $X$  and  $Y$  have a jointly continuous distribution with joint density  $f(x, y)$ . For constants  $a, b, c, d$ , define

$$U = aX + bY \quad \text{and} \quad V = cX + dY$$

Find the joint density function  $\psi(u, v)$  for  $(U, V)$ , under the assumption that the quantity  $\kappa = ad - bc$  is nonzero.

Think of the pair  $(U, V)$  as defining a new random point in  $\mathbb{R}^2$ . That is  $(U, V) = T(X, Y)$ , where  $T$  maps the point  $(x, y) \in \mathbb{R}^2$  to the point  $(u, v) \in \mathbb{R}^2$  with

$$u = ax + by \quad \text{and} \quad v = cx + dy,$$

or in matrix notation,

$$(u, v) = (x, y)A \quad \text{where} \quad A = \begin{pmatrix} a & c \\ b & d \end{pmatrix}.$$

Notice that  $\det A = ad - bc = \kappa$ . The assumption that  $\kappa \neq 0$  ensures that the transformation is invertible:

$$(u, v)A^{-1} = (x, y) \quad \text{where} \quad A^{-1} = \frac{1}{\kappa} \begin{pmatrix} d & -c \\ -b & a \end{pmatrix}.$$

That is,

$$\frac{du - bv}{\kappa} = x \quad \text{and} \quad \frac{-cu + av}{\kappa} = y.$$

Notice that  $\det(A^{-1}) = 1/\kappa = 1/(\det A)$ .

It helps to distinguish between the two roles for  $\mathbb{R}^2$ , referring to the domain of  $T$  as the  $(X, Y)$ -plane and the range as the  $(U, V)$ -plane.

The joint density function  $\psi(u, v)$  is characterized by the property that

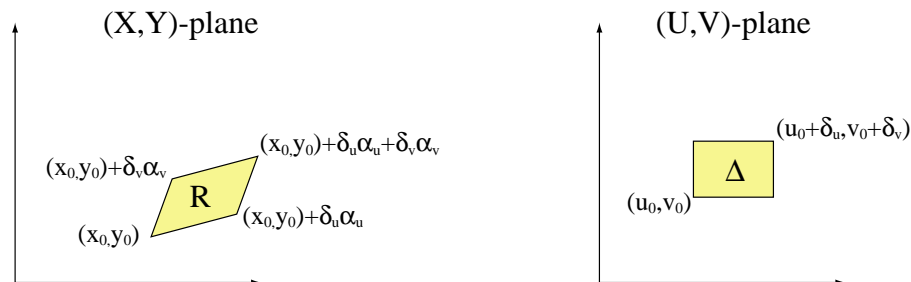
$$\mathbb{P}\{u_0 \leq U \leq u_0 + \delta_u, v_0 \leq V \leq v_0 + \delta_v\} \approx \psi(u_0, v_0)\delta_u\delta_v$$

for each  $(u_0, v_0)$  in the  $(U, V)$ -plane, and small  $(\delta_u, \delta_v)$ . To calculate the probability on the left-hand side we need to find the region  $R$  in the  $(X, Y)$ -plane corresponding to the small rectangle  $\Delta$ , with corners at  $(u_0, v_0)$  and  $(u_0 + \delta_u, v_0 + \delta_v)$ , in the  $(U, V)$ -plane.

The linear transformation  $A^{-1}$  maps parallel straight lines in the  $(U, V)$ -plane into parallel straight lines in the  $(X, Y)$ -plane. The region  $R$  must be a parallelogram, with vertices

$$\begin{aligned} (u_0, v_0 + \delta_v)A^{-1} & \quad \text{and} \quad (x_0 + \delta_x, y_0 + \delta_y) = (u_0 + \delta_u, v_0 + \delta_v)A^{-1} \\ (x_0, y_0) & = (u_0, v_0)A^{-1} \quad \text{and} \quad (u_0 + \delta_u, v_0)A^{-1} \end{aligned}$$

More succinctly,  $(\delta_x, \delta_y) = (\delta_u, \delta_v)A^{-1} = \delta_u\alpha_u + \delta_v\alpha_v$ , where  $\alpha_u = (d, -c)/\kappa$  and  $\alpha_v = (-b, a)/\kappa$  denote the two rows of  $A^{-1}$ .





From the formula in the Appendix to the Chapter, the parallelogram  $R$  has area

$$|\det(\delta_u \alpha'_u, \delta_v \alpha'_v)| = \delta_u \delta_v |\det(A^{-1})'| = \frac{\delta_u \delta_v}{|\det A|}.$$

For small  $\delta_u > 0$  and  $\delta_v > 0$ ,

$$\begin{aligned} \psi(u_0, v_0) \delta_u \delta_v &\approx \mathbb{P}\{(U, V) \in \Delta\} \\ &= \mathbb{P}\{(X, Y) \in R\} \\ &\approx (\text{area of } R) f(x_0, y_0) \\ &\approx \delta_u \delta_v f(x_0, y_0) / |\det(A)| \end{aligned}$$

It follows that  $(U, V)$  have joint density

$$\psi(u, v) = \frac{1}{|\det A|} f(x, y) \quad \text{where } (x, y) = (u, v)A^{-1}.$$

On the right-hand side you should substitute  $(du - bv)/\kappa$  for  $x$  and  $(-cu + av)/\kappa$  for  $y$ , in order to get an expression involving only  $u$  and  $v$ .

In effect, we have calculated a Jacobian by first principles. □

## Chapter 11

# Conditional densities

Density functions determine continuous distributions. If the distribution is calculated conditionally on some information, then the density is called a **conditional density**. When the conditioning information involves a random variable with a continuous distribution, the calculation of the conditional density involves arguments like those of Chapter 10.

Suppose  $X$  and  $Y$  have a jointly continuous distribution with joint density  $f(x, y)$ . From Chapter 10, we know that the marginal distribution of  $Y$  has density

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

The conditional distribution for  $X$  given  $Y = y$  has a (conditional) density, which I will denote by  $f_X(x | Y = y)$ , for which

$$\mathbb{P}\{x \leq X \leq x + \delta | Y = y\} \approx f_X(x | Y = y)\delta, \quad \text{if } \delta \text{ is small and positive.}$$

The subscript  $X$  on the  $f_X$  is intended to remind you that we are working with a density for the (conditional) distribution of  $X$ .

The conditioning information corresponds to an event  $\{Y = y\}$  with zero probability. An attempt to invoke the formula  $(\mathbb{P}A) \mathbb{P}(B | A) = \mathbb{P}(AB)$  would lead to the meaningless ratio

$$\frac{\mathbb{P}\{x \leq X \leq x + \delta, Y = y\}}{\mathbb{P}Y = y} \stackrel{?}{=} \frac{0}{0}.$$

Instead we must replace  $\{Y = y\}$  by an event  $\{y \leq Y \leq y + \epsilon\}$ , for a small  $\epsilon > 0$ , that provides almost the same conditioning information. Then

$$\begin{aligned} f_X(x | Y = y)\delta &\approx \mathbb{P}\{x \leq X \leq x + \delta | y \leq Y \leq y + \epsilon\} \\ &= \frac{\mathbb{P}\{x \leq X \leq x + \delta, y \leq Y \leq y + \epsilon\}}{\mathbb{P}\{y \leq Y \leq y + \epsilon\}} \approx \frac{f(x, y)\delta\epsilon}{g(y)\epsilon}. \end{aligned}$$

