Chapter 1

Probabilities and random variables

1.1 Overview

Probability theory provides 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 one) 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

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

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.

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". If we wish to contemplate such events we need a more intricate sample space S. Indeed, the choice of S—the detail with which possible outcomes are described—depends on the sort of events we wish to study.

In general, a sample space can make it easier to think precisely about events, but it is not always essential. 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 one) is as a long run frequency: *in a very large number (N) of repetitions of the experiment,*

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

More about independence soon.

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 the Appendix to Chapter 2 for an alternative interpretation, based on fair prices.

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

Rules for probabilities

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

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

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

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

Example <3.4> Find $\mathbb{P}\{$at least two heads$\}$ for the tossing of three coins.

Note: The examples are collected together at the end of each chapter

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 ...*

> **Remark.** If $A = \{\text{at least two heads}\}$ and info denotes the assumptions (coins fair, “independence,” ...) the last display makes an assertion about $\mathbb{P}(A \mid \text{info})$. The symbol $\mathbb{P}\cdot \mid \text{info})$ denotes the conditional probability given the information; it is NOT the probability of a conditional event. I regard “$A \mid \text{info}$” without the $\mathbb{P}$ as meaningless.

If the conditioning information is held fixed throughout a calculation, 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. In that case 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 in any of the following ways:

$$\begin{aligned}
&\mathbb{P}(\dots \mid \text{ event } A \text{ \underline{and} event } B \text{ have occurred \underline{and} previous info})\\
&\mathbb{P}(\dots \mid A\cap B \text{ \underline{and} previous info}) \qquad \text{where } \cap \text{ means intersection}\\
&\mathbb{P}(\dots \mid A,\ B, \text{ previous info})\\
&\mathbb{P}(\dots \mid A\cap B) \text{ or } \mathbb{P}(\dots \mid AB) \qquad \text{if “previous info” is understood.}
\end{aligned}$$

if the “previous info” is understood. I often write AB instead of $A \cap B$ for an intersection of two sets. 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 $\mathbb{P}\{\cdots \mid \text{info}\}$ 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 and $A \cap B$ occurs $N_{A\cap B}$ times. Then, for N large,

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

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

Remark. Many textbooks *define* $\mathbb{P}(B \mid A)$ as the ratio $\mathbb{P}(BA)/\mathbb{P}A$, which is just a rearrangement of (P5) without the info. That definition, not surprisingly, gives students the idea that conditional probabilities are always determined by taking ratios, which is not true. Often the assignment of conditional probabilities is part of the modelling. See Example <1.3> for example.

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 <1.2> 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 statement of the next example is taken verbatim from the delightful *Fifty Challenging Problems in Probability* by Frederick Mosteller, one of my favourite sources for elegant examples. One could learn a lot of probability by trying to solve all fifty problems. The underlying question has resurfaced in recent years in various guises. See

http://en.wikipedia.org/wiki/Monty_Hall_problem

http://en.wikipedia.org/wiki/Marilyn_vos_Savant#The_Monty_Hall_problem

to understand why probabilistic notation is so valuable. The lesson is: Be prepared to defend your assignments of conditional probabilities.

Example <1.3> 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 an explicit listing of the sample space specification of a sample space. In fact that is seldom the case. 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 construction of a sample space would be a nontrivial exercise but conditioning helps to break a complex random mechanism into a sequence of simpler stages.

Example <1.4> 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 <1.3> and <1.4> we had situations where particular pieces of information could be ignored in the calculation of some conditional probabilities,

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

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

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

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

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

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

for conditionally independent events E and F. Except for the conditioning information, the last equality 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 <1.5> 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, show that the probability that the first head appears on the kth toss is $(1-p)^{k-1}p$ for $k = 1, 2, \dots$.

The discussion for the Examples would have been slightly neater if I had had a name for the toss on which the first head occurs. Define

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

Then I could write

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

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, a random variable Z is said to have a ***geometric(p) distribution*** if it can take values 1, 2, 3, ... with probabilities

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

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

> **Remark.** Be warned. Some authors 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,

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

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

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

What does that tell you about coin tossing?

The final example for this Chapter, whose statement is also borrowed verbatim from the Mosteller book, is built around a "geometric" mechanism.

Example <1.6> 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?

1.2 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?

1.3 The examples

<1.1> **Example.** *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

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

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

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

□

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

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

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

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

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

where

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

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

$$\mathbb{P}\{\text{four of a kind}\} = \sum\nolimits_1^{13} \mathbb{P}F_i = 13\mathbb{P}F_1 \qquad \text{by rule P4.}$$

Now break F_1 into simpler pieces, $F_1 = \cup_{j=1}^5 F_{1j}$, where

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

Again by disjointness and symmetry, $\mathbb{P}F_1 = 5\mathbb{P}F_{1,1}$.

Decompose the event $F_{1,1}$ into five "stages", $F_{1,1} = N_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5$, where

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

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

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

Thus

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

Can you see any hidden assumptions in this analysis?

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? □

<1.3> **Example.** (The Prisoner's Dilemma—verbatim from Mosteller, 1987)

> 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

$\boxed{\text{aB}}$ = both A and B to be released, warder must say B

$\boxed{\text{aC}}$ = both A and C to be released, warder must say C

$\boxed{\text{Bc}}$ = both B and C to be released, warder says B

$\boxed{\text{bC}}$ = both B and C to be released, warder says C.

There are three events to be considered

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

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,

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

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

With those probabilities we calculate

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

from which we deduce (via rule P5) that

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

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

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

he would calculate a conditional probability

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

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

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

M waits for hhh and R waits for tthh.

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

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

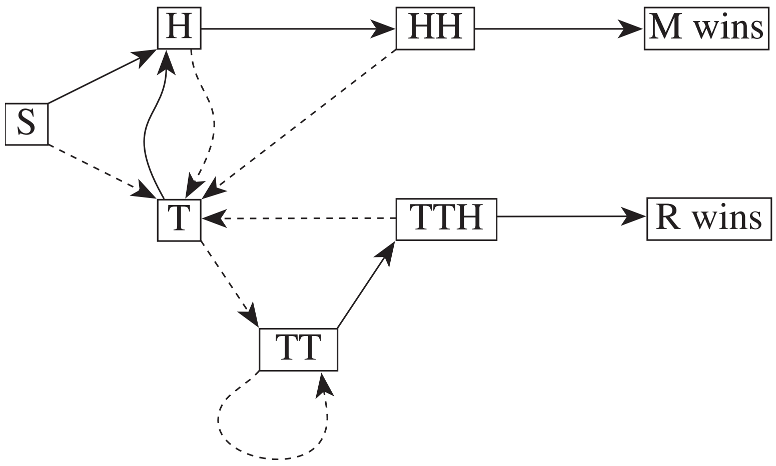
You might imagine that M has the advantage. After all, surely it must be easier to get a pattern of length 3 than a pattern of length 4. You'll discover that the solution is not that straightforward.

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

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

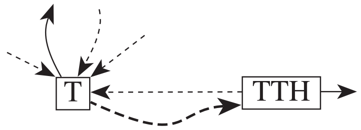
By claiming that these states summarize the game I am tacitly assuming that the coin has no "memory", in the sense that the conditional probability of a head given any particular past sequence of heads and tails is 1/2 (for a fair coin). The past history leading to a particular state does not matter;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

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



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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

If we similarly define a column vector,

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

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

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

with the boundary conditions $P_{\text{M wins}} = 1$ and $P_{\text{R wins}} = 0$.

> **Remark.** Write e'_M and e'_R for the last two rows of P and Q for the 6×8 matrix made up of the first 6 rows of $I - P$. Then π is the unique solution to the equation
> $$\begin{bmatrix} Q \\ e'_M \\ e'_R \end{bmatrix} \pi = e_M$$

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

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

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

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

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

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

a two-step transition possibility.

> **Remark.** What do the elements of the matrix P^n represent? What happens to this matrix as n tends to infinity? If you are interested in computation, look at the file HHH.TTHH.R, or try similar calculations with Matlab or Mathematica.

The name ***Markov chain*** is given to any process representable as the movement of a mouse (or 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.

<1.5> **Example.** *Suppose a coin has probability p of landing heads on any particular toss, independent of outcomes of other tosses. In a sequence of such tosses, what is the probability that the first head appears on the kth toss (for $k = 1, 2, \dots$)?*

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

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

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

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

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

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

□

<1.6> **Example.** (The Three-Cornered Duel—also borrowed from Mosteller, 1987) 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 oppenent, C, leaving A one shot before B finishes him off too. That single shot from A at B would have to succeed:

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

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,

$$\{\text{C misses, A hits}\}$$

or

$$\{\text{C misses, A misses, C misses, A hits}\}$$

or

$$\{\text{C misses, (A misses, C misses) twice, A hits}\}$$

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

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

In summary:

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

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

References

Mosteller, F. (1987). *Fifty Challenging Problems in Probability with Solutions*. New York: Dover.

Chapter 2

Expectations

2.1 Overview

Recall from Chapter 1 that a random variable is just a function that attaches a number to each item in the sample space. Less formally, a random variable corresponds to a numerical quantity whose value is determined by some chance mechanism.

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 (info), the symbol

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

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. For example, if the info forces a random variable X to always take values larger than 16 then $\mathbb{E}\,(X \mid \text{info})$ will be larger than 16.

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

$$\mathbb{E}\,(X \mid \text{info and "}F\text{ has occurred"}\,)$$
$$\mathbb{E}\,(X \mid \text{info, }F\,)$$

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. I will provide an interpretation for just one of the rules, using long-run averages of values generated by independent repetitions of random experiments. You should provide analogous interpretations for the other rules.

> **Remark.** See the Appendix to this Chapter for another interpretation, which does not depend on a preliminary concept of independent repetitions of an experiment. The expected value $\mathbb{E}X$ can be interpreted as a"fair price" to pay up-front, in exchange for a random return X later—something like an insurance premium.

Rules for (conditional) expectations

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

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

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

(E3) if the given "info" forces $X \le Y$ then $\mathbb{E}\left(X \mid \text{info}\right) \le \mathbb{E}\left(Y \mid \text{info}\right)$;

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

$$\mathbb{E}\left(X \mid \text{info}\right) = \sum\nolimits_i \mathbb{E}\left(X \mid F_i, \text{info}\right) \mathbb{P}\left(F_i \mid \text{info}\right).$$

Rule (E4) 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 equal to the whole sample space: Repeat the experiment (independently) a very large number (n) of times, each time with the same conditioning info, 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	

By the frequency interpretation of probabilities, $\mathbb{P}(F_1 \mid \text{info}) \approx n_1/n$ and $\mathbb{P}(F_2 \mid \text{info}) \approx n_2/n$. 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,

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

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.

Modulo some fine print regarding convergence of infinite series, rule (E1) extends to sums of infinite sequences of random variables,

$$(E1)' \qquad \mathbb{E}(X_1 + X_2 + \dots) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \dots$$

(For mathematical purists: the asserted equality holds if $\sum_i \mathbb{E}|X_i| < \infty$.)

> **Remark.** The rules for conditional expectations actually include all the rules for conditional probabilities as special cases. This delightfully convenient fact can be established by systematic use of particularly simple random variables. For each event A the ***indicator function of A*** is defined by
>
> $$\mathbb{I}_A = \begin{cases} 1 & \text{if the event } A \text{ occurs,} \\ 0 & \text{if the event } A^c \text{ occurs.} \end{cases}$$
>
> Each $\mathbb{I}_A$ is a random variable.
>
> Rule (E4) with $F_1 = A$ and $F_2 = A^c$ gives
>
> $$\begin{aligned} \mathbb{E}(\mathbb{I}_A \mid \text{info}) &= \mathbb{E}(\mathbb{I}_A \mid A, \text{ info}) \mathbb{P}(A \mid \text{info}) + \\ &\quad + \mathbb{E}(\mathbb{I}_A \mid A^c, \text{ info}) \mathbb{P}(A^c \mid \text{info}) \\ &= 1 \times \mathbb{P}(A \mid \text{info}) + 0 \times \mathbb{P}(A^c \mid \text{info}) \qquad \text{by (E2).} \end{aligned}$$

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 version of (E1) for infinite sums 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 recall 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.

Rules (E2) and (E4) can be used to calculate expectations from probabilities, for random variables that take values in "discrete" set. 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 $c_1, c_2, \dots$. 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,

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

More succinctly,

$$(E5) \qquad \mathbb{E}(g(X) \mid \text{info}) = \sum\nolimits_i g(c_i)\mathbb{P}(X = c_i \mid \text{info}).$$

In particular,

$$(E5)' \qquad \mathbb{E}(X \mid \text{info}) = \sum\nolimits_i c_i\mathbb{P}(X = c_i \mid \text{info}).$$

Both (E5) and (E5)' apply to random variables X that take values in the "discrete set" $\{c_1, c_2, \dots\}$.

Remark. For random variables that take a continuous range of values an approximation argument (see Chapter 7) will provide us with an analog of (E5) with the sum replaced by an integral.

You will find it helpful to remember expectations for a few standard mechanisms, such as coin tossing, rather than have to rederive them repeatedly.

Example <2.1> Expected value for the geometric(p) distribution is $1/p$.

The calculation of an expectation is often a good way to get a rough feel for the behaviour of a random process, but it doesn't tell the whole story.

Example <2.2> Expected number of tosses to get tthh is 16 with fair coin.

Compare with the next Example.

Example <2.3> Expected number of tosses to get hhh is 14 with fair coin.

Don't the last two results seem strange? On average it takes longer to reach tthh than hhh, but also on average the pattern tthh appears first.

> **Remark.** You should also be able to show that the expected number of tosses for the completion of the game with competition between hhh and tthh is $9\,^1\!/_3$. 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?

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 simplification.

Example <2.8> [Coupon collector problem] In order to encourage consumers to buy many packets of cereal, a manufacurer 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 will 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 $\mathbb{P}A_1$, $\mathbb{P}A_2$, ... via the formula

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

Example <2.9> Suppose an urn contains r red balls and b black balls, all identical except for color. Suppose you remove one ball at a time, without replacement, at each step selecting at random from the urn: if k balls remain then each has probability $1/k$ of being chosen.Show that the expected number of red balls removed before the firstblack ball equals $r/(b+1)$.

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 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?

The conditioning property (E5) can be used in a subtle way to solve the classical gambler's ruin problem. The method of solution invented by Abraham de Moivre, over two hundred years ago, has grown into one of the main technical tools of modern probability.

Example <2.10> 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 α dollar bills, and Betamax starts with β dollars bills (both α and β 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. The next Example provides the details. You could safely skip it if you understand the tricky idea behind Example <2.10>.

Example <2.11> Same problem as in Example <2.10>, except that the coin they toss has probability $p \neq 1/2$ of landing heads. (Could be skipped.)

You could also 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 <2.12> Big pills, little pills. (Tricky. Should be skipped.)

2.2 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,

  $$\mathbb{E}\left(X \mid \text{info}\right) = \sum\nolimits_i \mathbb{E}\left(X \mid F_i, \text{info}\right) \mathbb{P}\left(F_i \mid \text{info}\right),$$

  where the disjoint events $F_1, F_2, \ldots$ 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, $\mathbb{E}\left(\mathbb{I}_A \mid \text{info}\right)$, is the same as the probability of the corresponding event, $\mathbb{P}\left(A \mid \text{info}\right)$.

- As a consequence of the rules,

  $$\mathbb{E}\left(g(X) \mid \text{info}\right) = \sum\nolimits_i g(c_i) \mathbb{P}\left(X = c_i \mid \text{info}\right),$$

  if X can take only values $c_1, c_2, \ldots$.

2.3 The examples

<2.1> **Example.** For independent coin tossing, what is the expected value of X, the 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: a Markov argument without the picture
Condition on whether the first toss lands heads (H_1) or tails (T_1).

$$\begin{aligned}\mathbb{E}X &= \mathbb{E}(X \mid H_1)\mathbb{P}H_1 + \mathbb{E}(X \mid T_1)\mathbb{P}T_1 \\ &= (1)p + (1 + \mathbb{E}X)(1-p).\end{aligned}$$

The reasoning behind the equality

$$\mathbb{E}(X \mid T_1) = 1 + \mathbb{E}X$$

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 decrease as p increases? What happens as p tends to zero or one?)

Method B
By the formula (E5),

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

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

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

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. □

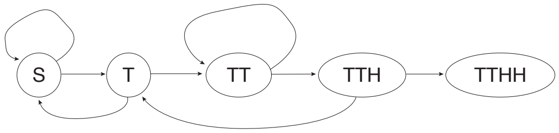
<2.2> **Example.** The "HHH versus TTHH" Example in Chapter 1 solved the following problem:

> 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. The corresponding transition matrix is:

$$P = \begin{array}{c} \\ \boxed{\text{S}} \\ \boxed{\text{T}} \\ \boxed{\text{TT}} \\ \boxed{\text{TTH}} \\ \boxed{\text{TTHH}} \end{array} \begin{array}{c} \begin{array}{ccccc} \boxed{\text{S}} & \boxed{\text{T}} & \boxed{\text{TT}} & \boxed{\text{TTH}} & \boxed{\text{TTHH}} \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

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

$$\begin{aligned} \mathcal{E}_S &= \mathbb{E}(X \mid \text{start at S}) \\ \mathcal{E}_H &= \mathbb{E}(X \mid \text{start at H}) \\ &\vdots \end{aligned}$$

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

To solve gthe problem, 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,

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

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

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

Similarly,

$$\begin{aligned} \mathcal{E}_T &= \tfrac{1}{2}(1 + \mathcal{E}_{TT}) + \tfrac{1}{2}(1 + \mathcal{E}_S) \\ \mathcal{E}_{TT} &= \tfrac{1}{2}(1 + \mathcal{E}_{TT}) + \tfrac{1}{2}(1 + \mathcal{E}_{TTH}) \\ \mathcal{E}_{TTH} &= \tfrac{1}{2}(1 + 0) + \tfrac{1}{2}(1 + \mathcal{E}_T) \end{aligned}$$

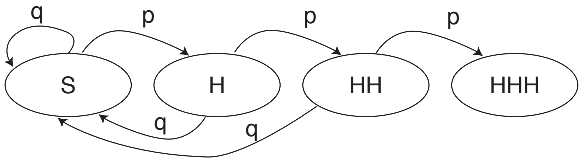
What does the zero in the last equation represent?