The  $\epsilon$  factors cancel. In the limit, as  $\delta$  and  $\epsilon$  tend to zero, we are left with

$$f_X(x | Y = y) = f(x, y)/g(y).$$

Symbolically,

$$\text{conditional density for } X \text{ given } \{Y = y\} = \frac{\text{joint density at } (x, y)}{\text{marginal density at } y}$$

---

**Example 52:** Let  $X$  and  $Y$  be independent random variables, each distributed  $N(0, 1)$ . Define  $R = \sqrt{X^2 + Y^2}$ . Show that, for each  $r > 0$ , the conditional distribution of  $X$  given  $R = r$  has density

$$f_X(x | R = r) = \frac{1}{\pi \sqrt{r^2 - x^2}} \quad \text{for } |x| < r \text{ and } r > 0.$$

The conditional density has a more intuitive interpretation via a transformation to polar coordinates:  $X = R \cos(\Theta)$  and  $Y = R \sin(\Theta)$ , where the polar angle  $\Theta$  is taken to lie in the range  $[0, 2\pi)$ . Remember from Chapter 9 that  $(X^2 + Y^2)/2$  has a gamma(1) distribution, with density  $e^{-t}$  for  $t > 0$ . That is,  $R^2/2$  has a standard exponential distribution. Thus, for  $|t| < 1$  and small  $\delta > 0$ ,

$$\begin{aligned}\mathbb{P}\{t \leq \cos(\Theta) \leq t + \delta \mid R = r\} &= \mathbb{P}\{rt \leq X \leq r(t + \delta) \mid R = r\} \\ &\approx (r\delta) f_X(rt \mid R = r) = \frac{1}{\pi \sqrt{1 - t^2}} \delta.\end{aligned}$$

That is, the conditional distribution of  $\cos(\Theta)$  does not depend on the value taken by  $R$ , which suggests that  $\Theta$  might be independent of  $R$ .

---

**Example 53:** Let  $W$  have a standard exponential distribution independent of  $\Theta$ , which is distributed Uniform $[0, 2\pi)$ . Put  $R = \sqrt{2W}$ . Show that the random variables  $X = R \cos \Theta$  and  $Y = R \sin \Theta$  are independent, with each variable having a  $N(0, 1)$  distribution.

---

In Example 52, we could have taken  $X = \sqrt{2W} \cos \Theta$  and  $Y = \sqrt{2W} \sin \Theta$ , with  $W$  and  $\Theta$  as in Example 53. Then  $X^2 + Y^2 = 2W$ , and the problem asks for the conditional distribution of  $\sqrt{2W} \cos(\Theta)$  given that  $W = r^2/2$ . The conditioning lets us put  $\sqrt{2W}$  equal to the constant  $r$ . The independence of  $W$  and  $\Theta$  lets us ignore the effects on  $\cos(\Theta)$  of the conditioning; the conditional density for  $\cos(\Theta)$  is the same as its marginal density, which you have already calculated in homework Problem 6.1 to be

$$\frac{1}{\pi \sqrt{1 - t^2}} \quad \text{for } |t| < 1,$$

in agreement with the calculation at the top of the page.

**REMARK.** The Box-Muller method generates independent  $N(0, 1)$  variates  $X_1$  and  $X_2$ , from two independent Uniform $(0, 1)$  variates,  $U_1$  and  $U_2$ , by

$$X_1 = \sqrt{-2 \log U_1} \cos(2\pi U_2) \quad \text{and} \quad X_2 = \sqrt{-2 \log U_1} \sin(2\pi U_2).$$

Why does the method work?

### Things to remember

- Suppose  $Y$  has a continuous distribution, with density  $f_Y(y)$ . For small positive  $\epsilon$ ,

$$\mathbb{E}(Z \mid Y = y) \approx \mathbb{E}(Z \mid y \leq Y \leq y + \epsilon),$$

for each random variable  $Z$ . With  $Z = h(Y)$ , a function of  $Y$ , we have used the approximation to calculate

$$\mathbb{E}h(Y) = \int_{-\infty}^{\infty} h(y) f_Y(y) dy.$$

With  $Z$  as the indicator function of an event  $\{x \leq X \leq x + \delta\}$ , we have used the approximation to find the conditional density  $f_X(x \mid Y = y)$ .

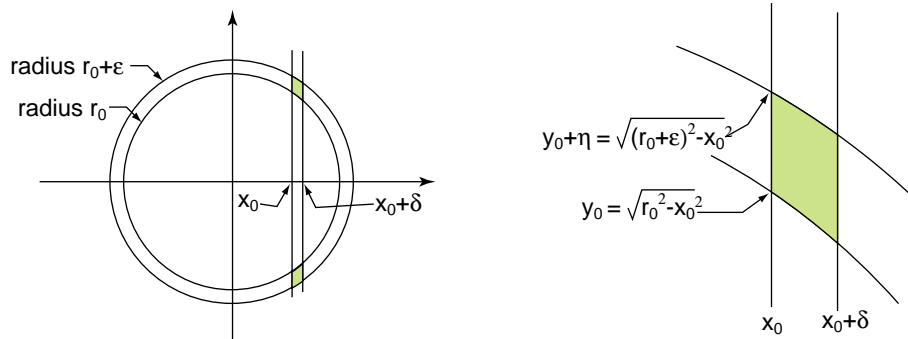
EXAMPLE 52: CONDITIONAL DENSITY FROM JOINT DENSITY

Let  $X$  and  $Y$  be independent random variables, each distributed  $N(0, 1)$ . Define  $R = \sqrt{X^2 + Y^2}$ . For each  $r > 0$ , find the density for the conditional distribution of  $X$  given  $R = r$ .

The joint density for  $(X, Y)$  equals

$$f(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right)$$

To find the joint density for  $X$  and  $R$ , calculate  $\mathbb{P}\{x_0 \leq X \leq x_0 + \delta, r_0 \leq R \leq r_0 + \epsilon\}$  for small, positive  $\delta$  and  $\epsilon$ . For  $|x_0| < r_0$ , the event corresponds to the two small regions in the  $(X, Y)$ -plane lying between the lines  $x = x_0$  and  $x = x_0 + \delta$ , and between the circles centered at the origin with radii  $r_0$  and  $r_0 + \epsilon$ .



By symmetry, both regions contribute the same probability. Consider the upper region. For small  $\delta$  and  $\epsilon$ , the region is approximately a parallelogram, with side length

$$\eta = \sqrt{(r_0 + \epsilon)^2 - x_0^2} - \sqrt{r_0^2 - x_0^2}$$

and width  $\delta$ . We could expand the expression for  $\eta$  as a power series in  $\epsilon$  by multiple applications of Taylor's theorem. It is easier to argue less directly, starting from the equalities

$$x_0^2 + (y_0 + \eta)^2 = (r_0 + \epsilon)^2 \quad \text{and} \quad x_0^2 + y_0^2 = r_0^2.$$

Expand the square on both sides of the first equality, discarding terms ( $\eta^2$  and  $\epsilon^2$ ) of smaller order, to get

$$x_0^2 + y_0^2 + 2\eta y_0 \approx r_0^2 + 2r_0\epsilon,$$

then invoke the second equality to deduce that  $\eta \approx (r_0\epsilon/y_0)$ . The upper region has approximate area  $r_0\epsilon\delta/y_0$ . Thus

$$\begin{aligned} \mathbb{P}\{x_0 \leq X \leq x_0 + \delta, r_0 \leq R \leq r_0 + \epsilon\} &= 2 \frac{r_0\epsilon\delta}{y_0} f(x_0, y_0) + \text{smaller order terms} \\ &\approx \frac{2r_0}{\sqrt{r_0^2 - x_0^2}} \frac{\exp(-r_0^2/2)}{2\pi} \epsilon\delta. \end{aligned}$$

The random variables  $X$  and  $R$  have joint density

$$\psi(x, r) = \frac{r \exp(-r^2/2)}{\pi \sqrt{r^2 - x^2}} \quad \text{for } |x| \leq r \text{ and } r > 0.$$

Once again I have omitted the subscript on the dummy variables, to indicate that the argument works for every  $x, r$  in the specified range.

The random variable  $R$  has marginal density

$$\begin{aligned} g(r) &= \int_{-r}^r \psi(x, r) dx = \frac{r \exp(-r^2/2)}{\pi} \int_{-r}^r \frac{dx}{\sqrt{r^2 - x^2}} \quad \text{put } x = r \cos \theta \\ &= \frac{r \exp(-r^2/2)}{\pi} \int_{\pi}^0 \frac{-r \sin \theta}{r \sin \theta} d\theta = r \exp(-r^2/2) \quad \text{for } r > 0. \end{aligned}$$

The conditional density equals

$$f_X(x \mid R = r) = \frac{\psi(x, r)}{g(r)} = \frac{1}{\pi \sqrt{r^2 - x^2}} \quad \text{for } |x| < r \text{ and } r > 0.$$

□

REMARK. For  $t > 0$  and small  $\delta > 0$ ,

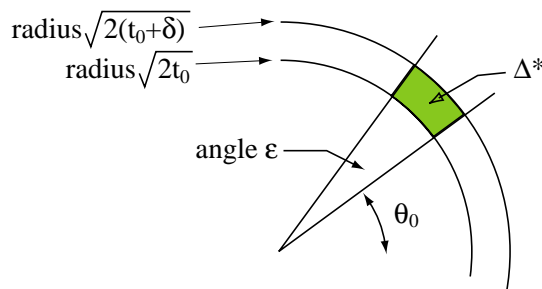
$$\begin{aligned} \mathbb{P}\{t \leq R^2/2 \leq t + \delta\} &= \mathbb{P}\{\sqrt{2t} \leq R \leq \sqrt{2t + 2\delta}\} \\ &\approx (\sqrt{2t + 2\delta} - \sqrt{2t}) g(\sqrt{2t}) \\ &\approx \frac{2\delta}{2\sqrt{2t}} \sqrt{2t} e^{-t}, \end{aligned}$$

showing that  $R^2/2$  has a standard exponential distribution, in agreement with the calculation in Example 51.

EXAMPLE 53: TRANSFORMATION TO POLAR COORDINATES

Let  $W$  have a standard exponential distribution independent of  $\Theta$ , which is distributed  $\text{Uniform}[0, 2\pi)$ . Put  $R = \sqrt{2W}$ . Show that the random variables  $X = R \cos \Theta$  and  $Y = R \sin \Theta$  are independent, with each variable having a  $N(0, 1)$  distribution.

The rectangle  $\Delta$  with corners  $(t_0, \theta_0)$ , and  $(t_0 + \delta, \theta_0 + \epsilon)$  in the  $(W, \Theta)$  strip corresponds to a region  $\Delta^*$  in the  $(X, Y)$ -plane bounded by radial lines at angles  $\theta_0$  and  $\theta_0 + \epsilon$  from the  $X$ -axis and two circles, of radii  $\sqrt{2t_0}$  and  $\sqrt{2(t_0 + \delta)}$ , centered at the origin.



Simple geometry will give the area of  $\Delta^*$ . (You might calculate the Jacobian as a cross-check.) The annular region between the two circles has area  $\pi 2(t_0 + \delta) - \pi(2t_0)$ . The two radial lines carve out a proportion  $\epsilon/(2\pi)$  of that area. That is,

$$\text{area of } \Delta^* = \frac{\epsilon}{2\pi} 2\pi \delta = \epsilon \delta.$$

The joint density  $f(x, y)$  for  $(X, Y)$  at the point  $(x_0, y_0) = (\sqrt{2t_0} \cos \theta_0, \sqrt{2t_0} \sin \theta_0)$  is given by

$$\begin{aligned} \epsilon \delta f(x_0, y_0) &\approx \mathbb{P}\{(X, Y) \in \Delta^*\} \\ &= \mathbb{P}\{\theta_0 \leq \Theta \leq \theta_0 + \epsilon, t_0 \leq W \leq t_0 + \delta\} \\ &= \mathbb{P}\{\theta_0 \leq \Theta \leq \theta_0 + \epsilon\} \mathbb{P}\{t_0 \leq W \leq t_0 + \delta\} && \text{by independence} \\ &\approx \frac{\epsilon}{2\pi} \delta \exp(-t_0) && \text{where } t_0 = \frac{x_0^2 + y_0^2}{2}. \end{aligned}$$

That is

$$f(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right).$$

The random variables  $X$ , and  $Y$  have the joint density of a pair of independent  $N(0, 1)$  distributed variates.  $\square$

## Chapter 12

# Multivariate normal distributions

The multivariate normal is the most useful, and most studied, of the standard joint distributions in probability. A huge body of statistical theory depends on the properties of families of random variables whose joint distributions are at least approximately multivariate normal. The bivariate case (two variables) is the easiest to understand, because it requires a minimum of notation. Vector notation and matrix algebra becomes necessities when many random variables are involved.

**Example 54:** The standard bivariate normal with correlation  $\rho$ ,

$$\psi(u, v) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{u^2 - 2\rho uv + v^2}{2(1-\rho^2)}\right) \quad \text{for all } u, v,$$

the joint density for  $U = X$  and  $V = \rho X + \sqrt{1-\rho^2} Y$ , where  $X$  and  $Y$  have independent  $N(0, 1)$  distributions and  $\rho$  is a constant with  $-1 < \rho < 1$ .

The construction of  $U$  and  $V$  from the independent random variables  $X$  and  $Y$  makes the calculation of the conditional distribution of  $V$  given  $U = u$  a triviality: the conditional distribution  $\rho X + \sqrt{1-\rho^2} Y \mid X = x$  is the same as the marginal distribution of the random variable  $\rho x + \sqrt{1-\rho^2} Y$ . That is,

$$V \mid U = u \sim N(\rho u, 1 - \rho^2).$$

Symmetry of the joint distribution of  $U$  and  $V$  implies that

$$U \mid V = v \sim N(\rho v, 1 - \rho^2),$$

a fact that you could check by explicit calculation of the ratio of joint to marginal densities,

$$\psi_U(u \mid V = v) = \psi(u, v) / \left( e^{-v^2/2} / \sqrt{2\pi} \right) = \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{(u - \rho v)^2}{2(1-\rho^2)}\right).$$

With various rescalings, we can manufacture more general bivariate normal distributions, involving five parameters. I feel it is much easier to think in terms of the standardized distribution.

**Definition.** Random variables  $S$  and  $T$  are said to have a bivariate normal distribution, with parameters  $\mathbb{E}S = \mu_S$ ,  $\mathbb{E}T = \mu_T$ ,  $\text{var}(S) = \sigma_S^2$ ,  $\text{var}(T) = \sigma_T^2$ , and correlation  $\rho$ , if the standardized random variables  $(S - \mu_S)/\sigma_S$  and  $(T - \mu_T)/\sigma_T$  have a standard bivariate normal distribution with correlation  $\rho$ .

The general bivariate normal is often used to model pairs of dependent random variables, such as: the height and weight of an individual; or (as an approximation) the score a student gets on a final exam and the total score she gets on the problem sets; or the heights of father and son; and so on. Many fancy statistical procedures implicitly require bivariate (or multivariate, for more than two random variables) normality.