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. □

<2.3> **Example.** *The expected number of tosses to get hhh with fair coin is 14. For a coin that lands heads with probability p the expected number to get hhh is* $p^{-1} + p^{-2} + p^{-3}$.

I could use the same method as for the tthh problem but I want to show you a variation on the method that is easier to generalize. It involves a lot more notation, but it captures better the recursive nature of the problem.

> **Remark.** In the original version of the notes I assumed $p = 1/2$ but in class I considered the more general case for which the coin lands heads with probability p and tails with probability $q = 1 - p$. The general p makes it easier to recognize a pattern. In class I also presented a slight variation on the original method.



First relabel the states: $S_0 = S$, $S_1 = H$, $S_2 = HH$, and $S_3 = HHH$. Then write X_k for the number of steps to reach state S_k and define

$$\tau_k = \mathbb{E}(X_k \mid \text{start at } S_0).$$

For each integer i write $\mathcal{H}_i$ for the event that the ith toss results in a head and $\mathcal{T}_i$ for $\mathcal{H}_i^c$. I claim that

<2.4> $$\mathbb{E}(X_{k+1} \mid X_k = m, \mathcal{H}_{m+1}, \text{start at } S_0) = m + 1$$

<2.5> $$\mathbb{E}(Y_{k+1} \mid X_k = m, \mathcal{T}_{m+1}, \text{start at } S_0) = m + 1 + \tau_{k+1}.$$

The second equality reflects the fact that the tail sends us right back to the start: the m comes from the X_k contribution, the 1 from the $(m+1)$st toss, and the τ_{k+1} from the fact that, given the conditioning information, the rest of the task to reach S_{k+1} is just like starting from state S_0.

We need to take a weighted average of these conditional expectations to get τ_{k+1}. The weights are

$$\begin{aligned}
&\mathbb{P}\{X_k = m, \mathcal{H}_{m+1} \mid \text{start at } S_0\} \\
&\quad = \mathbb{P}\{X_k = m \mid \text{start at } S_0\}\mathbb{P}\{\mathcal{H}_{m+1} \mid \text{start at } S_0, X_k = m\} \\
&\quad = \mathbb{P}\{X_k = m \mid \text{start at } S_0\}p
\end{aligned}$$

and

$$\mathbb{P}\{X_k = m, \mathcal{T}_{m+1} \mid \text{start at } S_0\} = \mathbb{P}\{X_k = m \mid \text{start at } S_0\}q.$$

By rule (E4),

$$\begin{aligned}\tau_{k+1} = \mathbb{E}(X_{k+1} \mid \text{start at } S_0)& \\ &= \sum\nolimits_m \mathbb{E}(X_{k+1} \mid X_k = m, \mathcal{H}_{m+1}, \text{start at } S_0)\times \\ &\qquad \mathbb{P}(X_k = m, \mathcal{H}_{m+1} \mid \text{start at } S_0) \\ &\quad + \sum\nolimits_m \mathbb{E}(X_{k+1} \mid X_k = m, \mathcal{T}_{m+1}, \text{start at } S_0)\times \\ &\qquad \mathbb{P}(X_k = m, \mathcal{T}_{m+1} \mid \text{start at } S_0) \\ &= \sum\nolimits_m \mathbb{P}(X_k = m \mid \text{start at } S_0)p(m+1) \\ &\qquad + \sum\nolimits_m \mathbb{P}(X_k = m \mid \text{start at } S_0)q(m+1+\tau_{k+1}) \\ &= \sum\nolimits_m (m+1)\mathbb{P}(X_k = m \mid \text{start at } S_0) \\ &\qquad q\tau_{k+1}\sum\nolimits_m \mathbb{P}(X_k = m \mid \text{start at } S_0).\end{aligned} \quad <2.6>$$

At this point you need to notice that

$$\begin{aligned}\tau_k + 1 &= \mathbb{E}(X_k + 1 \mid \text{start at } S_0) \\ &= \sum\nolimits_m \mathbb{E}(X_k + 1 \mid \text{start at } S_0, X_k = m)\mathbb{P}(X_k = m \mid \text{start at } S_0) \\ &= \sum\nolimits_m (m+1)\mathbb{P}(X_k = m \mid \text{start at } S_0)\end{aligned}$$

and

$$1 = \sum\nolimits_m \mathbb{P}(X_k = m \mid \text{start at } S_0).$$

Substitution of these expressions into <2.6> leaves

$$\tau_{k+1} = \tau_k + 1 + q\tau_{k+1} \qquad \text{or} \qquad \tau_{k+1} = \frac{\tau_k + 1}{p}$$

If you are very brave you might use $\tau_0 = 0$, otherwise you would appeal to Example <2.1> to get $\tau_1 = 1/p$. The recursive equality then gives $\tau_2 = 1/p + 1/p^2$ and $\tau_3 = 1/p + 1/p^2 + 1/p^3$, as asserted.

In the original version of this Example I was a bit trickier, effectively writing the argument for <2.6> as

$$\begin{aligned}\mathbb{E}(X_{k+1} - X_k \mid \text{start at } S_0)\\ &= \sum\nolimits_m \mathbb{E}(X_{k+1} - X_k \mid X_k = m, \mathcal{H}_{m+1}, \text{start at } S_0)\times\\ &\qquad \mathbb{P}(X_k = m, \mathcal{H}_{m+1} \mid \text{start at } S_0)\\ &\quad + \sum\nolimits_m \mathbb{E}(X_{k+1} - X_k \mid X_k = m, \mathcal{T}_{m+1}, \text{start at } S_0)\times\\ &\qquad \mathbb{P}(X_k = m, \mathcal{T}_{m+1} \mid \text{start at } S_0)\\ &= \sum\nolimits_m \mathbb{P}(X_k = m \mid \text{start at } S_0)p(1)\\ &\qquad + \sum\nolimits_m \mathbb{P}(X_k = m \mid \text{start at } S_0)q(1 + \tau_{k+1})\\ &= (1 + q\tau_{k+1}) \sum\nolimits_m \mathbb{P}(X_k = m \mid \text{start at } S_0)\end{aligned}$$

so that

$$\tau_{k+1} - \tau_k = \mathbb{E}(X_{k+1} - X_k \mid \text{start at } S_0) = 1 + q\tau_{k+1}.$$

That is, I realized that the m's on the right-hand sides of <2.4> and <2.5> were going to end up contributing $\mathbb{E}(X_k \mid \text{start at } S_0)$. □

<2.7> **Example.** (Brief summary) Consider a machine that sells tickets for random prices $M_1, M_2, \dots$ where

$$\mathbb{E}(M_i \mid \text{ all available past info } \} = \lambda$$

(More succinctly, the prices have identical distributions independent of all past information.) The price M_i entitles Sam to a shot at the prize by tossing a coin that heads with probability p. (A head wins.) Sam keeps buying tickets until he wins. What is the expected value of T, the total cost to Sam?

Define W_i as the event that the ith toss gives a head. Argue that

$$T = M_1 + M_2 \mathbf{1}_{W_1^c} + M_3 \mathbf{1}_{W_1^c W_2^c} + \dots.$$

By the linearity rule for expectations,

$$\mathbb{E}T = \mathbb{E}M_1 + \mathbb{E}(M_2 \mathbf{1}_{W_1^c}) + \mathbb{E}(M_3 \mathbf{1}_{W_1^c W_2^c}) + \dots.$$

To calculate $\mathbb{E}(M_3 \mathbf{1}_{W_1^c W_2^c})$ split according to the four possible combinations W_1W_2, $W_1^cW_2$, $W_1W_2^c$, and $W_1^cW_2^c$, only one of which makes a nonzero contribution to get

$$\mathbb{E}(M_3 \mathbf{1}_{W_1^c W_2^c}) = \lambda q^2.$$

And so on. Eventually,

$$\mathbb{E}T = \lambda + \lambda q + \lambda q^2 + \cdots = \lambda/p.$$

□

I asked in class what the last example had to do with the hhh Example.

<2.8> **Example.** *In order to encourage consumers to buy many packets of cereal, a manufacurer 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, $T_1 + T_2$, ..., $T_1 + T_2 + \cdots + T_{10}$, 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 $\mathbb{E}(T_1 + T_2 + \cdots + T_{10})$, which rule E1 (applied repeatedly) reexpresses as $\mathbb{E}T_1 + \mathbb{E}T_2 + \cdots + \mathbb{E}T_{10}$.

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

$$T_2 \text{ has a geometric}(9/10) \text{ distribution.}$$

Deduce from Example <2.1> that $\mathbb{E}T_2 = 10/9$, a value 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 acually irrelevent; 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:

$$T_3 \text{ has geometric } (8/10) \text{ distribution,}$$

and thus $\mathbb{E}T_3 = 10/8$.

Remark. More precisely, T_3 is independent of T_2 with conditional probability distribution geometric (8/10). That is, with $p = 8/10$,

$$\mathbb{P}\{T_3 = k \mid T_2 = \ell\} = (1-p)^{k-1}p \qquad \text{for } k = 1, 2, \dots$$

for every possible ℓ.

And so on, leading to

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

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 me 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.

<2.9> **Example.** *Suppose an urn contains r red balls and b black balls, all identical except for color. Suppose you remove one ball at a time, without replacement, at each step selecting at random from the urn: if k balls remain then each has probability $1/k$ of being chosen.Show that the expected number of red balls removed before the firstblack ball equals $r/(b+1)$.*

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

$$\mathbb{E}T = \sum\nolimits_k k\mathbb{P}\{T = k\}.$$

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,

$$\begin{aligned}\mathbb{P}\{T = k\} &= \mathbb{P}\{\text{first k balls red, (k+1)st ball is black}\}\\ &= \mathbb{P}(R_1R_2\dots R_kB_{k+1})\\ &= (\mathbb{P}R_1)\mathbb{P}(R_2 \mid R_1)\mathbb{P}(R_3 \mid R_1R_2)\dots\mathbb{P}(B_{k+1} \mid R_1\dots R_k)\\ &= \frac{r}{r+b}.\frac{r-1}{r+b-1}\dots\frac{b}{r+b-k}.\end{aligned}$$

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 $1, \ldots, r$. Let T_i equal 1 if red ball number i is sampled before the first black ball, zero otherwise. That is, T_i is the indicator for the event

$$\{\text{red ball number } i \text{ is removed before any of the black balls}\}.$$

(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 $T = T_1 + \cdots + T_r$. 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

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

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 $r/(b+1)$. □

<2.10> **Example.** *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 α dollar bills, and Betamax starts with β dollars bills (both α and β 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 $a+b$ or 0, depending on whether A won or not.) It is a random variable taking values in the range $\{0, 1, 2, \dots, a+b\}$. 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 $\theta = \mathbb{P}\{\text{A wins}\}$.

First calculation

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

$$\begin{aligned} A_n &= \{\text{A wins at, or before, the } n\text{th toss}\},\\ B_n &= \{\text{B wins at, or before, the } n\text{th toss}\},\\ C_n &= \{\text{game still going after the } n\text{th toss}\}. \end{aligned}$$

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

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

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 $k = 0, 1, \dots, a+b$, then work towards another appeal to rule (E4). For $k = 0$ or $k = a+b$, the game will be over, and X_n must take the same value as X_{n-1}. That is,

$$\mathbb{E}(X_n \mid F_0) = 0 \qquad \text{AND} \qquad \mathbb{E}(X_n \mid F_{\alpha+\beta}) = \alpha+\beta.$$

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 $k = 1, 2, \ldots, \alpha + \underline{-}1$,

$$\mathbb{E}(X_n \mid F_k) = {}^1\!/\!_2(k+1) + {}^1\!/\!_2(k-1) = k.$$

Now invoke (E4).

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

Compare with the direct application of (E5)' to the calculation of EX_{n-1}:

$$\begin{aligned}\mathbb{E}(X_{n-1}) = \big(0 \times \mathbb{P}\{X_{n-1} = 0\}\big) + \big(1 \times \mathbb{P}\{X_{n-1} = 1\}\big) + \\ \cdots + \big((\alpha+\beta) \times \mathbb{P}\{X_{n-1} = \alpha+\beta\}\big)\,,\end{aligned}$$

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

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

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

Putting the two answers together

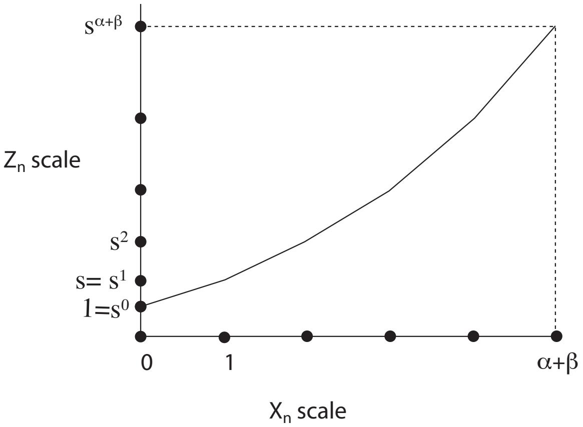
We have two results: $\mathbb{E}X_n = \alpha$, no matter how large n is; and $\mathbb{E}X_n$ gets arbitrarily close to $\theta(\alpha+\beta)$ as n gets larger. We must have $\alpha = \theta(\alpha+\beta)$. That is, Alf has probability $\alpha/(\alpha+\beta)$ of eventually winning all the money. □

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.

Danger: The next two Examples are harder. They can be skipped.

<2.11> **Example.** Same problem as in Example <2.10>, 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 $0 < p < 1$ (and $p \neq 1/2$). Essentially De Moivre's idea was that we could use almost the same method as in Example <2.10> 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 θ for $\mathbb{P}\{\text{A wins}\}$, and give the events A_n, B_n, and C_n the same meaning as in Example <2.10>.

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

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

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

$$\mathbb{E}(Z_n \mid X_{n-1} = 0) = s^0 \qquad \text{AND} \qquad \mathbb{E}(Z_n \mid X_{n-1} = \alpha+\beta) = s^{\alpha+\beta}.$$

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

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

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

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

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

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

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

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

What goes wrong with this calculation if $p = 1/2$? As a check we could let p tend to $1/2$, getting

$$\begin{aligned}\frac{1-s^\alpha}{1-s^{\alpha+\beta}} &= \frac{(1-s)(1+s+\dots+s^{\alpha-1})}{(1-s)(1+s+\dots+s^{\alpha+\beta-1})} \qquad \text{for } s \neq 1\\ &= \frac{1+s+\dots+s^{\alpha-1}}{1+s+\dots+s^{\alpha+\beta-1}}\\ &\to \frac{\alpha}{\alpha+\beta} \qquad \text{as } s \to 1.\end{aligned}$$

Comforted? □

<2.12> **Example.** My interest in the calculations in Example <2.9> 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. The following brown paragraphs are taken hyper-verbatim from the Monthly; I was seeing how closely LaTeX could reproduce the original text.

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 kth 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):

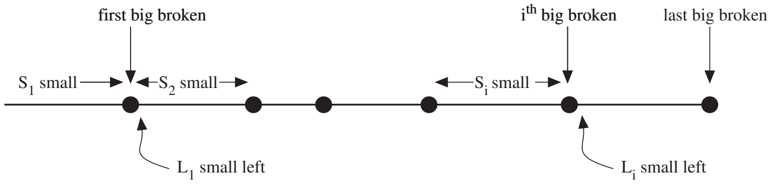
> *Editorial comment.* Most solvers derived a recurrence relation, guessed the answer, and verified it by induction. Several commented on the origins of the problem. Robert High saw a version of it in the MIT Technology Review of April, 1990. Helmut Prodinger reports that

> 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 ith big pill, and the $+1$ representing the half of the big pill that is returned to the bottle. Taking expectations we get

<2.13> $$\mathbb{E}L_i = \mathbb{E}L_{i-1} - \mathbb{E}S_i + 1.$$

The result from Example <2.9> 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 <2.9>, with

$r = L_{i-1}$ red balls (= small pills)
$b = m-(i-1)$ black balls (= big pills).