**Example 55: (Regression to the mean)** Let  $S$  denote the height (in inches) of a randomly chosen father, and  $T$  denote the height (in inches) of his son at maturity. Suppose each of  $S$  and  $T$  has a  $N(\mu, \sigma^2)$  distribution with  $\mu = 69$  and  $\sigma = 2$ . Suppose also that  $S$  and  $T$  have a bivariate normal distribution with correlation  $\rho = .3$ .

If Ulysses has a height of 74 inches, what would one predict about the ultimate height of his son Victor?

The joint density  $f(x, y) = \exp(-(x^2 + y^2)/2)/(2\pi)$  for a pair of independent  $N(0, 1)$  random variables is radially symmetric, that is,  $f$  is a function of the radial distance  $\sqrt{x^2 + y^2}$ . This fact has far reaching consequences.

**Example 56:** Let  $Z_1$  and  $Z_2$  have independent  $N(0, 1)$  distributions, defining a random point  $\mathbf{Z} = (Z_1, Z_2)$  in the plane. Rotate the coordinate axes through an angle  $\alpha$ , writing  $(W_1, W_2)$  for the coordinates of the random point in the new coordinate system. Show that  $W_1$  and  $W_2$  are also independent  $N(0, 1)$  distributed random variables.

### More than two variables

When we deal with many variables  $X_1, X_2, \dots$  it becomes convenient to use vector notation, writing  $\mathbf{X}$  for the **random vector**  $(X_1, \dots, X_n)$ , and  $\mathbf{x}$  for the generic point  $(x_1, \dots, x_n)$  in  $\mathbb{R}^n$ .

**Definition.** Random variables  $X_1, X_2, \dots, X_n$  are said to have a jointly continuous distribution with joint density function  $f(x_1, x_2, \dots, x_n)$  if, for each subset  $A$  of  $\mathbb{R}^n$ ,

$$\begin{aligned}\mathbb{P}\{\mathbf{X} \in A\} &= \iint \dots \int \{(x_1, x_2, \dots, x_n) \in A\} f(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \\ &= \int \{\mathbf{x} \in A\} f(\mathbf{x}) d\mathbf{x},\end{aligned}$$

where  $\int \dots d\mathbf{x}$  is an abbreviation for the  $n$ -fold integral. For small regions  $\Delta$  containing a point  $\mathbf{x}^\circ$ , the probability  $\mathbb{P}\{\mathbf{X} \in \Delta\}$  is approximately  $\text{vol}(\Delta) \times f(\mathbf{x}^\circ)$ , where  $\text{vol}(\Delta)$  denotes the  $n$ -dimensional volume of  $\Delta$ .

The density  $f$  must be nonnegative and integrate to 1 over  $\mathbb{R}^n$ .

If the random variables  $X_1, \dots, X_n$  are independent, the joint density function is equal to the product of the marginal densities for each  $X_i$ , and conversely. The proof is similar to the proof for the bivariate case.

For example, if  $X_1, \dots, X_n$  are independent and each  $X_i$  has a  $N(0, 1)$  distribution, the joint density is

$$\begin{aligned}f(x_1, \dots, x_n) &= \frac{1}{(2\pi)^{n/2}} \exp\left(-\sum_{i=1}^n x_i^2/2\right) \quad \text{for all } x_1, \dots, x_n \\ &= \frac{1}{(2\pi)^{n/2}} \exp(-\|\mathbf{x}\|^2/2) \quad \text{for all } \mathbf{x}.\end{aligned}$$

This joint distribution is denoted by  $N(\mathbf{0}, I_n)$ . It is often referred to as the **spherical normal distribution**, because of the spherical symmetry of the density.

The distance of the random vector  $\mathbf{Z}$  from the origin is  $\|\mathbf{Z}\| = \sqrt{Z_1^2 + \dots + Z_n^2}$ . From Example 51, we know that  $\|\mathbf{Z}\|^2/2$  has a gamma( $n/2$ ) distribution. The distribution of  $\|\mathbf{Z}\|^2$  is given a special name, because of its great importance in the theory of statistics.



**Definition.** Let  $\mathbf{Z} = (Z_1, Z_2, \dots, Z_n)$  have a spherical normal distribution  $N(\mathbf{0}, I_n)$ . The **chi-square**,  $\chi_n^2$ , is defined as the distribution of  $\|\mathbf{Z}\|^2 = Z_1^2 + \dots + Z_n^2$ .

The methods for finding joint densities for random variables defined as functions of other random variables with jointly continuous distributions—as explained in the last two Chapters—extend to multivariate distributions. There is a problem with the drawing of pictures in  $n$  dimensions, to keep track of the transformations, and one must remember to say “ $n$ -dimensional volume” instead of area, but otherwise calculations are not much more complicated than in two dimensions.

The spherical symmetry of the density  $f(\cdot)$  is responsible for an important property of multivariate normals, the obvious analog of Example 56.

---

**Example 57:** Let  $\mathbf{q}_1, \dots, \mathbf{q}_n$  be a new orthonormal basis for  $\mathbb{R}^n$ , and let

$$\mathbf{Z} = W_1 \mathbf{q}_1 + \dots + W_n \mathbf{q}_n$$

be the representation for  $\mathbf{Z}$  in the new basis. Then the  $W_1, \dots, W_n$  are also independent  $N(0, 1)$  distributed random variables.

---

To prove results about the spherical normal it is often merely a matter of transforming to an appropriate orthonormal basis.

---

**Example 58:** Suppose  $Z_1, Z_2, \dots, Z_n$  are independent, each distributed  $N(0, 1)$ . Define  $\bar{Z} = (Z_1 + \dots + Z_n)/n$  and  $T = \sum_{i \leq n} (Z_i - \bar{Z})^2$ . Show that  $\bar{Z}$  has a  $N(0, 1/n)$  distribution independently of  $T$ , which has a  $\chi_{n-1}^2$  distribution.

---

EXAMPLE 54: THE STANDARD BIVARIATE NORMAL WITH CORRELATION  $\rho$

The most general bivariate normal can be built from a pair of independent random variables,  $X$  and  $Y$ , each distributed  $N(0, 1)$ . For a constant  $\rho$  with  $-1 < \rho < 1$ , define random variables

$$U = X \quad \text{and} \quad V = \rho X + \sqrt{1 - \rho^2} Y.$$

That is,

$$(U, V) = (X, Y)A \quad \text{where } A = \begin{pmatrix} 1 & \rho \\ 0 & \sqrt{1 - \rho^2} \end{pmatrix}.$$

Notice that  $\mathbb{E}U = \mathbb{E}V = 0$ , and

$$\text{var}(V) = \rho^2 \text{var}(X) + (1 - \rho^2) \text{var}(Y) = 1 = \text{var}(U),$$

and

$$\text{cov}(U, V) = \rho \text{cov}(X, X) + \sqrt{1 - \rho^2} \text{cov}(X, Y) = \rho.$$

Consequently,

$$\text{correlation}(U, V) = \text{cov}(U, V) / \sqrt{\text{var}(U)\text{var}(V)} = \rho.$$

From Example 48, the joint density for  $(U, V)$  is

$$\psi(u, v) = \frac{1}{|\det A|} f((u, v)A^{-1}),$$

where

$$f(x, y) = \frac{1}{2\pi} \exp\left(-\frac{x^2 + y^2}{2}\right) \quad \text{all } x, y.$$

The matrix  $A$  has determinant  $\sqrt{1 - \rho^2}$  and inverse

$$A^{-1} = \begin{pmatrix} \sqrt{1 - \rho^2} & -\rho \\ 0 & 1 \end{pmatrix} / \sqrt{1 - \rho^2}$$

If  $(x, y) = (u, v)A^{-1}$  then

$$\begin{aligned} x^2 + y^2 &= (u, v)A^{-1}(A^{-1})'(u, v)' \\ &= (u, v) \begin{pmatrix} 1 & -\rho \\ -\rho & 0 \end{pmatrix} (u, v)' / (1 - \rho^2) \\ &= \frac{u^2 - 2\rho uv + v^2}{1 - \rho^2}. \end{aligned}$$

Thus  $U$  and  $V$  have joint density

$$\psi(u, v) = \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left(-\frac{u^2 - 2\rho uv + v^2}{2(1 - \rho^2)}\right) \quad \text{for all } u, v.$$

The joint distribution is often called the **standard bivariate normal** distribution with correlation  $\rho$ .

The symmetry of  $\psi$  in  $u$  and  $v$  implies that  $V$  has the same marginal distribution as  $U$ , that is,  $V$  is also  $N(0, 1)$  distributed. The calculation of the marginals densities involves the same integration for both variables.  $\square$

REMARK. When  $\rho$  equals zero, the joint density for  $U$  and  $V$  factorizes into

$$\frac{1}{\sqrt{2\pi}} \exp(-u^2/2) \frac{1}{\sqrt{2\pi}} \exp(-v^2/2)$$

which implies independence of  $U$  and  $V$ . That is, for random variables with a bivariate normal distribution, zero correlation is equivalent to independence. This equivalence for bivariate normals probably accounts for the widespread confusion between the properties of independence and zero correlation. In general, independence implies zero correlation, but not conversely.

# EXAMPLE 55: REGRESSION TO THE MEAN

Let  $S$  denote the height (in inches) of a randomly chosen father, and  $T$  denote the height (in inches) of his son at maturity. Suppose each of  $S$  and  $T$  has a  $N(\mu, \sigma^2)$  distribution with  $\mu = 69$  and  $\sigma = 2$ . Suppose also that  $S$  and  $T$  have a bivariate normal distribution with correlation  $\rho = .3$ .

If Ulysses has a height of 74 inches, what would one predict about the ultimate height of his son Victor?

In standardized units, Ulysses has height  $U = (S - \mu)/\sigma$ , which we are given to equal 2.5. Victor's ultimate standardized height is  $V = (T - \mu)/\sigma$ . By assumption, before the value of  $U$  was known, the pair  $(U, V)$  has a standard bivariate normal distribution with correlation  $\rho$ . The conditional distribution of  $V$  given that  $U = 2.5$  is

$$V \mid U = 2.5 \sim N(2.5\rho, 1 - \rho^2)$$

In the original units, the conditional distribution of  $T$  given  $S = 74$  is normal with mean  $\mu + 2.5\rho\sigma$  and variance  $(1 - \rho^2)\sigma^2$ , that is,

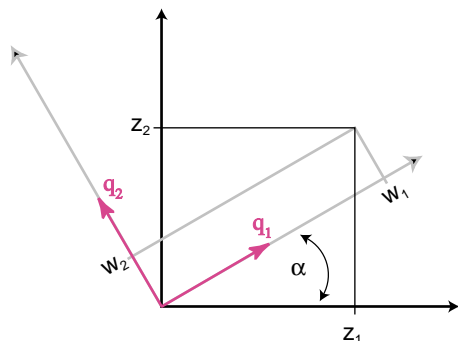
$$\text{Victor's ultimate height} \mid \text{Ulysses's height} = 74 \text{ inches} \sim N(70.5, 3.64)$$

If I had to make a guess, I would predict that Victor would ultimately reach a height of 70.5 inches. □

REMARK. Notice that Victor's expected height (given that Ulysses is 74 inches) is less than his father's height. This fact is an example of a general phenomenon called "regression towards the mean". The term **regression**, as a synonym for conditional expectation, has become commonplace in Statistics.

# EXAMPLE 56: ROTATION OF COORDINATE AXES

Let  $Z_1$  and  $Z_2$  have independent  $N(0, 1)$  distributions, defining a random point  $\mathbf{Z} = (Z_1, Z_2)$  in the plane. Rotate the coordinate axes through an angle  $\alpha$ , writing  $(W_1, W_2)$  for the coordinates of the random point in the new coordinate system. Show that  $W_1$  and  $W_2$  are also independent  $N(0, 1)$  distributed random variables.



The new axes are defined by the unit vectors

$$\mathbf{q}_1 = (\cos \alpha, \sin \alpha) \quad \text{and} \quad \mathbf{q}_2 = (-\sin \alpha, \cos \alpha).$$

From the representation

$$\mathbf{Z} = (Z_1, Z_2) = W_1 \mathbf{q}_1 + W_2 \mathbf{q}_2$$

we get

$$W_1 = \mathbf{Z} \cdot \mathbf{q}_1 = Z_1 \cos \alpha + Z_2 \sin \alpha \quad \text{and} \quad W_2 = \mathbf{Z} \cdot \mathbf{q}_2 = -Z_1 \sin \alpha + Z_2 \cos \alpha.$$

More succinctly,

$$(W_1, W_2) = (Z_1, Z_2) A_\alpha \quad \text{where} \quad A_\alpha = \begin{pmatrix} \cos \alpha & -\sin \alpha \\ \sin \alpha & \cos \alpha \end{pmatrix}.$$

The joint density for  $W_1$  and  $W_2$  again comes from the formula derived in Example 48. The matrix  $A_\alpha$  has determinant 1 and inverse  $A_{-\alpha}$ . It is an orthogonal matrix; it preserves lengths. The joint density of  $(W_1, W_2)$  is

$$\frac{1}{2\pi} \exp(-\|(w_1, w_2) A^{-1}\|^2/2) = \frac{1}{2\pi} \exp(-(w_1^2 + w_2^2)/2),$$

the product of two marginal  $N(0, 1)$  densities.

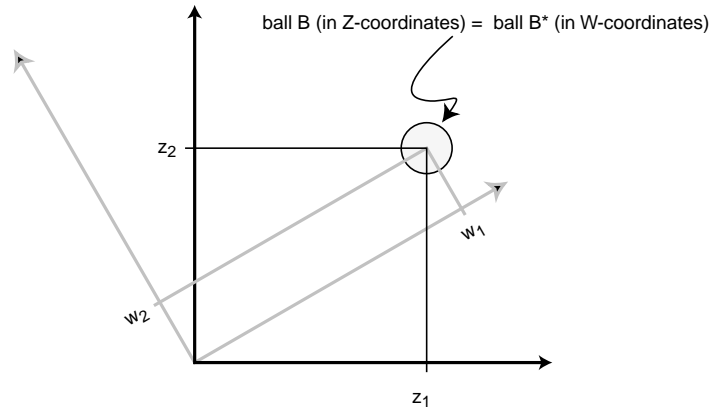
EXAMPLE 57: ROTATION TO NEW COORDINATES: MULTIVARIATE CASE

Let  $\mathbf{q}_1, \dots, \mathbf{q}_n$  be a new orthonormal basis for  $\mathbb{R}^n$ , and let

$$\mathbf{Z} = W_1 \mathbf{q}_1 + \dots + W_n \mathbf{q}_n$$

be the representation for  $\mathbf{Z}$  in the new basis. Then the  $W_1, \dots, W_n$  are also independent  $N(0, 1)$  distributed random variables.