From that Example,

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

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 <2.13> 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

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

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

<2.14> $$n\mathbb{E}H_1 + m\mathbb{E}B_1 = n\mathbb{P}\{H_1 = 1\} + m\mathbb{P}\{B_1 = 1\}.$$

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 kth 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

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

Also by symmetry,

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

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 <2.14>, the asserted answer to part (a) results. □

2.4 Appendix: The 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 × 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 ≥ 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 α and β.*

Consider the combined effect of the following fair bets:

you pay me $\alpha p(X)$ to receive αX
you pay me $\beta p(Y)$ to receive βY
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 $p(X \mid F)$. 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 $(X - p(X \mid F))\, Z$. 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 $F_1, \ldots, F_k$, then

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

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

(a) For each i, pay $c_i p(F_i)$ in order to receive c_i if F_i occurs.

ritem[(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, and $\mathbb{E}(X \mid F)$ for $p(X \mid F)$, 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 repetions from which to calculate a long-run frequency or average.

See de Finetti (1974) for a detailed discussion of expectations as fair prices.

References

de Finetti, B. (1974). *Theory of Probability*, Volume 1. New York: Wiley.

Chapter 3

Things binomial

3.1 Overview

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".) Clearly, $\binom{n}{0} = 1 = \binom{n}{n}$: there is only one empty subset and only one subset containing everything.

Here is a neat probabilistic way to determine $\binom{n}{k}$, for integers $1 \le k \le n$. 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, the event

$$A = \{\text{the sample consists of the red balls}\}$$

has probabilty $1/\binom{n}{k}$. We can also calculate this probability using a conditioning argument. Given that the first i balls are red, the probability that the $(i+1)$st is red is $(k-i)/(n-i)$. Thus

$$\mathbb{P}A = \frac{k}{n}.\frac{k-1}{n-1}.\frac{k-2}{n-2}\dots\frac{1}{n-k+1} = \frac{k!(n-k)!}{n!}.$$

Equating the two values for $\mathbb{P}A$ we get

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

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 α define

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

Then

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

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

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

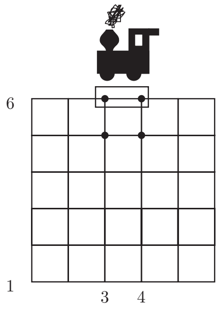
Compare with the binomial expansion,

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

Example <3.1> For n independent tosses of a coin that lands heads with probability p, show that the total number of heads has a Bin(n,p) distribution, with expected value np.

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.

Example <3.2> 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 conditioning method you already know.

Example <3.3> Suppose a multiple-choice exam consists of a string of unrelated questions, each having three possible answers. Suppose also that 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. Finally, 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?

As a method of solving statistical problems, Bayesian inference is advocated 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 previous Example—that is controversial in a general setting.

The Bayesian message comes through more strongly in the next Example.

Example <3.4> Suppose we have three coins, which land heads with probabilities p_1, p_2, and p_3. Choose a coin according to the ***prior distribution*** $\theta_i = \mathbb{P}\{\text{ choose coin } i \}$, for $i = 1, 2, 3$, then toss that coin n times. Find the posterior probabilities $\mathbb{P}\{\text{ chose coin } i \mid k \text{ heads with } n \text{ tosses} \}$, for $k = 0, 1, \dots, n$.

We will meet the Binomial again.

3.2 The examples

<3.1> **Example.** *For n independent tosses of a coin that lands heads with probability p, show that the total number of heads has a Bin(n, p) distribution, with expected value np.*

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

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

Here i runs from 1 to $\binom{n}{k}$, because each F_i corresponds to a different choice of the k positions for the heads to occur.

Remark. The indexing on the F_i is most uninformative; it gives no indication of the corresponding pattern of heads and tails. Maybe you can think of something better.

Write H_j for the event {jth toss is a head}. Then

$$
\begin{aligned}
\mathbb{P}F_1 &= \mathbb{P}\left(H_1H_2\dots H_kH_{k+1}^c\dots H_n^c\right)\\
&= (\mathbb{P}H_1)(\mathbb{P}H_2)\dots(\mathbb{P}H_n^c) \qquad \text{by independence}\\
&= p^k(1-p)^{n-k}.
\end{aligned}
$$

A similar calculation gives $\mathbb{P}F_i = p^k(1-p)^{n-k}$ for every other i; all that changes is the order in which the p and $(1-p)$ factors appear. Thus

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

which is the asserted Binomial distribution.

It is possible to calculate $\mathbb{E}X$ by the summation formula

$$
\begin{aligned}
\mathbb{E}X &= \sum\nolimits_{k=0}^{n} \mathbb{E}(X|X = k)\mathbb{P}\{X = k\}\\
&= \sum\nolimits_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k}\\
&= \sum\nolimits_{k=1}^{n} \frac{n(n-1)!}{(k-1)!(n-k)!}p^k(1-p)^{n-k}\\
&= np\sum\nolimits_{k-1=0}^{n-1}\binom{n-1}{k-1}p^{k-1}(1-p)^{(n-1)-(k-1)}\\
&= np \qquad \text{cf. binomial expansion of } (p + (1-p))^{n-1}.
\end{aligned}
$$

The manipulations of the sums was only slightly tedious, but why endure even a little tedium when the method of indicators is so much simpler? Define

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

Then $X = X_1 + \dots X_n$, which gives $\mathbb{E}X = \mathbb{E}X_1 + \dots \mathbb{E}X_n = n\mathbb{E}X_1$. Calculate.

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

Thus $\mathbb{E}X = np$.

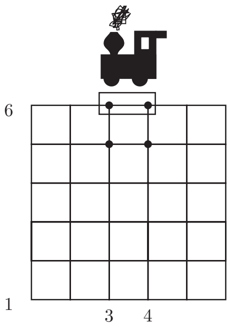
> **Remark.** The calculation of the expected value made no use of the independence. If each X_i has ***marginal*** distribution Ber(p), that is, if
>
> $$\mathbb{P}\{X_i = 1\} = p = 1 - \mathbb{P}\{X_i = 0\} \qquad \text{for each } i,$$
>
> then $\mathbb{E}(X_1 + \dots X_n) = np$, regardless of possible dependence between the tosses. The expectation of a sum is the sum of the expectations, no matter how dependent the summands might be.

The symbol Ber stands for "Bernoulli".

□

<3.2> **Example.** *An unwary visitor to the Big City is standing at the corner of 1st Street and 1st Avenue. He wishes to reach the railroad station, which actually occupies the block on 6th Street from 3rd to 4th Avenue. (The Street numbers increase as one moves north; the Avenue numbers increase as one moves east.) He is unaware that he is certain to be mugged as soon as he steps onto 6th Street or 6th Avenue.*

Being unsure of the exact location of the railroad station, the visitor lets himself be guided by the tosses of a fair coin: at each intersection he goes east, with probability $1/2$*, 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.

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

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

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

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

□

<3.3> **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) conditional on being a guesser, X has a Bin$(6, 1/3)$ distribution (sometimes abbreviated to $X \mid G \sim \text{Bin}(6, 1/3)$)

(ii) conditional on being a skilled candidate, X has a Bin$(6, 1/2)$ distribution (sometimes abbreviated to $X \mid S \sim \text{Bin}(6, 1/2)$).

(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.

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

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

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

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:

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

Apply the usual splitting/conditioning argument.

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

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

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

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

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

of answering the next question correctly.

Remark. 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”.

□

<3.4> **Example.** *Suppose we have three coins, which land heads with probabilities p_1, p_2, and p_3. Choose a coin according to the* ***prior distribution*** $\theta_i = \mathbb{P}\{\text{choose coin } i\}$*, for* $i = 1, 2, 3$*, then toss that coin* n *times. Find the posterior probabilities*

$$\mathbb{P}\{\text{chose coin } i \mid \; k \text{ heads with } n \text{ tosses }\} \qquad \text{for } k = 0, 1, \dots, n.$$

Let C_i denote the event $\{\text{chose coin } i\}$ and D_k denote the event that we get k heads from the n tosses. Then $\mathbb{P}C_i = \theta_i$ and

$$\mathbb{P}(D_k \mid C_i) = \binom{n}{k} p_i^k (1-p_i)^{n-k} \qquad \text{for } k = 0, 1, \dots, n.$$

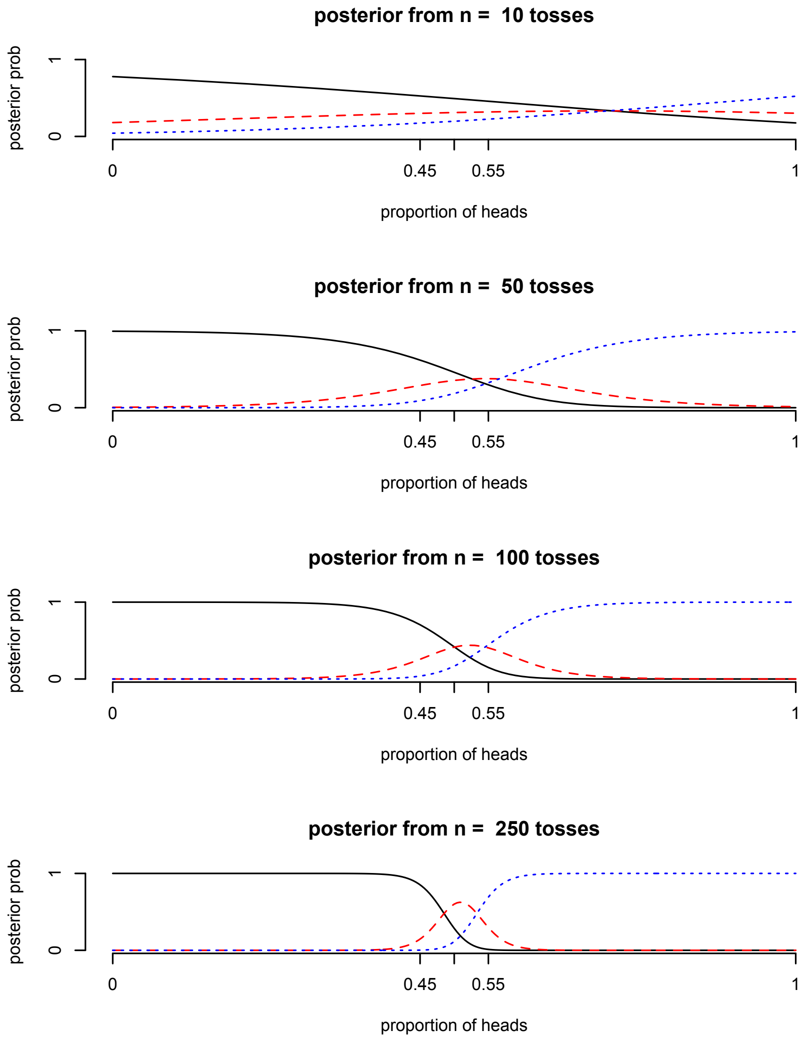
Condition.

$$\begin{aligned} \mathbb{P}(C_i \mid D_k) &= \frac{\mathbb{P}(C_i D_k)}{\mathbb{P}D_k} \\ &= \frac{\mathbb{P}(D_k \mid C_i)\mathbb{P}(C_i)}{\sum_{j=1}^3 \mathbb{P}(D_k \mid C_j)\mathbb{P}(C_j)} \\ &= \frac{p_i^k (1-p_i)^{n-k}\theta_i}{\sum_{j=1}^3 p_j^k (1-p_j)^{n-k}\theta_j} \end{aligned}$$

Notice that the $\binom{n}{k}$ factors have cancelled. In fact, we would get the same posterior probabilities if we conditioned on any particular pattern of k heads and $n-k$ tails.

The R-script Bayes.R defines functions to plot the posterior probabilities as a function of k/n, for various choices of the p_i's and the θ_i's and n. The $\mathbb{P}(C_1 \mid D_k)$ are in solid black, the $\mathbb{P}(C_2 \mid D_k)$ are in dashed red, and the $\mathbb{P}(C_3 \mid D_k)$ are in dotted blue. For the pictures I chose $p_1 = 0.45$, $p_2 = 0.5$ and $p_3 = 0.55$ with prior probabilities $\theta_1 = 0.5$, $\theta_2 = 0.3$, and $\theta_3 = 0.2$. The pictures were produced by running:

```
draw.posterior(
        p = c(0.45,0.5,0.55),
        tosses=c(10,50,100,250),
        prior = c(0.5,0.3,0.2)
        )
```



When n gets large, the posterior probability $\mathbb{P}(C_i \mid D_k)$ gets closer to 1 for values of k/n close to p_i, a comforting fact. □

Chapter 4

Variances and covariances

4.1 Overview

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 μ then the expected value equals μ. To refine the picture of a distribution about its "center of location" we need some measure of spread (or concentration) around that value. For many distributions the simplest measure to calculate is the variance (or, more precisely, the square root of the variance).

Definition. The ***variance*** of a random variable X with expected value $\mathbb{E}X = \mu$ is defined as $\text{var}(X) = \mathbb{E}\left((X-\mu)^2\right)$. The square root of the variance of a random variable is called its ***standard deviation***, sometimes denoted by $\text{sd}(X)$.

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: for constants a and b,

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

> **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 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.

There is an enormous 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 <4.1> The Tchebychev inequality: $\mathbb{P}\{|X-\mu| \geq \epsilon\} \leq \text{var}(X)/\epsilon^2$, where $\mu = \mathbb{E}X$ and $\epsilon > 0$.

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

The Tchebychev inequality gives the right insight when dealing with sums of random variables, for which variances are easy to calculate. Suppose $\mathbb{E}Y = \mu_Y$ and $\mathbb{E}Z = \mu_Z$. Then

$$\begin{aligned}\text{var}(Y + Z) &= \mathbb{E}\left[Y - \mu_Y + Z - \mu_Z\right]^2 \\ &= \mathbb{E}\left[(Y - \mu_Y)^2 + 2(Y - \mu_Y)(Z - \mu_Z) + (Z - \mu_Z)^2\right] \\ &= \text{var}(Y) + 2\text{cov}(Y, Z) + \text{var}(Z)\end{aligned}$$

where $\text{cov}(Y, Z)$ denotes the ***covariance*** between Y and Z:

$$\text{cov}(Y, Z) := \mathbb{E}\left[(Y - \mu_Y)(Z - \mu_Z)\right].$$

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

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

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

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

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

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

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:

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

Rescaled covariances define correlations, a concept that is much abused by those who do not understand probability.

Definition. 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 random variables Y and Z are said to be ***uncorrelated*** if $\text{corr}(Y,Z) = 0$.

> **Remark.** Strictly speaking, the variance of a random variable is not well defined unless it has a finite expectation. Similarly, we should not talk about $\text{corr}(Y,Z)$ unless both random variables have well defined variances for which $0 < \text{var}(Y) < \infty$ and $0 < \text{var}(Z) < \infty$.

Example <4.2> When well defined, correlations always lie between $+1$ and -1.

Variances for sums of uncorrelated random variables grow more slowly than might be anticipated. 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

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

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. These facts lead to a useful concentration property.

Example <4.3> Concentration of averages around expected value

Zero correlation is often deduced from independence. A pair of random variables X and Y is said to be ***independent*** if *every event determined by X is independent of every event determined by Y*. For example, independence implies that events such as $\{X \le 5\}$ and $\{7 \le Y \le 18\}$ 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. For the purposes of Stat241, you should not fret about the definition of independence: Just remember to explain why you regard some pieces of information as irrelevant when you calculate conditional probabilities and conditional expectations.

For example, suppose a random variable X can take values $x_1, x_2, \dots$ and that X is independent of another random variable Y. Consider the expected value of a product $g(X)h(Y)$, for any functions g and h. Calculate by conditioning on the possible values taken by X:

$$\mathbb{E}g(X)h(Y) = \sum\nolimits_i \mathbb{P}\{X = x_i\}\mathbb{E}(g(X)h(Y) \mid X = x_i).$$

Given that $X = x_i$, we know that $g(X) = g(x_i)$ but we get no help with understanding the behavior of $h(Y)$. Thus, independence implies

$$\mathbb{E}(g(X)h(Y) \mid X = x_i) = g(x_i)\mathbb{E}(h(Y) \mid X = x_i) = g(x_i)\mathbb{E}h(Y).$$

Deduce that

$$\mathbb{E}g(X)h(Y) = \sum\nolimits_i \mathbb{P}\{X = x_i\}g(x_i)\mathbb{E}h(Y) = \mathbb{E}g(X)\mathbb{E}h(Y).$$

Put another way, if X and Y are independent random variables

$$\text{cov}\big(g(X), h(Y)\big) = \mathbb{E}\big(g(X)h(Y)\big) - (\mathbb{E}g(X))\,(\mathbb{E}h(Y)) = 0.$$

That is, each function of X is uncorrelated with each function of Y. In particular, if X and Y are independent then they are uncorrelated. The converse is not usually true: uncorrelated random variables need not be independent.

Example <4.4> An example of uncorrelated random variables that are dependent

The concentration phenomenon can also hold for averages of dependent random variables.

Example <4.5> Comparison of spread in sample averages for sampling with and without replacement: the Decennial Census.

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 know as the "analysis of variance".

Example <4.6> An example to show how variances can sometimes be decomposed into components attributable to difference sources. (Can be skipped.)

4.2 Things to remember

- $\mathbb{E}g(X)h(Y) = \mathbb{E}g(X)\mathbb{E}h(Y)$ if X and Y are independent random variables
- the 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.

- For constants a, b, c, d, and random variables U, V, Y, Z,

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

- Sampling without replacement gives smaller variances than sampling with replacement.

4.3 The examples

<4.1> **Example.** The Tchebychev inequality asserts: for a random variable X with expected value μ,

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

The inequality becomes obvious if we write F for the event $\{|X - \mu| > \epsilon\}$. First note that $\mathbb{I}_F \le |X-\mu|^2/\epsilon^2$: 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 ϵ^2. It follows that

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

□

<4.2> **Example.** *When well defined, correlations always lies between* $+1$ *and* -1.

Suppose

$$\begin{aligned}&\mathbb{E}Y = \mu_Y \qquad \text{and} \qquad \text{var}(Y) = \sigma_Y^2\\ &\mathbb{E}Z = \mu_Y \qquad \text{and} \qquad \text{var}(Z) = \sigma_Z^2\end{aligned}$$

Define standardized variables

$$Y' = \frac{Y - \mu_Y}{\sigma_Y} \qquad \text{and} \qquad Z' = \frac{Z - \mu_Z}{\sigma_Z}.$$

Note that $\mathbb{E}Y' = \mathbb{E}Z' = 0$ and $\text{var}(Y') = \text{var}(Z') = 1$. Also

$$\text{corr}(Y, Z) = \text{cov}(Y'Z') = \mathbb{E}(Y'Z').$$

Use the fact that variances are always nonnegative to deduce that

$$0 \le \text{var}(Y' + Z') = \text{var}(Y') + 2\text{cov}(Y', Z') + \text{var}(Z') = 2 + 2\text{cov}(Y', Z'),$$

which rearranges to $\text{cov}(Y', Z') \ge -1$. Similarly

$$0 \le \text{var}(Y' - Z') = \text{var}(Y') - 2\text{cov}(Y', Z') + \text{var}(Z') = 2 - 2\text{cov}(Y', Z'),$$

which rearranges to $\text{cov}(Y', Z') \le +1$. □

<4.3> **Example.** Suppose $X_1, \dots, X_n$ are uncorrelated random variables, each with expected value μ and variance σ^2. By repeated application of the formula for the variance of a sum of variables with zero covariances,

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

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

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

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

From the Tchebychev inequality,

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

In particular, for each positive constant C,

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

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

<4.4> **Example.** Consider 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:

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

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

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

□

<4.5> **Example.** Until quite recently, in the Decennial Census of Housing and Population the Census Bureau would obtain some more detailed about the population via 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 would receive the so-called "long form".

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

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

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

Suppose a sample of n housing units is 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

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

Notice that

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

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

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

a quantity that I will denote by σ^2.

If the sample is taken without replacement—which, of course, the Census Bureau had to 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

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

so that 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 <4.3>, we would have $\operatorname{var}\left(\bar{Y}\right) = \sigma^2/n$. 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 inefficiciency in sampling with replacement, as shown by a calculation of variance for sampling without replacement:

$$\begin{aligned}
\operatorname{var}\left(\bar{Y}\right) &= \mathbb{E}\left(\bar{Y} - \bar{y}\right)^2 \\
&= \mathbb{E}\left(\frac{1}{n}\sum\nolimits_{i=1}^{n}(Y_i - \bar{y})\right)^2 \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum\nolimits_{i=1}^{n}(Y_i - \bar{y})^2 + 2\sum\nolimits_{1\le i<j\le n}(Y_i - \bar{y})(Y_j - \bar{y})\right) \\
&= \frac{1}{n^2}\left(\sum\nolimits_{i=1}^{n}\mathbb{E}(Y_i - \bar{y})^2 + 2\sum\nolimits_{1\le i<j\le n}\mathbb{E}\left((Y_i - \bar{y})(Y_j - \bar{y})\right)\right) \\
&= \frac{1}{n^2}\left(\sum\nolimits_{i=1}^{n}\operatorname{var}(Y_i) + \sum\nolimits_{1\le i\ne j\le n}\operatorname{cov}(Y_i, Y_j)\right)
\end{aligned}$$

What formula did I just rederive?

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

$$(*) \qquad \operatorname{var}\left(\bar{Y}\right) = \frac{1}{n^2}\left(n\sigma^2 + n(n-1)c\right) = \frac{\sigma^2}{n} + \frac{(n-1)c}{n}.$$

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

$$c = \operatorname{cov}(Y_1, Y_2) = \frac{1}{N(N-1)}\sum\nolimits_{i\ne j}(y_i - \bar{y})(y_j - \bar{y}).$$

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

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

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

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

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 σ^2/n for $\text{var}(\overline{Y})$ 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 $\overline{Y}$ 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, $\text{var}(\overline{Y}) = 0$ 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. □

<4.6> **Example.** Consider a two stage method for generating a random variable Z. 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 $p_1 + p_2 + \cdots + p_k = 1$. 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

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

From the first formula we get

$$\mathbb{E}Z = \sum\nolimits_i \mathbb{P}\{X = i\}\mathbb{E}\left(Z \mid X = i\right) = \sum\nolimits_i p_i\mu_i,$$

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

$$\text{var}(Z) = \mathbb{E}\left(Z - \bar{\mu}\right)^2 = \sum\nolimits_i p_i\mathbb{E}\left((Z - \bar{\mu})^2 \mid X = i\right).$$

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

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

Taking conditional expectations, we then get

$$\begin{aligned} &\mathbb{E}\left((Z - \bar{\mu})^2 \mid X = i\right) \\ &\quad = \mathbb{E}\left((Z - \bar{\mu}_i)^2 \mid X = i\right) + 2(\mu_i - \bar{\mu})\mathbb{E}\left(Z - \mu_i \mid X = i\right) + (\mu_i - \bar{\mu})^2. \end{aligned}$$

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

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

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

$$\text{var}(Z) = \mathbb{E}\left(\text{var}(Z \mid X)\right) + \text{var}\left(\mathbb{E}(Z \mid X)\right),$$

where $E(Z \mid X)$ denotes the random variable that takes the value μ_i when X takes the value i, and $\text{var}(Z \mid X)$ denotes the random variable that takes the value σ_i^2 when X takes the value i. □

Chapter 5

Normal approximation to the Binomial

5.1 History

In 1733, Abraham de Moivre presented an approximation to the Binomial distribution. He later (de Moivre, 1756, page 242) 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 of how we might show that experimental proportions should be close to their expected values:

> *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.*

De Moivre then stated and proved what is now known as the normal approximation to the Binomial distribution. The approximation itself has subsequently been generalized to give normal approximations for many other distributions. Nevertheless, de Moivre's elegant method of proof is still worth understanding. This Chapter will explain de Moivre's approximation, using modern notation.

A Method of approximating the Sum of the Terms of the Binomial $\overline{a+b}\backslash^{n}$ 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 $\overline{a+b}\backslash^{n}$ 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 ...*

5.2 Pictures of the binomial

Suppose X_n has a Bin(n,p) distribution. That is,

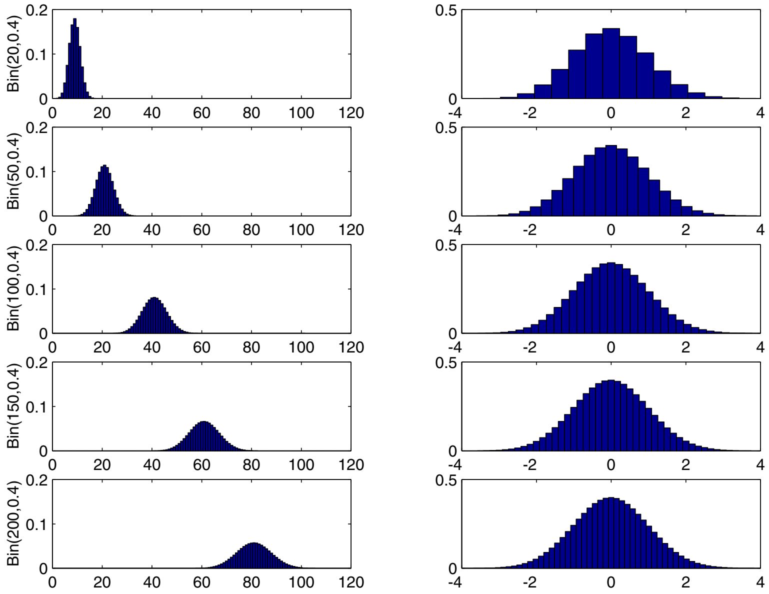
$$b_n(k) := \mathbb{P}\{X_n = k\} = \binom{n}{k} p^k q^{n-k} \qquad \text{for } k = 0, 1, \ldots, n, \text{ where } q = 1 - p,$$

Recall that we can think of X_n as a sum of independent random variables $Y_1+\cdots+Y_n$ with $\mathbb{P}\{Y_i=1\}=p$ and $\mathbb{P}\{Y_i=0\}=q$. From this representation it follows that

$$\begin{aligned}\mathbb{E}X_n &= \sum\nolimits_i \mathbb{E}Y_i = n\mathbb{E}Y_1 = np\\ \operatorname{var}(X_n) &= \sum\nolimits_i \operatorname{var}(Y_i) = n\times\operatorname{var}(Y_1) = npq\end{aligned}$$

Recall also that Tchebychev's inequality suggests the distribution should be clustered around np, with a spread determined by the standard deviation, $\sigma_n := \sqrt{npq}$.

What does the Binomial distribution look like? The plots in the next display, for the Bin$(n, 0.4)$ distribution with $n = 20, 50, 100, 150, 200$, are typical. Each plot on the left shows bars of height $b_n(k)$ and width 1, centered at k. The maxima occur near $n\times 0.4$ for each plot. As n increases, the spread also increases, reflecting the increase in the standard deviations $\sigma_n = \sqrt{npq}$ for $p = 0.4$. Each of the shaded regions on the left has area to one because $\sum_{k=0}^{n} b_n(k) = 1$ for each n.



The plots on the right show represent the distributions of the standardized random variables $Z_n = (X_n - np)/\sigma_n$. The location and scaling effects of the increasing expected values and standard deviations (with $p = 0.4$ and various n) are now removed. Each plot is shifted to bring the location of the maximum close to 0 and the horizontal scale is multiplied by a factor $1/\sigma_n$.

A bar of height $\sigma_n \times b_n(k)$ with width $1/\sigma_n$ is now centered at $(k - np)/\sigma_n$. The plots all have similar shapes. Each shaded region still has area 1.

5.3 De Moivre's argument

Notice how the standardized plots in the last picture settle down to a symmetric 'bell-shaped' curve. You can understand this effect by looking at the ratio of successive terms:

$$\begin{aligned} &b_n(k)/b_n(k-1) \\ &\quad = \left(\frac{n!}{k!(n-k)!}p^k q^{n-k}\right) \Big/ \left(\frac{n!}{(k-1)!(n-k+1)!}p^{k-1}q^{n-k+1}\right) \\ &\quad = (n-k+1)p/(kq) \qquad \text{for } k = 1, 2, \ldots, n. \end{aligned}$$

As a consequence, $b_n(k) \ge b_n(k-1)$ if and only if $(n-k+1)p \ge kq$, that is, iff $(n+1)p \ge k$. For fixed n, the probability $b_n(k)$ achieves its largest value at $k_{\max} = \lfloor (n+1)p \rfloor \approx np$. The probabilities $b_n(k)$ increase with k for $k \le k_{\max}$ then decrease for $k > k_{\max}$. That explains why each plot on the left has a peak near np.

Now for the shape. At least for $k = k_{\max} + i$ near $k_{\max}$ we get a good approximation for the logarithm of the ratio of successive terms using the Taylor approximation: $\log(1+x) \approx x$ for x near 0. Indeed,

$$\begin{aligned} b(k_{\max}+i)/b(k_{\max}+i-1) &= \frac{(n-k_{\max}-i+1)p}{(k_{\max}+i)q} \\ &\approx \frac{(nq-i)p}{(np+i)q} \\ &= \frac{1-i/(nq)}{1+i/(np)} \qquad \text{after dividing through by } npq. \end{aligned}$$

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 taking a product of successive ratios we get the ratio of the individual Binomial probabilities to their largest term. On a log scale the calculation is even simpler. For example, if $m \geq 1$ and $k_{\max} + m \leq n$,

$$
\begin{aligned}
\log & \frac{b(k_{\max}+m)}{b(k_{\max})} \\
&= \log \left(\frac{b(k_{\max}+1)}{b(k_{\max})} \times \frac{b(k_{\max}+2)}{b(k_{\max}+1)} \times \cdots \times \frac{b(k_{\max}+m)}{b(k_{\max}+m-1)} \right) \\
&= \log \frac{b(k_{\max}+1)}{b(k_{\max})} + \log \frac{b(k_{\max}+2)}{b(k_{\max}+1)} + \cdots + \log \frac{b(k_{\max}+m)}{b(k_{\max}+m-1)} \\
&\approx \frac{-1-2-\cdots-m}{npq} \\
&\approx -\tfrac{1}{2}\frac{m^2}{npq}.
\end{aligned}
$$

The last line used the fact that $1+2+3+\cdots+m = \frac{1}{2}m(m+1) \approx \frac{1}{2}m^2$.

In summary,

$$
\mathbb{P}\{X = k_{\max}+m\} \approx b(k_{\max}) \exp\left(-\frac{m^2}{2npq}\right) \qquad \text{for } m \text{ not too large.}
$$

An analogous approximation holds for $0 \leq k_{\max} + m \leq k_{\max}$.

5.4 The largest binomial probability

Using the fact that the probabilities sum to 1, for $p = 1/2$ de Moivre was able to show 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 (de Moivre, 1756, page 244)

> *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 $2\sqrt{nc}$.*

In modern notation, the vital fact discovered by the learned Mr. James Stirling asserts that

$$n! \approx \sqrt{2\pi}\, n^{n+1/2} e^{-n} \qquad \text{for } n = 1, 2, \ldots$$

in the sense that the ratio of both sides tends to 1 (very rapidly) as n goes to infinity. See Feller (1968, pp52-53) for an elegant, modern derivation of the Stirling formula.

By Stirling's formula, for $k = k_{\max} \approx np$,

$$\begin{aligned} b_n(k) &= \frac{n!}{k!(n-k)!} p^k q^{n-k} \\ &\approx \frac{1}{\sqrt{2\pi}} \frac{n^{n+1/2}}{(np)^{np+1/2}(nq)^{nq+1/2}} p^{np} q^{nq} \\ &= \frac{1}{\sqrt{2\pi npq}}. \end{aligned}$$

De Moivre's approximation becomes

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

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

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

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

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

for the interval $k - 1/2 \le x \le k + 1/2$. (The length of the interval is 1, so it does not appear in the previous display.)

Similarly, for each pair of integers with $0 \le a < b \le n$,

$$\mathbb{P}\{a \le X_n \le b\} = \sum\nolimits_{k=a}^{b} b_n(k) \approx \sum\nolimits_{k=a}^{b} \int_{k-1/2}^{k+1/2} f(x)\,dx = \int_{a-1/2}^{b+1/2} f(x)\,dx.$$

A change of variables, $y = (x - np)/\sigma_n$, simplifies the last integral to

$$\frac{1}{\sqrt{2\pi}} \int_{\alpha}^{\beta} e^{-y^2/2} dy \qquad \text{where } \alpha = \frac{a - np - 1/2}{\sigma_n} \text{ and } \beta = \frac{b - np + 1/2}{\sigma_n}.$$

Remark. It usually makes little difference to the approximation if we omit the ±1/2 terms from the definitions of α and β.

5.5 Normal approximations

How does one actually perform a normal approximation? Back in the olden days, I would have interpolated from a table of values for the function

$$\Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-y^2/2} dy,$$

which was found in most statistics texts. For example, if X has a $\text{Bin}(100, 1/2)$ distribution,

$$\begin{aligned}\mathbb{P}\{45 \le X \le 55\} &\approx \Phi\left(\frac{55.5 - 50}{5}\right) - \Phi\left(\frac{44.5 - 50}{5}\right) \\ &\approx 0.8643 - 0.1356 = 0.7287\end{aligned}$$

These days, I would just calculate in R:

```
> pnorm(55.5, mean = 50, sd = 5) - pnorm(44.5, mean = 50, sd = 5)
[1] 0.7286679
```

or use another very accurate, built-in approximation:

```
> pbinom(55,size = 100, prob = 0.5) - pbinom(44,size = 100, prob = 0.5)
[1] 0.728747
```

5.6 Continuous distributions

At this point, the integral in the definition of $\Phi(x)$ is merely a reflection of the Calculus trick of approximating a sum by an integral. Probabilists have taken a leap into abstraction by regarding Φ, or its derivative $\phi(y) := \exp(-y^2/2)/\sqrt{2\pi}$, as a way to define a probability distribution

<5.1> **Definition.** *A random variable Y is said to have a* ***continuous distribution*** *(on $\mathbb{R}$) with* ***density function*** *$f(\cdot)$ if*

$$\mathbb{P}\{a \le Y \le b\} = \int_a^b f(y)\,dy \qquad \text{for all intervals } [a, b] \subseteq \mathbb{R}.$$

Equivalently, for each subset A of the real line,

$$\mathbb{P}\{Y \in A\} = \int_A f(y)\,dy = \int_{-\infty}^{\infty} \mathbb{I}\{y \in A\} f(y)\,dy$$

Notice that f should be a nonnegative function, for otherwise it might get awkward when calculating $\mathbb{P}\{Y \in A\}$ for the set $A = \{y \in \mathbb{R} : f(y) < 0\}$:

$$0 \le \mathbb{P}\{Y \in A\} = \int_A f(y)\,dy \le 0.$$

Remark. By putting A equal to $\mathbb{R}$ we get

$$1 = \mathbb{P}\{-\infty < Y < +\infty\} = \int_{-\infty}^{\infty} f(y)\,dy$$

That is, the integral of a density function over the whole real line equals one.