The picture shows only two of the  $n$  coordinates:



For a small ball  $B$  centered at  $\mathbf{z}$ ,

$$\mathbb{P}\{\mathbf{Z} \in B\} \approx f(\mathbf{z})(\text{volume of } B).$$

The corresponding region for  $\mathbf{W}$  is  $B^*$ , a ball of the same radius, but centered at the point  $\mathbf{w} = (w_1, \dots, w_n)$  for which  $w_1 \mathbf{q}_1 + \dots + w_n \mathbf{q}_n = \mathbf{z}$ . Thus

$$\mathbb{P}\{\mathbf{W} \in B^*\} = \mathbb{P}\{\mathbf{Z} \in B\} \approx (2\pi)^{-n/2} \exp(-\frac{1}{2}\|\mathbf{x}\|^2)(\text{volume of } B).$$

From the equalities

$$\|\mathbf{w}\| = \|\mathbf{z}\| \quad \text{and} \quad \text{volume of } B = \text{volume of } B^*,$$

we get

$$\mathbb{P}\{\mathbf{W} \in B^*\} \approx (2\pi)^{-n/2} \exp(-\frac{1}{2}\|\mathbf{w}\|^2)(\text{volume of } B^*).$$

That is,  $\mathbf{W}$  has the asserted spherical normal density.

EXAMPLE 58: INDEPENDENCE OF SAMPLE MEAN AND SAMPLE VARIANCE

Suppose  $Z_1, Z_2, \dots, Z_n$  are independent, each distributed  $N(0, 1)$ . Define

$$\bar{Z} = \frac{Z_1 + \dots + Z_n}{n} \quad \text{and} \quad T = \sum_{i \leq n} (Z_i - \bar{Z})^2$$

Show that  $\bar{Z}$  has a  $N(0, 1/n)$  distribution independently of  $T$ , which has a  $\chi_{n-1}^2$  distribution.

Choose the new orthonormal basis with  $\mathbf{q}_1 = (1, 1, \dots, 1)/\sqrt{n}$ . Choose  $\mathbf{q}_2, \dots, \mathbf{q}_n$  however you like, provided they are orthogonal unit vectors, all orthogonal to  $\mathbf{q}_1$ . In the new coordinate system,

$$\mathbf{Z} = W_1 \mathbf{q}_1 + \dots + W_n \mathbf{q}_n \quad \text{where } W_i = \mathbf{Z} \cdot \mathbf{q}_i \text{ for each } i.$$

In particular,

$$W_1 = \mathbf{Z} \cdot \mathbf{q}_1 = \frac{Z_1 + \dots + Z_n}{\sqrt{n}} = \sqrt{n} \bar{Z}$$

From Example 57 we know that  $W_1$  has a  $N(0, 1)$  distribution. It follows that  $\bar{Z}$  has a  $N(0, 1/n)$  distribution.

The random variable  $T$  equals the squared length of the vector

$$(Z_1 - \bar{Z}, \dots, Z_n - \bar{Z}) = \mathbf{Z} - \bar{Z}(\sqrt{n}\mathbf{q}_1) = \mathbf{Z} - W_1 \mathbf{q}_1 = W_2 \mathbf{q}_2 + \dots + W_n \mathbf{q}_n.$$

That is,

$$T = \|W_2 \mathbf{q}_2 + \dots + W_n \mathbf{q}_n\|^2 = W_2^2 + \dots + W_n^2,$$

a sum of squares of  $n - 1$  independent  $N(0, 1)$  random variables, which has a  $\chi_{n-1}^2$  distribution.

Finally, notice that  $\bar{Z}$  is a function of  $W_1$ , whereas  $T$  is a function of the independent random variables  $W_2, \dots, W_n$ . The independence of  $\bar{Z}$  and  $T$  follows.  $\square$

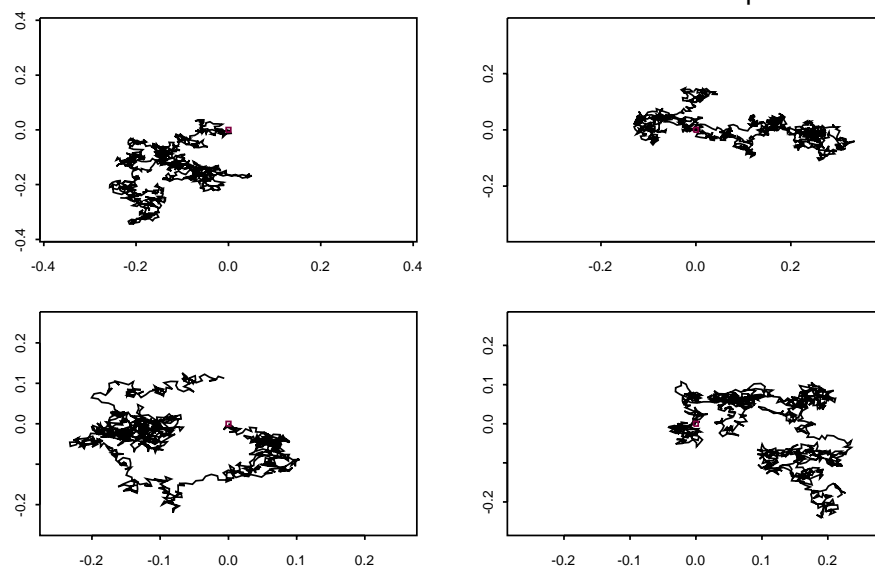
## Chapter 13

# Brownian motion

Apparently, when little particles suspended in water are observed under a microscope, they are seen to jiggle around in an irregular motion. The British botanist Robert Brown discovered the phenomenon in the early nineteenth century. Three quarters of a century later, Albert Einstein explained the motion as the consequence of large number of random impacts from molecules; the displacement over time is the accumulation of a large number of small, independent nudges from molecules that are miniscule in size compared with the particle.

To give you the idea of the sort of irregular motion that results from a large number of small independent increments, I have simulated (four times) the process using sequences of independent observations  $\Theta_i$  on the  $\text{Uniform}(0, 2\pi)$  distribution, taking the  $\delta \cos(\Theta_i)$  as the increments in the  $x$ -direction, and the  $\delta \sin(\Theta_i)$  as the increments in the  $y$ -direction, for a small  $\delta$ . That is, after  $n$  steps the particle is at position  $(X_n, Y_n)$ , where  $X_n = \delta \cos(\Theta_1) + \delta \cos(\Theta_2) + \dots + \delta \cos(\Theta_n)$  and  $Y_n = \delta \sin(\Theta_1) + \delta \sin(\Theta_2) + \dots + \delta \sin(\Theta_n)$ .

Simulated 2-d Brownian Motions with 1000 steps



In general, symmetry of the directions in which the small increments act, and a central limit effect (via the bivariate form of the CLT) suggest that the displacement  $(X_t, Y_t)$  of the particle after time  $t$  should have a  $N(0, \sigma_t^2 I_2)$  bivariate normal distribution. That is,  $X_t$  should be distributed  $N(0, \sigma_t^2)$  independently of  $Y_t$ , which also has a  $N(0, \sigma_t^2)$  distribution.

Homogeneity and independence arguments will show that  $\sigma_t^2$  must be proportional to  $t$ . The **increment**  $X_{t+s} - X_t$  represents the change in  $x$ -coordinate over a time interval of length  $s$ , and therefore it should have the same normal distribution as  $X_s$ . That is,  $X_{t+s} - X_t \sim N(0, \sigma_s^2)$  independently of  $X_t \sim N(0, \sigma_t^2)$ . By independence,

$$\sigma_{s+t}^2 = \text{var}(X_{t+s}) = \text{var}(X_{t+s} - X_t) + \text{var}(X_t) = \sigma_s^2 + \sigma_t^2.$$

For such an equality to hold for all  $s \geq 0$  and  $t \geq 0$  we must have  $\sigma_t^2 = ct$  for some positive constant  $c$ .

A family of random variables indexed by a parameter  $t$  that ranges over a continuous interval is usually called a **stochastic process** (in continuous time).

**Definition.** A stochastic process  $\{B_t : t \geq 0\}$  is called a **Brownian motion** (started from zero) if

- (i)  $B_0 = 0$
- (ii)  $B_{t+s} - B_t \sim N(0, s)$  for each  $t \geq 0$  and  $s \geq 0$
- (iii) the increments of the process over disjoint time intervals are independent
- (iv) the process changes continuously with  $t$

The symmetric random walk process is obtained by repeated tosses of a fair coin, with a move of one unit to the right for each head, and a move of one unit to the left for each tail. Brownian motion can be thought of as a continuous time analog of the symmetric random walk.

---

**Example 59:** For positive constants  $\alpha$  and  $\beta$ , what is the probability that a Brownian motion started at zero reaches  $\beta$  before it reaches  $-\alpha$ ?

---

Many interesting stochastic processes have been built from Brownian motion. In recent years, perhaps the most famous example has been a model for the changes in a stock price over time. Interpretation of the model brings out a surprising property about the increments of Brownian motion.

---

**Example 60:** A model for stock prices

---

Under the model from the previous Example, it is possible to reduce problems in the pricing of options to the solution of partial differential equations. The method depends heavily on the properties of Brownian motion.

---

**Example 61:** The Black-Scholes differential equation

---



# EXAMPLE 59: CONTINUOUS ANALOG OF GAMBLER'S RUIN PROBLEM

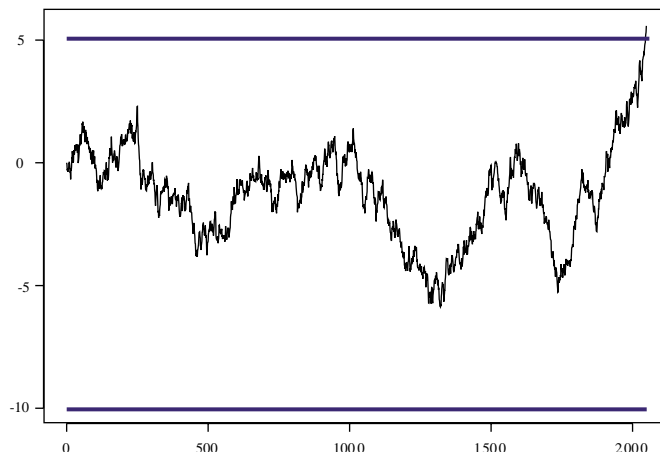
The following problem was solved in Example 12:

Suppose two players, Alf (A for short) and Betamax (B for short), bet on the tosses of a fair coin: for a head, Alf pays Betamax one dollar; for a tail, Betamax pays Alf one dollar. They stop playing when one player runs out of money. If Alf starts with  $\alpha$  dollar bills, and Betamax starts with  $\beta$  dollars bills (both  $\alpha$  and  $\beta$  whole numbers), what is the probability that Alf ends up with all the money?

The solution was:  $\mathbb{P}\{\text{Alf wins}\} = \alpha/(\alpha + \beta)$ . There is an analogous problem for Brownian motion, with an analogous solution:

*For positive constants  $\alpha$  and  $\beta$ , what is the probability that a Brownian motion started at zero reaches  $\beta$  before it reaches  $-\alpha$ ?*

For the picture, I take  $\alpha = 10$  and  $\beta = 5$ , even though there was no need for either of them to be integers.



For each  $x$  in  $[-\alpha, \beta]$ , consider the analogous problem for a Brownian motion started at  $x$  (that is, for the process  $x + B_t$ ). Define

$$f(x) = \mathbb{P}\{x + B_t \text{ hits } \beta \text{ before } -\alpha\}$$

That is, if we define two hitting times

$$\tau_\beta = \min\{t : x + B_t = \beta\} \quad \text{and} \quad \tau_{-\alpha} = \min\{t : x + B_t = -\alpha\},$$

then  $f(x) = \mathbb{P}\{\tau_\beta < \tau_{-\alpha}\}$ .

For trivial reasons,  $f(\beta) = 1$  and  $f(-\alpha) = 0$ . For a small  $\delta > 0$ , write  $F$  for the event  $\{x + B_t \text{ hits } x + \delta \text{ before } x - \delta\}$ . By symmetry,  $\mathbb{P}F = \mathbb{P}F^c = 1/2$ . Conditioning on whether  $F$  or  $F^c$  occurs, we get

$$f(x) = \mathbb{P}\{x + B_t \text{ hits } \beta \text{ before } -\alpha \mid F\} \frac{1}{2} + \mathbb{P}\{x + B_t \text{ hits } \beta \text{ before } -\alpha \mid F^c\} \frac{1}{2}.$$

The conditional given  $F$  corresponds to starting the Brownian motion at  $x + \delta$ ; the conditional probability equals  $f(x + \delta)$ . Similarly, the conditional probability given  $F^c$  equals  $f(x - \delta)$ . Thus

$$f(x) = \frac{1}{2}f(x + \delta) + \frac{1}{2}f(x - \delta),$$

for arbitrarily small  $\delta$ . The right-hand side equals

$$\begin{aligned} & \frac{1}{2} (f(x) + \delta f'(x) + \frac{1}{2}\delta^2 f''(x) + \dots) + \frac{1}{2} (f(x) - \delta f'(x) + \frac{1}{2}\delta^2 f''(x) - \dots) \\ &= f(x) + \frac{1}{2}\delta^2 f''(x) + \dots \end{aligned}$$

That is, the function  $f$  satisfies the differential equation  $f''(x) = 0$  for  $-\alpha < x < \beta$ , with boundary conditions  $f(-\alpha) = 0$  and  $f(\beta) = 1$ , which has solution

$$f(x) = \frac{x + \alpha}{\beta + \alpha} \quad \text{for } -\alpha \leq x \leq \beta.$$

In particular, for the problem as originally posed, there is probability  $\alpha/(\beta + \alpha)$  that the Brownian motion reaches  $\beta$  before it reaches  $-\alpha$ .

□

# EXAMPLE 60: A MODEL FOR STOCK PRICES

Let  $S_t$  denote the price at time  $t \geq 0$  of one share in a particular company, starting from a given price  $S_0$  at time  $t = 0$ . One standard theory models the process  $\{S_t : t \geq 0\}$  by means of an underlying Brownian motion  $\{B_t : t \geq 0\}$  and two parameters,  $\mu$  and  $\sigma > 0$ ,

$$S_t = S_0 \exp\left(\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma B_t\right) \quad \text{for } t \geq 0.$$

The strange form of the parametrization will make more sense if we consider the increment  $\Delta S = S_{t+\delta} - S_t$  in the stock price over a small time interval  $[t, t + \delta]$  as a proportion of the price at time  $t$ .

Write  $\Delta B$  for the corresponding increment  $B_{t+\delta} - B_t$  in the underlying Brownian motion. Remember that  $\Delta B \sim N(0, \delta)$ , so that  $\mathbb{E}(\Delta B)^2 = \delta$  and  $\mathbb{E}\left((\Delta B)^2 - \delta\right)^2 = 2\delta^2$ . (You should check these calculations.)