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 in appropriate places. If a range of integration is not indicated explicitly, it can then always be understood as $-\infty$ to ∞, with the zero density killing off unwanted contributions.

Distributions defined by densities have both similarities to and differences from the sort of distributions I have been considering up to this point in Stat 241/541. All the distributions before now were ***discrete***. They were described by a (countable) discrete set of possible values $\{x_i : i = 1, 2, \dots\}$ that could be taken by a random variable X and the probabilities with which X took those values:

$$\mathbb{P}\{X = x_i\} = p_i \qquad \text{for } i = 1, 2, \dots.$$

For any subset A of the real line

$$\mathbb{P}\{X \in A\} = \sum\nolimits_i \mathbb{I}_{\{x_i \in A\}} \mathbb{P}\{X = x_i\} = \sum\nolimits_i \mathbb{I}_{\{x_i \in A\}} p_i$$

Expectations, variances, and things like $\mathbb{E}g(X)$ for various functions g, could all be calculated by conditioning on the possible values for X.

For a random variable X with a continuous distribution defined by a density f, we have

$$\mathbb{P}\{X = x\} = \int_x^x f(y)\,dy = 0$$

for every $x \in \mathbb{R}$. We cannot hope to calculate a probability by adding up (an uncountable set of) zeros. Instead, as you will see in Chapter 7, we must pass to a limit and replace sums by integrals when a random variable X has a continuous distribution.

5.7 Appendix: The mysterious $\sqrt{2\pi}$

The $\sqrt{2\pi}$ appeared in de Moivre's approximation by way of Stirling's formula. It is slightly mysterious why it appears in that formula. The reason for both appearances is the fact that the constant

$$C := \int_{-\infty}^{\infty} \exp(-x^2/2)\,dx$$

is exactly equal to $\sqrt{2\pi}$, as I now explain.

Equivalently, the constant $C^2 = \iint \exp(-(x^2+y^2)/2)\,dx\,dy$ equal to 2π. (Here, and subsequently, the double integral runs over the whole plane.) We can evaluate this double integral by using a small Calculus trick.

Using the fact that

$$\int_0^{\infty} \mathbb{I}\{r \le z\} e^{-z}\,dz = e^{-r} \qquad \text{for } r > 0,$$

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

$$\begin{aligned} C^2 &= \iint \left(\int_0^{\infty} \mathbb{I}\{x^2+y^2 \le 2z\} e^{-z}\,dz \right) dx\,dy \\ &= \int_0^{\infty} \left(\iint \mathbb{I}\{x^2+y^2 \le 2z\}\,dx\,dy \right) e^{-z}\,dz. \end{aligned}$$

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, $C = \sqrt{2\pi}$.

References

de Moivre, A. (1756). *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play* (Third ed.). New York: Chelsea. Third edition (fuller, clearer, and more correct than the former), reprinted in 1967. First edition 1718.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (third ed.), Volume 1. New York: Wiley.

Chapter 6

Central limit theorems

6.1 Overview

Recall that a random variable Z is said to have a ***standard normal*** distribution, denoted by $N(0,1)$, if it has a continuous distribution with density

$$\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2) \qquad \text{for } -\infty < z < \infty.$$

That is, for all intervals $[a,b]$,

$$\mathbb{P}\{a \le Z \le b\} = \int_a^b \phi(z)\,dz,$$

and, for each subset A of the real line, $\mathbb{P}\{Z \in A\} = \int_A \phi(z)\,dz$. In particular, for each fixed b we must have $\mathbb{P}\{Z = b\} = \int_b^b \phi(z)\,dz = 0$.

More generally, for $\mu \in \mathbb{R}$ and $\sigma > 0$, a random variable X is said to have a $N(\mu,\sigma^2)$ ***distribution*** if $(X-\mu)/\sigma$ has a $N(0,1)$ distribution. That is,

$$\begin{aligned}\mathbb{P}\{a \le X \le b\} &= \mathbb{P}\{(a-\mu)/\sigma \le (X-\mu)/\sigma) \le (b-\mu)/\sigma\}\\ &= \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \phi(z)\,dz\\ &= \int_a^b f_{\mu,\sigma}(x)\,dx\end{aligned}$$

where

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

In other words, X has a continuous distribution with density $f_{\mu,\sigma}(x)$.

> **Remark.** In Chapter 7 you will see that if Z has a $N(0,1)$ distribution then $\mathbb{E}Z = 0$ and $\text{var}(Z) = 1$. Consequently, if X has a $N(\mu,\sigma^2)$ distribution then $\mathbb{E}X = \mu$ and $\text{var}(X) = \sigma^2$.

The normal approximation to the Binomial distribution also implies a normal approximation for the distribution of some other random variables.

Example <6.1> A normal approximation for a sample median

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, and if $\mathbb{E}X = \mu$ and $\text{var}(X) = \sigma^2$, then the standardized variable $(X-\mu)/\sigma$ has approximately a standard normal distribution. Equivalently, X is approximately $N(\mu,\sigma^2)$ distributed.

If you are interested in the reasons behind the success of normal approximation, see the Appendix to Chapter 8 for an outline of a proof of the central limit theorem.

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 offhand references to the central limit theorem to mollify anyone who doubts that all distributions are normal. That said, let me also note that 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 <6.2> A hidden normal approximation—the boxplot

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

Example <6.3> Normal approximations for sample means

6.2 The examples

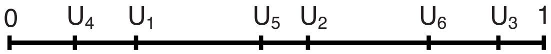
<6.1> **Example.** Suppose $U_1, \dots, U_n$ are independent random variables each distributed Uniform$(0,1)$. That is,

$$\mathbb{P}\{a \le U_i \le b\} = b - a \qquad \text{for all } 0 < a \le b < 1.$$

The corresponding density function is $f(z) = \mathbf{1}\{0 < z < 1\}$.

For simplicity suppose n is even, $n = 2k$. The sample median M_n is defined as the kth smallest when the U_i's are arranged in increasing order.

> **Remark.** Some authors would define M_n as the $(k+1)$st smallest or as some value between the kth and $(k+1)$st. It doesn't make much difference when n is large.



For example, if $n = 6$ and the U_i's are as shown then M_n would be equal to U_5. For another realization it would probably be equal to another U_i.

Now consider any fixed y in $(0,1)$. Write N_y for the number of U_i's that are $\le y$. More formally,

$$N_y = \sum\nolimits_{i \le n} \mathbf{1}\{U_i \le y\}.$$

The random variable N_y counts the number of "successes" (the number of U_i's that are $\le y$) in n independent trials; N_y has Bin(n, y) distribution, with expected value ny and variance $ny(1-y)$. The key thing to notice is:

$$N_y \ge k \qquad \text{iff} \quad \text{“at least } k \text{ of the } U_i\text{'s are} \le y \quad \text{iff} \quad M_n \le y.$$

Thus

$$\begin{aligned}\mathbb{P}\{M_n \le y\} &= \mathbb{P}\{N_y \ge k\} \\ &= \mathbb{P}\left\{\frac{N_y - ny}{\sqrt{ny(1-y)}} \ge \frac{k - ny}{\sqrt{ny(1-y)}}\right\}.\end{aligned}$$

Use the normal approximation for the distribution of the standardized variable $(N_y - ny)/\sqrt{ny(1-y)}$ to deduce that the last probability is approximately equal to

$$\int_\gamma^\infty \phi(y)\,dy = 1 - \Phi(\gamma) \qquad \text{where } \gamma := (k - ny)/\sqrt{ny(1-y)}.$$

Now consider a special value, $y = (1 + x/\sqrt{n})/2$, for a fixed x. When n is large enough we certainly have $y \in (0, 1)$. This choice also gives

$$ny(1-y) = \frac{n}{4}\left(1 - \frac{x^2}{n}\right) \approx \frac{n}{4}$$

and

$$k - ny = -x\sqrt{n}/2,$$

implying $\gamma \approx -x$ and

$$\mathbb{P}\{M_n \le (1 + x/\sqrt{n})/2\} \approx 1 - \Phi(-x) = \Phi(x).$$

For the last equality I have used the symmetry of ϕ around zero to deduce that $\int_{-x}^{\infty} \phi(y)\,dy = \int_{-\infty}^{x} \phi(y)\,dy$.

Put another way,

$$\mathbb{P}\{2\sqrt{n}(M_n - 1/2) \le x\} \approx \Phi(x)$$

which shows that $2\sqrt{n}(M_n - 1/2)$ is approximately $N(0, 1)$ distributed.

> **Remark.** It might be more convincing to use the approximation twice, first with $x = b$ and then with $x = a$, where $a < b$, then subtract.

That is, M_n has approximately a $N(1/2, 1/(4n))$ distribution. □

<6.2> **Example.** The boxplot provides a convenient way of summarizing data (such as grades in Statistics 241/541). The method is:

(i) arrange the data in increasing order

(ii) find the split points

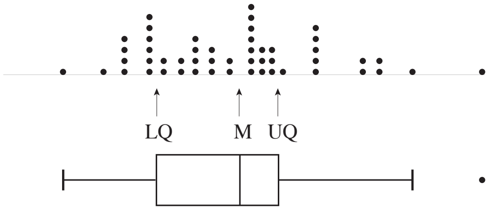
$$\begin{aligned} \text{LQ} &= \text{lower quartile: 25\% of the data smaller than LQ} \\ \text{M} &= \text{median: 50\% of the data smaller than M} \\ \text{UQ} &= \text{upper quartile: 75\% of the data smaller than UQ} \end{aligned}$$

(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 × IQR) and LQ − (1.5 × 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 $\mathbb{P}\{W \le x\}$, where W has a $N(\mu, \sigma^2)$ distribution. From normal tables (or a computer),

$$\mathbb{P}\{W \le \mu + .675\sigma\} \approx .75 \qquad \text{and} \qquad \mathbb{P}\{W \le \mu - .675\sigma\} \approx .25$$

and, of course, $\mathbb{P}\{W \le \mu\} = .5$. For the sample we should expect

$$\text{LQ} \approx \mu - .675\sigma \qquad \text{and} \qquad \text{UQ} \approx \mu + .675\sigma \qquad \text{and} \qquad \text{M} \approx \mu$$

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),

$$\mathbb{P}\{W \le \mu - 2.70\sigma\} = \mathbb{P}\{W \ge \mu + 2.70\sigma\} = .003$$

Only about 0.6% of the sample should be out beyond the whiskers. □

<6.3> **Example.** Chapter 4 gave the expected value and variance of a sample mean $\bar{Y}$ for a sample of size n (with replacement) from a finite population labelled $1, \dots, N$ with "values of interest" $y_1, y_2, \dots, y_N$:

$$\mathbb{E}\overline{Y} = \overline{y} = \frac{1}{N} \sum\nolimits_{i=1}^{N} y_i.$$

For sampling with replacement,

$$\text{var}(\overline{Y}) = \sigma^2/n \qquad \text{where } \sigma^2 = \sum\nolimits_{i=1}^{N} (y_i - \overline{y})^2 / N.$$

The standardized random variable $(\overline{Y}-\overline{y})/\sqrt{\sigma^2/n}$ is well approximated by the $N(0,1)$. Thus

$$\mathbb{P}\left\{-\frac{1.96\sigma}{\sqrt{n}} \le \overline{Y}-\overline{y} \le \frac{1.96\sigma}{\sqrt{n}}\right\} \approx \Phi(1.96)-\Phi(-1.96) \approx 0.95.$$

Before we sample, we can assert that we have about a 95% chance of getting a value of $\overline{Y}$ in the range $\overline{y} \pm 1.96\sigma/\sqrt{n}$. (For the post-sampling interpretation of the approximation, you should take Statistics 242/542.)

Of course, we would not know the value σ, so it must be estimated. How?

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 me too far beyond 241/541.) □

Chapter 7

Continuous Distributions

7.1 Overview

In Chapter 5 you met your first example of a continuous distribution, the normal. Recall the general definition.

Densities
A random variable X is said to have a ***continuous distribution*** (on $\mathbb{R}$) with ***density function*** $f(\cdot)$ if

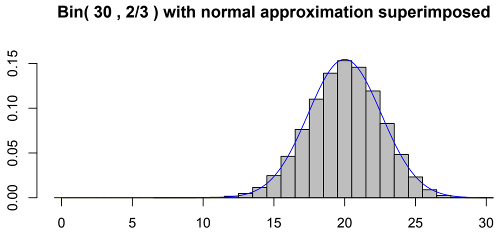
(i) f is a nonnegative function on the real line for which $\int_{-\infty}^{+\infty} f(x)\,dx = 1$

(ii) for each subset A of the real line,

$$\mathbb{P}\{X \in A\} = \int_A f(x)\,dx = \int_{-\infty}^{\infty} \mathbb{I}\{x \in A\} f(x)\,dy$$

Assumption (ii) is actually equivalent to its special case:

$$\mathbb{P}\{a \le X \le b\} = \int_a^b f(x)\,dx \qquad \text{for all intervals } [a,b] \subseteq \mathbb{R}.$$


Bin(30 , 2/3) with normal approximation superimposed

For the normal approximation to the $\text{Bin}(n, p)$ the density was

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

with $\mu = np$ and $\sigma^2 = npq$. That is, f is the $N(\mu, \sigma^2)$ density.

> **Remark.** As you will soon learn, the $N(\mu, \sigma^2)$ distribution has expected value μ and variance σ^2.

Notice that a change of variable $y = (x - \mu)/\sigma$ gives

$$\int_{-\infty}^{\infty} f(x)\,dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}\,dy,$$

which (see Chapter 5) equals 1.

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

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

Equivalently,

$$\mathbb{P}\{a \le U \le b\} = \int_a^b f(x)\,dx \qquad \text{for all real } a, b,$$

where

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

I will use the Uniform to illustrate several general facts about continuous distributions.

> **Remark.** Of course, to actually simulate a 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 Uniform[0,1] by a discrete distribution placing probabilities of about 10^{-7} on a fine grid of about 10^7 equi-spaced points in the interval. You might think of the Uniform$[0, 1]$ as a convenient idealization of the discrete approximation.

Be careful not to confuse the density $f(x)$ with the probabilities $p(y) = \mathbb{P}\{Y = y\}$ used to specify ***discrete distributions***, that is, distributions for random variables that can take on only a finite or countably infinite set of different values. The Bin(n, p) and the geometric(p) are both discrete distributions. Continuous distributions smear the probability out over a continuous range of values. In particular, if X has a continuous distribution with density f then

$$\mathbb{P}\{X = t\} = \int_t^t f(x)\,dx = 0 \qquad \text{for each fixed } t.$$

The value $f(x)$ does not represent a probability. Instead, the values taken by the density function could be thought of as constants of proportionality. At least at points where the density function f is continuous and when δ is small,

$$\mathbb{P}\{t \le X \le t + \delta\} = \int_t^{t+\delta} f(x)\,dy = f(t)\delta + \text{ terms of order } o(\delta).$$

Remark. Remember that $g(\delta) = o(\delta)$ means that $g(\delta)/\delta \to 0$ as $\delta \to 0$.

Equivalently,

$$\lim_{\delta\to 0} \frac{1}{\delta}\mathbb{P}\{t \le X \le t + \delta\} = f(t).$$

Some texts define the density as the derivative of the ***cumulative distribution function***

$$F(t) = \mathbb{P}\{-\infty < X \le t\} \qquad \text{for } -\infty < t < \infty.$$

That is,

$$f(t) = \lim_{\delta\to 0} \frac{1}{\delta}\big(F(t + \delta) - F(t)\big)$$

This approach works because

$$\begin{aligned}
&\mathbb{P}\{t \le X \le t + \delta\}\\
&\qquad = \mathbb{P}\{X \le t + \delta\} - \mathbb{P}\{X < t\}\\
&\qquad = F(t + \delta) - F(t) \qquad \text{because } \mathbb{P}\{X = t\} = 0.
\end{aligned}$$

> **Remark.** Evil probability books often refer to random variables X that have continuous distributions as "continuous random variables", which is misleading. If you are thinking of a random variable as a function defined on a sample space, the so-called continuous random variables need not be continuous as functions.
>
> Evil probability books often also explain that distributions are called continuous if their distribution functions are continuous. A better name would be non-atomic: if X has distribution function F and if F has a jump of size p at x then $\mathbb{P}\{X = x\} = p$. Continuity of F (no jumps) implies no atoms, that is, $\mathbb{P}\{X = x\} = 0$ for all x. It is sad fact of real analysis life that continuity of F does not imply that the corresponding distribution is given by a density. Fortunately, you won't be meeting such strange beasts in this course.

When we are trying to determine a density function, 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 <7.1> The distribution of $\tan(X)$ if $X \sim \text{Uniform}(-\pi/2, \pi/2)$

I recommend that you remember the method used in the previous 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 <7.2> Smooth functions of a random variable with a continuous distribution

Calculations with continuous distributions typically involve integrals or derivatives where discrete distribution involve sums or probabilities attached to individual points. The formulae developed in previous chapters for expectations and variances of random variables have analogs for continuous distributions.