Recall the Taylor expansion

$$e^x = 1 + x + \frac{1}{2}x^2 + \dots$$

As you will see, when dealing with Brownian motion we must carry the expansion out to the quadratic term. Temporarily write  $\kappa$  for  $\mu - \frac{1}{2}\sigma^2$ . From the equality

$$\begin{aligned} S_{t+\delta} &= S_0 \exp(\kappa(t + \delta) + \sigma(B_t + \Delta B)) \\ &= S_t \exp(\kappa\delta + \sigma\Delta B) \end{aligned}$$

and the expansion for  $e^x$  with  $x$  replaced by  $\kappa\delta + \sigma\Delta B$  we get

$$\begin{aligned} \frac{\Delta S}{S_t} &= \exp(\kappa\delta + \sigma\Delta B) - 1 \\ &= \kappa\delta + \sigma\Delta B + \frac{1}{2}(\kappa\delta + \sigma\Delta B)^2 + \dots \\ &= \mu\delta + \sigma\Delta B + \frac{1}{2}\sigma^2((\Delta B)^2 - \delta) + \frac{1}{2}\kappa^2\delta^2 + \kappa\sigma\delta\Delta B + \dots \end{aligned}$$

Notice how the  $-\frac{1}{2}\sigma^2\delta$  has contributed the term to “center” the  $(\Delta B)^2$  by subtracting off its expected value. The centered quantity  $(\Delta B)^2 - \delta$ , which I will write as  $\Delta Q$ , has zero expected value and variance  $2\delta^2$ . The increment  $\sigma\Delta B$  also has zero expected value, but its variance equals  $\sigma^2\delta$ .

To determine the change in the stock price between two times,  $t_0$  and  $t_1$ , we could divide  $[t_0, t_1]$  into small subintervals of length  $\delta$ , then add up the  $\Delta S$  increments as a weighted sum of terms like

$$\mu S_t \delta + \sigma S_t \Delta B + \frac{1}{2}\sigma^2 S_t \Delta Q + \dots$$

Both the  $\sigma S_t \Delta B$  and the  $\frac{1}{2}\sigma^2 S_t \Delta Q$  contribute sums with zero expected value. For the second sum the variance goes to zero as  $\delta$  is made smaller, but the variance for the first sum does not go to zero. (A similar argument kills off the other random contributions.) In the limit, only the sum of the  $\mu S_t \delta + \sigma S_t \Delta B$  terms survives.

In summary: When dealing with approximations over short time intervals, we may ignore all except the drift term  $\mu\delta$  and the noise term  $\sigma\Delta B$ ,

$$\frac{\Delta S}{S_t} \approx \mu\delta + \sigma\Delta B.$$

Conditional on  $S_t$ , the relative increment has approximately a  $N(\mu\delta, \sigma^2\delta)$  distribution. Moreover, when adding up increments (and then passing to a limit), we may replace terms  $(\Delta B)^2$  by their expected values. □

REMARK. The model is sometimes written in symbolic form as  $dS_t = \mu dt + \sigma dB_t$ .

EXAMPLE 61: THE BLACK-SCHOLES DIFFERENTIAL EQUATION

Consider a stock whose price  $S_t$  at time  $t$ , for  $0 \leq t \leq 1$ , is driven by the process described in Example 60, that is,  $\Delta S \approx S_t(\mu\delta + \sigma\Delta B)$ . Assume that  $\sigma$  is known, but  $\mu$  is not. That is, assume we know the volatility but not the drift.

Suppose  $Y$  is a random quantity determined by the stock price  $S_1$  at time  $t = 1$ . I will give you the option of receiving the amount  $Y$  at time  $t = 1$  if you pay me an amount  $y_0$  at time 0. More adventurously, I specify the fair price  $Y_t$  to pay at time  $t$ , for  $0 \leq t \leq 1$ , in order to receive the amount  $Y$  at time  $t = 1$ . The price is given by a function  $Y_t = f(S_t, t)$  of the stock price at time  $t$  and of  $t$  (or, equivalently, of  $S_t$  and  $1 - t$ , the time remaining before the payoff). How should I determine the function  $f$ ?

There are various reasonable properties to require of  $f$ . For example, as  $t$  increases to 1, the price  $Y_t$  should converge to  $Y$ , for otherwise someone stands to make a nearly riskless profit by trading one blink of an eye before time  $t = 1$ . How?

There is another constraint that is less obvious. It comes from the Taylor expansion

$$f(x + \epsilon, t + \delta) \approx \delta h(x, t) + \epsilon g_1(x, t) + \frac{1}{2}\epsilon^2 g_2(x, t),$$

where, for simplicity of notation, I write  $h(x, t)$  for  $\partial f / \partial t$ , and  $g_1(x, t)$  for  $\partial f / \partial x$ , and  $g_2(x, t)$  for  $\partial^2 f / \partial^2 x$ . First consider how the option price responds to a small change in the stock price between time  $t$  and  $t + \delta$ .

$$Y_{t+\delta} = f(S_t + \Delta S, t + \delta) \approx f(S_t, t) + \delta h(S_t, t) + (\Delta S) g_1(S_t, t) + \frac{1}{2} (\Delta S)^2 g_2(S_t, t)$$

From the model,

$$(\Delta S)^2 \approx S_t^2 (\mu\delta + \sigma\Delta B)^2 \approx S_t^2 \sigma^2 (\Delta B)^2 + \dots$$

Remember that we may replace the  $(\Delta B)^2$  by a  $\delta$  when adding up the effect of changes over many small time intervals. That is, we may calculate as if

$$\begin{aligned} \Delta Y &\approx h(S_t, t)\delta + (\Delta S) g_1(S_t, t) + \frac{1}{2}\sigma^2 S_t^2 g_2(S_t, t)\delta \\ &= F(S_t, t)\delta + (\Delta S) g_1(S_t, t), \end{aligned}$$

where  $F(x, t) = h(x, t) + \frac{1}{2}\sigma^2 x^2 g_2(x, t)$ .

The function  $F$  plays an important role in determining the profit from a trading scheme designed to exploit any inconsistency in the pricing of the option. The scheme involves continuous trading, defined by some function  $p(S_t, t)$  (to be specified soon) of the stock price and time. Carry out the following trades:

- (i) purchase  $p(S_t, t)$  options at the price  $p(S_t, t)Y_t$  at time  $t$  then sell them at time  $t + \delta$  for the return  $p(S_t, t)(Y_t + \Delta Y)$
- (ii) purchase  $-p(S_t, t)g_1(S_t, t)$  stocks at the price  $-p(S_t, t)g_1(S_t, t)S_t$  at time  $t$  then sell them at time  $t + \delta$  for the return  $-p(S_t, t)g_1(S_t, t)(S_t + \Delta S)$ .

The net profit (possibly negative) from the trades is

$$p(S_t, t)\Delta Y - p(S_t, t)g_1(S_t, t)(\Delta S) \approx p(S_t, t)F(S_t, t)\delta.$$

Notice how the stock trade cancels out the noisy part of the option trade.

If we choose  $|p(x, t)|$  large and with same sign as  $F(x, t)$ , we slowly accumulate risk free profit over the times when  $F(S_t, t) \neq 0$ . To avoid such exploitation, I should ensure that my option pricing function  $f(x, t)$  is such that  $F(x, t) = 0$ . That is, I should be careful that

$$\frac{\partial f}{\partial t} + \frac{1}{2}\sigma^2 x^2 \frac{\partial^2 f}{\partial^2 x} = 0.$$

Together with various boundary conditions designed to eliminate other schemes for making a riskless profit at my expense, this differential equation can be used to find the function  $f$  that defines the option prices. □

To learn more about this story, you could take Statistics 251/551, *Stochastic Processes*.