Example <7.3> Expectations of functions of a random variable with a continuous distribution

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.

Example <7.4> Expected value and variance for the $N(\mu, \sigma^2)$.

Calculations for continuous distributions are often simpler than analogous calculations for discrete distributions because we are able to ignore some pesky cases.

Example <7.5> Zero probability for ties with continuous distributions.

Calculations are also greatly simplified by the fact that we can ignore contributions from higher order terms when working with continuous distributions and small intervals.

Example <7.6> 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

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

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(α, β) distribution is defined by the density function

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

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

As you just saw in Example <7.6>, the kth order statistic from a sample of n independently generated random variables with Uniform$[0, 1]$ distributions has a Beta$(k, n-k+1)$ distribution.

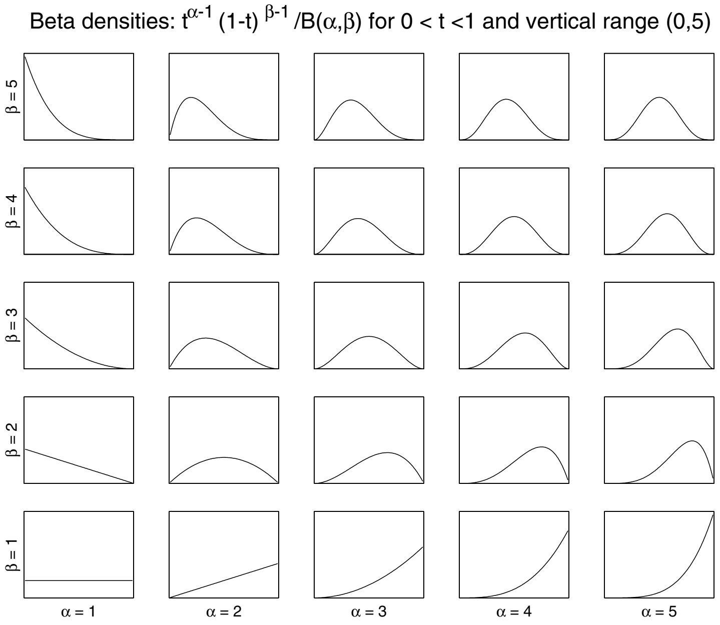
The function *beta()* in R calculates the value of the beta function:

```
> beta(5.5,2.7)
[1] 0.01069162
> ?beta   # get help for the beta() function
```

Also, there is a set of R functions that gives useful results for the beta density. For example, the pictures on the next page could be drawn by a series of R commands like:

```
> jj=(1:1000)/1000
> plot(jj,dbeta(jj,2,3),type="l")
```

The functions *dbeta()* calculates the values of the beta density at a fine grid of points. The *plot()* function is called with the option of joining the points by a smooth curve.



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

Example <7.7> Binomial tail probabilities from beta distributions.

7.2 Things to remember

- The density function $f(\cdot)$ gives the constants of proportionality, and not probabilities: $f(x)$ is not the same as $\mathbb{P}\{X = x\}$, which is zero for every x if X has a continuous distribution.
- A density function, f, must be non-negative and it must integrate to one over the whole line, $1 = \int_{-\infty}^{\infty} f(t)\,dt$.
- Expected value of a function of a random variable with a continuous distribution: if the distribution of X has density f then
$$\mathbb{E}H(X) = \int_{-\infty}^{\infty} H(x)f(x)\,dx$$
- 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.

7.3 Examples for Chapter 7

<7.1> **Example.** *The distribution of* $\tan(X)$ *if* $X \sim \text{Uniform}(-\pi/2, \pi/2)$

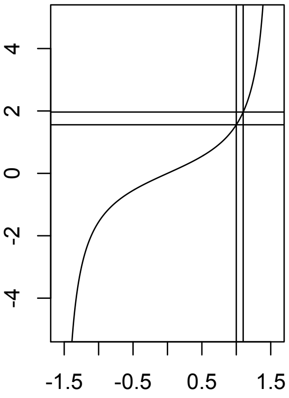
The distribution of X is continuous with density

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

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 positive δ, we have

$$y \le Y \le y + \delta \qquad \text{if and only if} \qquad x \le X \le x + \epsilon, \tag{$*$}$$

where x and ϵ are related to y and δ by the equalities



$$y = \tan(x) \qquad \text{AND} \qquad y + \delta = \tan(x + \epsilon).$$

By Calculus, for small δ,

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

Compare with the definition of the derivative:

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

Thus

$$\begin{aligned}\mathbb{P}\{y\le Y\le y+\delta\}&=\mathbb{P}\{x\le X\le x+\epsilon\}\\&\approx \epsilon f(x)\\&\approx\frac{\delta\cos^2 x}{\pi}.\end{aligned}$$

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

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

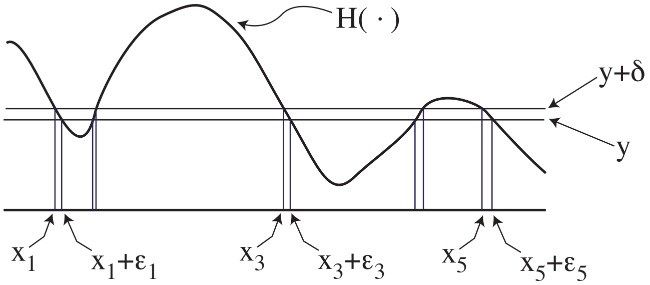
Thus Y has a continuous distribution with density

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

The continuous distribution with this density is called the ***Cauchy***. □

<7.2> **Example.** For functions that are not one-to-one, the analog of the method from Example <7.1> can require a little more work. In general, we can consider a random variable Y defined as $H(X)$, a function of another random variable. If X has a continous 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 for the *tan* function.

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 $[x_1, x_1+\epsilon_1]$, $[x_2, x_2+\epsilon_2]$, ... The number of pieces in the X range might be different for different values of y.



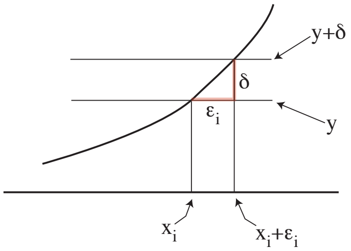
From the representation of $\{y \le Y \le y+\delta\}$ as a disjoint union of events

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

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

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

For each small interval, the ratio of δ/ϵ_i is close to the absolute value of 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,

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

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

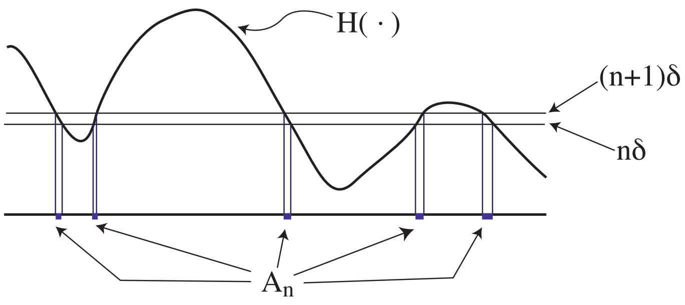
$$\frac{f(x_1)}{|H'(x_1)|} + \frac{f(x_2)}{|H'(x_2)|} + \dots$$

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

<7.3> **Example.** *Expectations of functions of a random variable with a continuous distribution*

Suppose X has a continuous distribution with density function f. Let $Y = H(X)$ be a new random variable, defined as a function of X. We can calculate $\mathbb{E}Y$ by an approximation argument similar to the one used in Example <7.2>. It will turn out that

$$\mathbb{E}H(X) = \int_{-\infty}^{+\infty} H(x)f(x)\,dx.$$



Cut the range of values that might be taken by Y into disjoint intervals of the form $n\delta \le y < (n+1)\delta$, for an arbitrarily small, positive δ. Write A_n for the corresponding set of x values. That is, for each x in $\mathbb{R}$,

$$n\delta \le H(x) < (n+1)\delta \qquad \text{if and only if} \qquad x \in A_n.$$

We now have simple upper and lower bounds for H:

$$H_\delta(x) \le H(x) \le \delta + H_\delta(x) \qquad \text{for every real } x$$
$$\text{where } H_\delta(x) = \sum\nolimits_n n\delta \mathbf{1}\{x \in A_n\}.$$

(You should check the inequalities when $x \in A_n$, for each possible integer n.) Consequently

$$\mathbb{E}H_\delta(X) \le \mathbb{E}H(X) \le \delta + \mathbb{E}H_\delta(X)$$

and

$$\int_{-\infty}^{+\infty} H_\delta(x)f(x)\,dx \le \int_{-\infty}^{+\infty} H(x)f(x)\,dx \le \delta + \int_{-\infty}^{+\infty} H_\delta(x)f(x)\,dx.$$

More concisely,

$$(\star) \qquad |\mathbb{E}H_\delta(X) - \mathbb{E}H(X)| \le \delta$$

and

$$(\star\star) \qquad |\int_{-\infty}^{+\infty} H_\delta(x)f(x)\,dx - \int_{-\infty}^{+\infty} H(x)f(x)\,dx| \le \delta.$$

The random variable $H_\delta(X)$ has a discrete distribution whose expectation you know how to calculate:

$$\begin{aligned}\mathbb{E}H_\delta(X) &= \mathbb{E}\sum\nolimits_n n\delta \mathbf{1}\{X \in A_n\} && \text{expectation of a countable sum}\\ &= \sum\nolimits_n n\delta \mathbb{P}\{X \in A_n\} && \text{because } \mathbb{E}\mathbf{1}\{X \in A_n\} = \mathbb{P}\{X \in A_n\}\\ &= \sum\nolimits_n n\delta \int_{-\infty}^{+\infty} \mathbf{1}\{x \in A_n\} f(x)\,dx && \text{definition of } f\\ &= \int_{-\infty}^{+\infty} H_\delta(x) f(x)\,dx && \text{take sum inside integral.}\end{aligned}$$

From the inequalities ($\star$) and ($\star\star$), the last equality deduce that

$$|\mathbb{E}H(X) = \int_{-\infty}^{+\infty} H(x)f(x)\,dx| \le 2\delta$$

for arbitrarily small $\delta > 0$. The asserted representation for $\mathbb{E}H(X)$ follows. □

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

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

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

<7.4> **Example.** *Expected value and variance* $N(\mu, \sigma^2)$.

If $X \sim N(\mu, \sigma^2)$ its density function is

$$\begin{aligned}f(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) && \text{for } -\infty < x < \infty\\ &= \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) && \text{where } \phi(y) := (2\pi)^{-1/2}\exp(-y^2/2).\end{aligned}$$

Calculate, using a change of variable $y = (x-\mu)/\sigma$.

$$\begin{aligned}\mathbb{E}X &= \int_{-\infty}^{+\infty} x f(x)\,dx\\ &= \int_{-\infty}^{+\infty} (\mu + \sigma y)\phi(y)\,dy\\ &= \mu\int_{-\infty}^{+\infty} \phi(y)\,dy + \sigma\int_{-\infty}^{+\infty} y\phi(y)\,dy\\ &= \mu.\end{aligned}$$

The second integral vanishes because $y\phi(y) = -(-y)\phi(-y)$.

Similarly

$$\begin{aligned}\text{var}(X) &= \mathbb{E}(X-\mu)^2 \\ &= \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,dx \\ &= \sigma^2 \int_{-\infty}^{+\infty} y^2\phi(y)\,dy \\ &= \sigma^2\end{aligned}$$

using integration by parts and $\frac{d}{dy}\phi(y) = -y\phi(y)$.

□

<7.5> **Example.** Suppose X and Y are independent random variables, each with a Uniform$[0,1]$ distribution. Show that $\mathbb{P}\{X = Y\} = 0$.

The event $\{X = Y = 1\}$ is a subset of $\{X = 1\}$, which has zero probability. The other possibilities are almost as easy to dispose of: for each positive integer n,

$$\{X = Y < 1\} \subset \cup_{i=0}^{n-1}\{i/n \le X < (i+1)/n \text{ and } i/n \le Y < (i+1)/n\}$$

a disjoint union of events each with probability $1/n^2$, by independence. Thus

$$\mathbb{P}\{X = Y < 1\} \le n(1/n^2) = 1/n \qquad \text{for every } n.$$

It follows that $\mathbb{P}\{X = Y\} = 0$.

A similar calculation shows that $\mathbb{P}\{X = Y\} = 0$ for independent random variables with any pair of continuous distributions. □

<7.6> **Example.** *The distribution of the order statistics from the uniform distribution.*

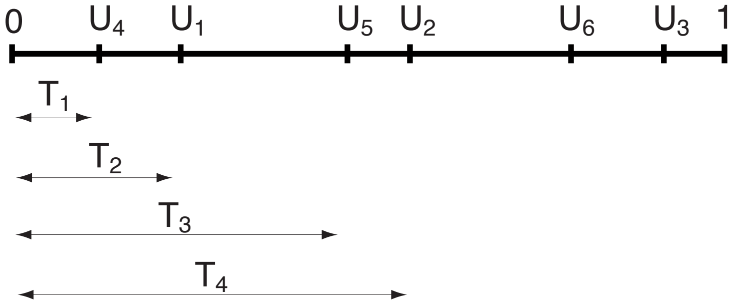
Suppose $U_1, U_2, \dots, U_n$ are independent random variables, each with distribution Uniform$(0,1)$. That is,

$$\mathbb{P}\{a \le U_i \le b\} = \int_a^b h(x)\,dx \qquad \text{for all real } a \le b,$$

where

$$h(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The U_i's define n points in the unit interval. If we measure the distance of each point from 0 we obtain random variables $0 \le T_1 < T_2 < \cdots < T_n \le 1$, the values $U_1, \dots, U_n$ rearranged into increasing order. (Example <7.5> lets me ignore ties.) 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.

The random variable T_k, the kth smallest of the ordered values, is usually called the kth ***order statistic***. It takes a continuous range of values. It has a continuous distribution. What is its density function?

For a very short interval $[t, t+\delta]$, with $0 < t < t+\delta < 1$ and δ very small, we need to show that $\mathbb{P}\{t \le T_k \le t+\delta\}$ is roughly proportional to δ, 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 $t \le T_k \le t+\delta$ we must have $N \ge 1$. If $N = 1$ then we must have exactly $k-1$ points in $[0,t)$ to get $t \le T_k \le t+\delta$. If $N \ge 2$ then it becomes more complicated to describe all the ways that we would get $t \le T_k \le t+\delta$. Luckily for us, the contributions from all those complicated expressions will turn out to be small enough to ignore if δ is small. Calculate.

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

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

$$F_2 = \{N \ge 2\} \cap \{t \le T_k \le t+\delta\}$$

is a subset of the union

$$\cup_{1<i<j\le n}\{U_i,\, U_j \text{ both in } [t, t+\delta]\,\}$$

Put another way,

$$\mathbb{I}_{F_2} \le \sum\nolimits_{1\le i<j\le n} \mathbb{I}\{U_i,\, U_j \text{ both in } [t, t+\delta]\,\}.$$

Take expectations of both sides to deduce that

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

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

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

Thus $\mathbb{P}F_2 \le \binom{n}{2}\delta^2$, which tends to zero much faster than δ as $\delta \to 0$. (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 F_1 into disjoint events 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 event has probability

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

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

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

contributions 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

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

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

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

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

> **Remark.** 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

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

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.

<7.7> **Example.** *Binomial tail probabilities from beta distributions.*

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

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

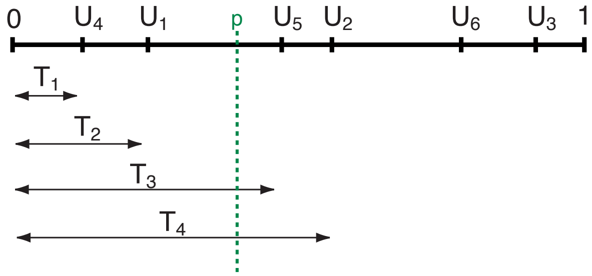
With a computer (using R, for example) such a task would not be as arduous as it used to be back in the days of hand calculation. We could also use a normal approximation (as in the example for the median in Chapter 6). However, there is another method based on the facts about the order statistics, which gives an exact integral expression for the Binomial tail probability.

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

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

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 Bin(n, p) distribution.

As in Example <7.6>, write T_k for the kth 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,

$X_n \ge k$ if and only if $T_k \le p$.

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

$$\mathbb{P}\{X_n \ge k\} = \mathbb{P}\{T_k \le p\} = \frac{n!}{(k-1)!(n-k)!} \int_0^p t^{k-1}(1-t)^{n-k}\, dt.$$

The density for the distribution of T_k comes from Example <7.6>. □

Chapter 8

Conditioning on a random variable with a continuous distribution

8.1 Overview

At this point in the course I hope you appreciate the usefulness of the discrete conditioning formula,

$$\mathbb{E}\big(Y\big) = \sum\nolimits_i \mathbb{P}\left(F_i\right) \mathbb{E}\left(Y \mid F_i\right),$$

for a finite or countably infinite collection of disjoint events $F_1, F_2, \dots$ with $\sum_i \mathbb{P}F_i = 1$. As a particular case, if X is a random variable that takes only a discrete set of values $\{x_1, x_2, \dots\}$ then

<8.1> $$\mathbb{E}\big(Y\big) = \sum\nolimits_i \mathbb{P}\{X = x_i\}\mathbb{E}\big(Y \mid X = x_i\big).$$

This formula can be made to look simpler by the introduction of the function $g(x) = \mathbb{E}\big(Y \mid X = x\big)$, for then

<8.2> $$\mathbb{E}\big(Y\big) = \sum\nolimits_i \mathbb{P}\{X = x_i\}g(x_i) = \mathbb{E}\big(g(X)\big).$$

Throughout the course I have been working with examples where you could figure out things like $\mathbb{E}(Y \mid X = x)$ or $\mathbb{P}(A \mid F)$ by identifying the probabilistic mechanism corresponding to the conditional probability distribution $\mathbb{P}(\cdot \mid X = x\}$ or $\mathbb{P}(\cdot \mid F)$. In a few cases, you could also have calculated directly from

<8.3> $$\mathbb{P}(A \mid F) = \frac{\mathbb{P}(AF)}{\mathbb{P}F} \text{ or } \mathbb{E}(Y \mid F) = \frac{\mathbb{E}(Y\mathbf{1}_F)}{\mathbb{P}F}.$$

Of course this formula only makes sense if $\mathbb{P}F \neq 0$.

If the random variable X has a continuous distribution, you still have the possibility of calculating things like $\mathbb{E}(Y \mid X = x)$ and $\mathbb{P}(A \mid X = x)$ by recognizing the probabilistic mechanism corresponding to $\mathbb{P}(\cdot \mid X = x)$. But you won't have much luck in putting $F = \{X = x\}$ in <8.3> because $\mathbb{P}\{X = x\} = 0$.

Nevertheless there is a formula similar to <8.2> that works when X has a continuous distribution with density function f. Section 8.2 (which could be skipped) explains in more detail why

<8.4> $$\mathbb{E}\left(Y\right) = \int_{-\infty}^{\infty} \mathbb{E}(Y \mid X = x) f(x)\, dx.$$

As a special case, when Y equals the indicator function of an event B the formula reduces to

$$\mathbb{P}B = \int_{-\infty}^{\infty} \mathbb{P}(B \mid X = x) f(x)\, dx.$$

Moreover, for most values of z, the conditional expectation can be calculated as a limit of a ratio,

<8.5> $$\mathbb{E}(Y \mid X = z) = \lim_{\epsilon \to 0} \mathbb{E}(Y \mid z \le X \le z + \epsilon) = \lim_{\epsilon \to 0} \frac{\mathbb{E}(Y\mathbf{1}\{z \le X \le z + \epsilon\}}{\mathbb{P}\{z \le X \le z + \epsilon\}}.$$

The final ratio is amenable to discretization methods like those introduced in Chapter 7.

*8.2 *Some justification for the conditioning formulae

You could skip this Section if you are prepared to believe <8.4> and <8.5>.

Here is an argument that many authors, including me, find convincing as a way to get the formula <8.4>. We start from an assumption about conditional expectations that has intuitive appeal:

For each subset A of the real line, if α and β are constants (depending on A, of course) for which

$$\alpha \le \mathbb{E}(Y \mid X = x) \le \beta \qquad \text{for all } x \in A,$$

then we should have

$$\alpha \le \mathbb{E}(Y \mid X \in A) \le \beta.$$

Suppose the assumption is valid. As before write $g(x)$ for $\mathbb{E}(Y \mid X = x)$. For a small $\delta > 0$ and each integer n define

$$A_n = \{x \in \mathbb{R} : n\delta \le g(x) < (n+1)\delta\}.$$

Define a 'step function' $g_\delta(x) = \sum_n n\delta \mathbf{1}\{x \in A_n\}$, which approximates g very well, in the sense that

<8.6> $$g_\delta(x) \le g(x) < g_\delta(x) + \delta \qquad \text{for every real number } x.$$

By definition of the set A_n,

$$n\delta \le \mathbb{E}(Y \mid X = x) \le (n+1)\delta \qquad \text{for all } x \in A_n.$$

The conditioning assumption (in the box) then implies

<8.7> $$n\delta \le \mathbb{E}(Y \mid X \in A_n) \le (n+1)\delta \qquad \text{for each } n.$$

By the discrete conditioning formula <8.1> with $F_n = \{X \in A_n\}$,

$$\mathbb{E}Y = \sum\nolimits_n \mathbb{P}\{X \in A_n\}\mathbb{E}(Y \mid X \in A_n).$$

Writing $\int_{A_n} f(x)\,dx = \int_{-\infty}^{+\infty} \mathbf{1}\{x \in A_n\}f(x)\,dx$ for $\mathbb{P}\{X \in A_n\}$ and using the lower bound from <8.7> we then get

$$\begin{aligned}\mathbb{E}Y &\ge \sum\nolimits_n \int_{-\infty}^{+\infty} n\delta\mathbf{1}\{x \in A_n\}f(x)\,dx \\ &= \int_{-\infty}^{+\infty} \sum\nolimits_n n\delta\mathbf{1}\{x \in A_n\}f(x)\,dx \qquad \text{legit?} \\ &= \int_{-\infty}^{+\infty} g_\delta(x)f(x)\,dx.\end{aligned}$$

A similar argument gives the upper bound

$$\mathbb{E}Y \le \int_{-\infty}^{+\infty} \left(g_\delta(x) + \delta\right) f(x)\,dx = \delta + \int_{-\infty}^{+\infty} g_\delta(x)f(x)\,dx.$$

In summary,

$$\int_{-\infty}^{+\infty} g_\delta(x)f(x)\,dx \le \mathbb{E}Y \le \delta + \int_{-\infty}^{+\infty} g_\delta(x)f(x)\,dx.$$

Compare with the inequality that results when we multiply both sides of <8.6> by $f(x)$ then integrate:

$$\int_{-\infty}^{+\infty} g_\delta(x)f(x)\,dx \le \int_{-\infty}^{+\infty} g(x)f(x)\,dx \le \delta + \int_{-\infty}^{+\infty} g_\delta(x)f(x)\,dx.$$

The upper and lower bounds in the last two inequalities are the same. Both $\mathbb{E}Y$ and $\int_{-\infty}^{+\infty} g(x)f(x)\,dx$ lie in an interval of length δ. The two quantities can differ by at most δ, no matter how small we choose $\delta > 0$. Equality <8.4> follows.

Equality <8.4> has a small extension based on the idea that functions of X should behave like constants when we condition on $X = x$:

<8.8>
$$\begin{aligned}\mathbb{E}\left(YH(X)\right) &= \int_{-\infty}^{+\infty} \mathbb{E}(YH(X) \mid X = x)f(x)\,dx \\ &= \int_{-\infty}^{+\infty} H(x)\mathbb{E}(Y \mid X = x)f(x)\,dx \\ &= \int_{-\infty}^{+\infty} H(x)g(x)f(x)\,dx.\end{aligned}$$

In particular, if $H(x)$ is the indicator function of an interval $[z, z+\epsilon]$, for a fixed z and a small $\epsilon > 0$ then

$$\mathbb{E}\left(Y\mathbf{1}\{z \le X \le z+\epsilon\}\right) = \int_z^{z+\epsilon} g(x)f(x)\,dx = g(z)\int_z^{z+\epsilon} f(x)\,dx + o(\epsilon)$$

if g is continuous at z. The inequality rearranges to give <8.5>.

Remark. In advanced probability theory, the abstract treatment of conditional expectations starts by taking <8.8> as a desirable property. One then shows, using measure theoretic arguments, that there exists a function $g(x)$ (which is uniquely determined up to trivial changes on very small sets) for which the desired property holds. One declares that function to be the definition of $\mathbb{E}(Y \mid X = x)$. From that starting point, one then goes on to verify the assumptions that I have taken as axiomatic for Stat 241/541.

The benefit is mathematical rigor; the cost is the need to work with an abstraction that has little intuitive appeal.

8.3 Convolutions

The conditioning formula <8.4> can be used to find the distribution for a sum of two independent random variables, each having a continuous distribution.

Example <8.10> Suppose X has a continuous distribution with density f and Y has a continuous distribution with density g. If X and Y are independent then the random variable $Z = X + Y$ has a continuous distribution with density

<8.9> $$h(z) = \int_{-\infty}^{\infty} g(z-x)f(x)\,dx \qquad \text{for all real } z.$$

Equality <8.9> is called the ***convolution formula***. The next Example shows the formula in action. It also serves as an advertisement for indicator functions.

Example <8.11> If X and Y are independent, each with the Uniform$(0,1)$ distribution, find the distribution of $X+Y$.

The convolution formula also establishes an important fact about the normal distribution, which lurks behind the central limit theorem. If you are interested, look at the Appendix to this chapter for a sketch of a beautiful proof of the CLT due to Lindeberg (1922).

Example <8.12> If X_1 and X_2 are independent random variables with $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

8.4 Examples for Chapter 8

<8.10> **Example.** *Suppose X has a continuous distribution with density f and Y has a continuous distribution with density g. If X and Y are independent show that the random variable $Z = X + Y$ has a continuous distribution with density*

$$h(z) = \int_{-\infty}^{\infty} g(z-x)f(x)\,dx \qquad \text{for all real } z.$$

As usual, consider a fixed z and a small, positive δ. Then

$$\begin{aligned}
&\mathbb{P}\{z \le Z \le z+\delta\} \\
&\quad = \int_{-\infty}^{+\infty} \mathbb{P}\{z \le X+Y \le z+\delta \mid X=x\} f(x)\,dx \qquad \text{by <8.4>} \\
&\quad = \int_{-\infty}^{+\infty} \mathbb{P}\{z \le x+Y \le z+\delta \mid X=x\} f(x)\,dx \\
&\quad = \int_{-\infty}^{+\infty} \mathbb{P}\{z-x \le Y \le z-x+\delta \mid X=x\} f(x)\,dx \\
&\quad = \int_{-\infty}^{+\infty} \mathbb{P}\{z-x \le Y \le z-x+\delta\} f(x)\,dx \qquad \text{by independence} \\
&\quad \approx \int_{-\infty}^{+\infty} \delta g(z-x) f(x)\,dx \qquad \text{density for } Y.
\end{aligned}$$

That is,

$$\mathbb{P}\{z \le Z \le z+\delta\} \approx \delta h(x)$$

as asserted. □

<8.11> **Example.** *If X and Y are independent, each with the Uniform$(0,1)$ distribution, find the distribution of $X+Y$.*

The Uniform$(0,1)$ has density function $f(x) = \mathbf{1}\{0 < x < 1\}$, that is,

$$f(x) = \begin{cases} 1 & \text{if } x \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

The density function h for the distribution of $X+Y$ is given by

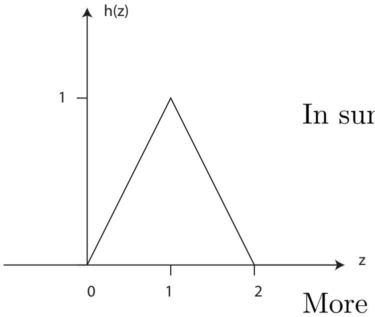
$$\begin{aligned}
h(z) &= \int_{-\infty}^{\infty} \mathbf{1}\{0 < z-x < 1\}\mathbf{1}\{0 < x < 1\}\,dx \\
&= \int_{-\infty}^{\infty} \mathbf{1}\{x < z,\ x > z-1,\ 0 < x < 1\}\,dx \\
&= \int_{-\infty}^{\infty} \mathbf{1}\{\max(0, z-1) < x < \min(1,z)\}\,dx.
\end{aligned}$$

If $z \le 0$ or $z \ge 2$ there are no values of x that satisfy the pair of inequalities in the final indicator function; for those cases the indicator function is zero. If $0 < z \le 1$ the indicator becomes $\mathbf{1}\{0 < x < z\}$, so that the corresponding integral equals

$$\int_{-\infty}^{\infty} \mathbf{1}\{0 < x < z\}\,dx = \int_0^z 1\,dx = z.$$

Similarly, if $1 < z < 2$ the integral becomes

$$\int_{-\infty}^{\infty} \mathbf{1}\{z-1 < x < 1\}\, dx = \int_{z-1}^{1} 1\, dx = 2 - z.$$



In summary,

$$h(z) = \begin{cases} 0 & \text{if } z \le 0 \text{ or } z \ge 2 \\ z & \text{if } 1 < z \le 1 \\ 2-z & \text{if } 1 < z < 2 \end{cases} .$$

More succinctly, $h(z) = \big(1 - |z-1|\big)^{+}$.

□

<8.12> **Example.** *If X_1 and X_2 are independent random variables with $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Let me simplify the algebra by writing $X_i = \mu_i + \sigma_i Z_i$, where Z_1 and Z_2 are independent standard normals. Then we have $X_1 + X_2 = \mu_1 + \mu_2 + \sigma_1 Z_1 + \sigma_2 Z_2$. It will suffice we show that $W = \sigma_1 Z_1 + \sigma_2 Z_2$ has a $N(0, \sigma_1^2 + \sigma_2^2)$ distribution.

The convolution formula gives the density for the distribution of W,

$$h(z) = \frac{1}{\sigma_1 \sigma_2 2\pi} \int_{-\infty}^{\infty} \exp\left(- \frac{(z-x)^2}{2\sigma_1^2} - \frac{x^2}{2\sigma_2^2} \right) dx.$$

The exponent expands to

$$-\tfrac{1}{2} x^2 \left(\sigma_1^{-2} + \sigma_2^{-2} \right) + zx/\sigma_1^2 - \tfrac{1}{2} z^2/\sigma_1^2.$$

Make the change of variable $y = x/c$, with

$$c = 1/\sqrt{\sigma_1^{-2} + \sigma_2^{-2}} = \sigma_1 \sigma_2/\tau \qquad \text{where } \tau = \sqrt{\sigma_1^2 + \sigma_2^2}.$$

The exponent becomes

$$\begin{aligned} &-\tfrac{1}{2}\left(y^2 - 2zcy/\sigma_1^2 + c^2 z^2/\sigma_1^4 \right) + \tfrac{1}{2} c^2 z^2/\sigma_1^4 - \tfrac{1}{2} z^2/\sigma_1^2 \\ &= -\tfrac{1}{2}\left(y - zc/\sigma_1^2 \right)^2 - \tfrac{1}{2} z^2/\tau^2. \end{aligned}$$

The expression for $h(z)$ simplifies to

$$\frac{1}{\tau 2\pi} \exp\left(- \frac{z^2}{2\tau^2} \right) \int_{-\infty}^{\infty} \exp\left(- \tfrac{1}{2} (y - zc/\sigma_1^2)^2 \right) dy.$$

The change of variable $w = y - zc/\sigma_1^2$ then leaves an integral that equals $\sqrt{2\pi}$.

What a mess!

All the sneaky changes of variable might leave you feeling that the argument is difficult. In fact I didn't have to be so careful. In the original convolution integral I had an exponent of the form $-C_1x^2 + C_2xz - C_3z^2$ for some constants C_1, C_2, C_3. I completed the square to rewrite the exponent as $-C_4(y - C_5z)^2 - C_6z^2$, where y is a linear function of x and C_4, C_5, C_6 were new constants. A change of variable allowed me to integrate out the y, leaving an expression of the form $C_7 \exp(-C_6z^2)$, which is clearly a $N(0, \tau^2)$ density for some τ. I could have calculated τ directly by $\tau^2 = \text{var}(W) = \sigma_1^2\text{var}(Z_1) + \sigma_2^2\text{var}(Z_2)$. □

8.5 Appendix: Lindeberg's method for the CLT

This Section should be skipped unless you have a burning desire to know how a general proof of the CLT works.

Suppose $X = X_1 + X_2 + \dots + X_n$, a sum of a independent random variables with $\mathbb{E}X_i = 0$ and $\text{var}(X_i) = \sigma_i^2$. We may assume the random variable X has been scaled so that $\sum_i \sigma_i^2 = 1$.

If all the X_i's are normally distributed, repeated appeals to Example <8.12> show that X is also normally distributed.

If the X_i's are not normal, we replace them one at a time by new independent random variables $Y_1, \dots, Y_n$ with $Y_i \sim N(0, \sigma_i^2)$. It is easy to use Taylor's theorem to track the effect of the replacement if we consider smoooth functions of the sum.

For example, suppose h has a lot of bounded, continuous derivatives. Write S for $X_1 + \dots + X_{n-1}$. Then

$$\begin{aligned}
&\mathbb{E}h(X)\\
&\quad = \mathbb{E}h(S + X_n)\\
&\quad = \mathbb{E}\Big[h(S) + X_nh'(S) + \tfrac{1}{2}X_n^2h''(S) + \tfrac{1}{6}X_n^3h'''(S) + \dots\Big]\\
&\quad = \mathbb{E}h(S) + \mathbb{E}X_n\mathbb{E}h'(S) + \tfrac{1}{2}\mathbb{E}(X_n^2)\mathbb{E}h''(S) + \tfrac{1}{6}\mathbb{E}(X_n^3)\mathbb{E}(h'''(S)) + \dots\\
&\quad = \mathbb{E}h(S) + 0 + \tfrac{1}{2}\sigma_i^2 + \text{ terms of higher order}
\end{aligned}$$

In the last two lines I used the independence to factorize a bunch of products. Exactly the same idea works for $h(S + Y_n)$. That is,

$$\mathbb{E}h(S + Y_n) = \mathbb{E}h(S) + 0 + \tfrac{1}{2}\sigma_i^2 + \text{ terms of higher order}$$

Subtract the two expansions, noting the cancellations.

$$\mathbb{E}h(S+X_n) - \mathbb{E}h(S+Y_n) = \text{ terms of higher order }.$$

A similar argument works if we replace the X_{n-1} in $\mathbb{E}h(S+Y_n)$ by its companion Y_{n-1}. And so on. After we swap out all the X_i's we are left with

$$Eh(X) - \mathbb{E}h(Y_1+Y_2+\dots Y_n) = \text{ a sum of quantities of third, or higher order.}$$

A formal theorem would give a precise meaning to how small the X_i's have to be in order to make the "sum of quantities of third, or higher order" small enough to ignore.

If you were interested in expectations $\mathbb{E}h(X)$ for functions that are not smooth, as happens with $\mathbb{P}\{X \le x\}$, you would need to approximate the non-smooth h by a smooth function for which Lindeberg's method can be applied.

References

Lindeberg, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift 15*, 211–225.

Chapter 9

Poisson approximations

9.1 Overview

The Bin(n, p) can be thought of as the distribution of a sum of independent indicator random variables $X_1 + \cdots + X_n$, with $\{X_i = 1\}$ denoting a head on the ith toss of a coin that lands heads with probability p. Each X_i has a Ber(p) distribution. 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 $\lambda := np$ is not too large, a different type of approximation to the Binomial is better. The traditional explanation uses an approximation to

$$\mathbb{P}\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

for a fixed k. For a k that is small compared with n, consider the contributions $\binom{n}{k}p^k$ and $(1-p)^{n-k}$ separately.

$$\begin{aligned}\binom{n}{k}p^k &= \frac{n(n-1)\dots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k \\ &= 1 \times \left(1 - \frac{1}{n}\right) \times \dots \left(1 - \frac{k-1}{n}\right)\frac{\lambda^k}{k!} \approx \frac{\lambda^k}{k!}\end{aligned}$$

and

$$\log(1-p)^{n-k} = (n-k)\log\left(1 - \lambda/n\right) \approx n\left(-\lambda/n\right).$$

That is, $(1-p)^{n-k} \approx e^{-\lambda}$. Together the two approximations give

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{\lambda^k}{k!} e^{-\lambda}.$$

For large k, both $\mathbb{P}\{X = k\}$ and $p'_k := e^{-\lambda}\lambda^k/k!$ are small. The p'_k define a new distribution.

Definition. A random variable Y is said to have a ***Poisson distribution*** with parameter λ if it can take values in $\mathbb{N}_0$, the set of nonnegative integers, with probabilities

$$\mathbb{P}\{Y = k\} = \frac{e^{-\lambda}\lambda^k}{k!} \qquad \text{for } k = 0, 1, 2, \ldots$$

The parameter λ must be positive. The distribution is denoted by Poisson(λ).

That is , for $\lambda = np$ not too large, the Bin(n, p) is (well?) approximated by the Poisson(λ).

> **Remark.** 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 modeled by Poisson distributions, at least as a first approximation.

The Poisson inherits several properties from the Binomial. For example, the Bin(n, p) has expected value np and variance $np(1 - p)$. One might suspect that the Poisson(λ) should therefore have expected value $\lambda = n(\lambda/n)$ and variance $\lambda = \lim_{n\to\infty} n(\lambda/n)(1 - \lambda/n)$. Also, the coin-tossing origins of the Binomial show that if X has a Bin(m, p) distribution and Y has a Bin(n, p) distribution independent of X, then $X + Y$ has a Bin$(n + m, p)$ distribution. Putting $\lambda = mp$ and $\mu = np$ one might then suspect that the sum of independent Poisson(λ) and Poisson(μ) distributed random variables is Poisson$(\lambda + \mu)$ distributed. These suspicions are correct.

Example <9.1> If X has a Poisson(λ) distribution, then $\mathbb{E}X = \text{var}(X) = \lambda$. If also Y has a Poisson(μ) distribution, and Y is independent of X, then $X + Y$ has a Poisson$(\lambda + \mu)$ distribution.

There is a clever way to simplify some of the calculations in the last Example using ***generating functions***, a way to code all the Poisson probabilities into a single function on $[0, 1]$.

Example <9.2> Calculate moments of the Poisson(λ) distribution using its generating function.

9.2 A more precise Poisson approximation

Modern probability methods have improved this rough approximation of the Binomial by the Poisson by giving useful upper bounds for the error of approximation. Using a technique known as the ***Chen-Stein method*** one can show that

$$d_{TV}(\text{Bin}(n,p), \text{Poisson}(np)) := \frac{1}{2}\sum\nolimits_{k\geq 0}\left|\mathbb{P}\{S=k\}-e^{-\lambda}\frac{\lambda^k}{k!}\right| \leq \min\left(p, np^2\right),$$

which makes precise the traditional advice that the Poisson approximation is good "when p is small and np is not too big". (In fact, the tradition was a bit conservative.)

Remark. The quantity $d_{TV}(P,Q)$ is called the ***total variation distance*** between two probabilities P and Q. It is also equal to $\max_A |PA - QA|$ where the maximum runs over all subsets A of the set where both P and Q are defined. For $P = \text{Bin}(n,p)$ and $Q = \text{Poisson}(np)$, the A runs over all subsets of the nonnegative integers.

The Chen-Stein method of approximation also works in situations where the rare events do not all have the same probability of occurrence. For example, suppose $S = X_1 + X_2 + \cdots + X_n$, a sum of independent random variables where X_i has a Ber(p_i) distribution, for constants $p_1, p_2, \ldots, p_n$ that are not necessarily all the same. The sum S has expected value $\lambda = p_1 + \cdots + p_n$. Using Chen-Stein it can also be shown that that

$$\frac{1}{2}\sum\nolimits_{k\geq 0}\left|\mathbb{P}\{S=k\} - e^{-\lambda}\frac{\lambda^k}{k!}\right| \leq \min\left(1, \frac{1}{\lambda}\right)\sum\nolimits_{i=1}^{n} p_i^2.$$

The Chen-Stein method of proof is elementary—in the sense that it makes use of probabilistic techniques only at the level of Statistics 241—but extremely subtle. See Barbour et al. (1992) for an extensive discussion of the method.

9.3 Poisson approximations under dependence

The Poisson approximation also applies in many settings where the trials are "almost independent", but not quite. Again the Chen-Stein method delivers impressively good bounds on the errors of approximation. For example, the method works well in two cases where the dependence takes an a simple form.

Once again suppose $S = X_1 + X_2 + \cdots + X_n$, where X_i has a $\text{Ber}(p_i)$ distribution, for constants $p_1, p_2, \ldots, p_n$ that are not necessarily all the same. Often X_i is interpreted as the indicator function for success in the ith in some finite set of trials. Define $S_{-i} = S - X_i = \sum_{1\le j\le n} \mathbb{I}\{j \ne i\}X_j$. The random variables $X_1, \ldots, X_n$ are said to be ***positively associated*** if

$$\mathbb{P}\{S_{-i} \ge k \mid X_i = 1\} \ge \mathbb{P}\{S_{-i} \ge k \mid X_i = 0\} \qquad \text{for each } i \text{ and } k = 0, 1, 2, \ldots$$

and ***negatively associated*** if

$$\mathbb{P}\{S_{-i} \ge k \mid X_i = 1\} \le \mathbb{P}\{S_{-i} \ge k \mid X_i = 0\} \qquad \text{for each } i \text{ and } k = 0, 1, 2, \ldots.$$

Intuitively, positive association means that success in the ith trial makes success in the other trials more likely; negative association means that success in the ith trial makes success in the other trials less likely.

With some work it can be shown (Barbour et al., 1992, page 20) that

$$\frac{1}{2}\sum\nolimits_{k\ge 0}\left|\mathbb{P}\{S = k\} - e^{-\lambda}\frac{\lambda^k}{k!}\right| \le \min\left(1, \frac{1}{\lambda}\right) \times \begin{cases} \left(\text{var}(S) - \lambda + 2\sum_{i=1}^n p_i^2\right) & \text{under positive association} \\ (\lambda - \text{var}(S)) & \text{under negative association} \end{cases}.$$

These bounds take advantage of the fact that $\text{var}(S)$ would be exactly equal to λ if S had a $\text{Poisson}(\lambda)$ distribution.

The next Example illustrates both the classical approach and the Chen-Stein approach (via positive association) to deriving a Poisson approximation for a matching problem.

Example <9.3> Poisson approximation for a matching problem: assignment of n letters at random to n envelopes, one per envelope.

9.4 Examples for Chapter 9

<9.1> **Example.** *If X has a* Poisson(λ) *distribution, then* $\mathbb{E}X = \text{var}(X) = \lambda$. *If also Y has a* Poisson(μ) *distribution, and Y is independent of X, then $X+Y$ has a* Poisson($\lambda+\mu$) *distribution.*

Assertion (i) comes from a routine application of the formula for the expectation of a random variable with a discrete distribution.

$$\begin{aligned}\mathbb{E}X = \textstyle\sum_{k=0}^{\infty} k\mathbb{P}\{X=k\} &= \textstyle\sum_{k=1}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} \qquad \text{What happens to } k=0?\\ &= e^{-\lambda}\lambda\sum_{k-1=0}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}\\ &= e^{-\lambda}\lambda e^{\lambda}\\ &= \lambda.\end{aligned}$$

Notice how the k cancelled out one factor from the $k!$ in the denominator.

If I 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?

$$\begin{aligned}\mathbb{E}(X^2-X) &= \sum_{k=0}^{\infty} k(k-1)\mathbb{P}\{X=k\}\\ &= e^{-\lambda}\sum_{k=2}^{\infty} k(k-1)\frac{\lambda^k}{k!} \qquad \text{What happens to } k=0 \text{ and } k=1?\\ &= e^{-\lambda}\lambda^2\sum_{k-2=0}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} = \lambda^2.\end{aligned}$$

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 $0, 1, 2 \ldots$. 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.

$$\begin{aligned}
&\mathbb{P}\{X+Y=k\} = \\
&\quad \mathbb{P}\{X=0, Y=k\} + \mathbb{P}\{X=1, Y=k-1\} + \dots + \mathbb{P}\{X=k, Y=0\} \\
&\quad = \mathbb{P}\{X=0\}\mathbb{P}\{Y=k\} + \dots + \mathbb{P}\{X=k\}\mathbb{P}\{Y=0\} \\
&\quad = \frac{e^{-\lambda}\lambda^0}{0!}\frac{e^{-\mu}\mu^k}{k!} + \dots + \frac{e^{-\lambda}\lambda^k}{k!}\frac{e^{-\mu}\mu^0}{0!} \\
&\quad = \frac{e^{-\lambda-\mu}}{k!}\left(\frac{k!}{0!k!}\lambda^0\mu^k + \frac{k!}{1!(k-1)!}\lambda^1\mu^{k-1} + \dots + \frac{k!}{k!0!}\lambda^k\mu^0\right) \\
&\quad = \frac{e^{-\lambda-\mu}}{k!}(\lambda+\mu)^k.
\end{aligned}$$

The bracketed sum in the second last line is just the binomial expansion of $(\lambda+\mu)^k$. □

> **Remark.** 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 $k=0$.

<9.2> **Example.** There is a sneakier way to calculate $\mathbb{E}X^m$ for $m = 1, 2, \dots$ when X has a Poisson(λ) distribution. Code the whole distribution into a function (the ***probability generating function***) of a dummy variable s:

$$g(s) := \mathbb{E}s^X = \sum\nolimits_{k\ge0} s^k e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum\nolimits_{k\ge0}\frac{(s\lambda)^k}{k!} = e^{-\lambda}e^{\lambda s}.$$

Given g, the individual probabilities $\mathbb{P}\{X=k\}$ could be recovered by expanding the function as a power series in s.

Other facts about the distribution can also be obtained from g. For example,

$$\frac{d}{ds}g(s) = \lim_{h\to0}\mathbb{E}\left(\frac{(s+h)^X - s^X}{h}\right) = \mathbb{E}\frac{\partial}{\partial s}s^X = \mathbb{E}Xs^{X-1}$$

and, by direct calculation, $g'(s) = e^{-\lambda}\lambda e^{\lambda s}$. Put $s=1$ in both expressions to deduce that $\mathbb{E}X = g'(1) = \lambda$.

Similarly, repeated differentiation inside the expectation sign gives

$$g^{(m)}(s) = \frac{\partial^m}{\partial s^m}\mathbb{E}(s^X) = \mathbb{E}\left(X(X-1)\dots(X-m+1)s^{X-m}\right),$$

and direct differentiation of g gives $g^{(m)}(s) = e^{-\lambda}\lambda^m e^{\lambda s}$. Again put $s=1$ to deduce that

$$\lambda^m = g^{(m)}(1) = \mathbb{E}\left(X(X-1)\dots(X-m+1)\right) \qquad \text{for } m = 1, 2, \dots$$

□

<9.3> **Example.** Suppose n letters are placed at random into n envelopes, one letter per envelope. The total number of correct matches, S, can be written as a sum $X_1 + \cdots + X_n$ 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

$$p_i = \mathbb{P}\{X_i = 1\} = 1/n \qquad \text{for each } i$$

and

$$\mathbb{P}\{X_i = 1 \mid X_1 = X_2 = \cdots = X_{i-1} = 1\} = \frac{1}{n-i+1}$$

Remark. If we eliminated 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 is approximated by Poisson(1) if n is large.

We can get some supporting evidence for S having something close to a Poisson(1) distribution under the original assumption (one letter per envelope) by calculating some moments.

$$\mathbb{E}S = \sum\nolimits_{i \le n} \mathbb{E}X_i = n\mathbb{P}\{X_i = 1\} = 1$$

and

$$\begin{aligned} \mathbb{E}S^2 &= \mathbb{E}\left(X_1^2 + \cdots + X_n^2 + 2\sum_{i<j} X_i X_j\right) \\ &= n\mathbb{E}X_1^2 + 2\binom{n}{2}\mathbb{E}X_1X_2 \qquad \text{by symmetry} \\ &= n\mathbb{P}\{X_1 = 1\} + (n^2 - n)\mathbb{P}\{X_1 = 1,\, X_2 = 1\} \\ &= \left(n \times \frac{1}{n}\right) + (n^2 - n) \times \frac{1}{n(n-1)} \\ &= 2. \end{aligned}$$

Thus $\text{var}(S) = \mathbb{E}S^2 - (\mathbb{E}S)^2 = 1$. Compare with Example <9.1>, which gives $\mathbb{E}Y = 1$ and $\text{var}(Y) = 1$ for a Y distributed Poisson(1).

Using the ***method of inclusion and exclusion***, it is possible (Feller, 1968, Chapter 4) to calculate the exact distribution of the number of correct matches,

$$\mathbb{P}\{S = k\} = \frac{1}{k!}\left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} - \dots \pm \frac{1}{(n-k)!}\right) \qquad \text{for } k = 0, 1, \dots, n.$$

For fixed k, as $n \to \infty$ 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 Poisson(1) distribution.

The Chen-Stein method is also effective in this problem. I claim that it is intuitively clear (although a rigorous proof might be tricky) that the X_i's are positively associated:

$$\mathbb{P}\{S_{-i} \geq k \mid X_i = 1\} \geq \mathbb{P}\{S_{-i} \geq k \mid X_i = 0\} \qquad \text{for each } i \text{ and each } k \in \mathbb{N}_0.$$

I feel that if $X_i = 1$, then it is more likely for the other letters to find their matching envelopes than if $X_i = 0$, which makes things harder by filling one of the envelopes with the incorrect letter i. Positive association gives

$$\frac{1}{2}\sum\nolimits_{k\geq 0}\left|\mathbb{P}\{S = k\} - e^{-\lambda}\frac{\lambda^k}{k!}\right| \leq 2\sum\nolimits_{i=1}^{n} p_i^2 + \mathrm{var}(S) - 1 = 2/n.$$

As n gets large, the distribution of S does get close to the Poisson(1) in the strong, total variation sense. □

References

Barbour, A. D., L. Holst, and S. Janson (1992). *Poisson Approximation.* Oxford University Press.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications* (third ed.), Volume 1. New York: Wiley.

Chapter 10

Poisson processes

10.1 Overview

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 $n \to \infty$, we get an idealization called a ***Poisson process***.

> **Remark.** 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. Conversely, I 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.

Definition. A Poisson process with rate λ 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.

It often helps to think of $[0, \infty)$ as time.

Note that, for a very short interval of length δ, the number of points N in the interval has a Poisson($\lambda\delta$) distribution, with

$$\begin{aligned}
\mathbb{P}\{N = 0\} &= e^{-\lambda\delta} = 1 - \lambda\delta + o(\delta) \\
\mathbb{P}\{N = 1\} &= \lambda\delta e^{-\lambda\delta} = \lambda\delta + o(\delta) \\
\mathbb{P}\{N \geq 2\} &= 1 - e^{-\lambda\delta} - \lambda\delta e^{-\lambda\delta} = o(\delta).
\end{aligned}$$

When we pass to the idealized mechanism of points generated in continuous time, several awkward details of discrete-time coin tossing disappear.

Example <10.1> (Gamma distribution from Poisson process) The waiting time W_k to the kth point in a Poisson process with rate λ has a continuous distribution, with density $g_k(w) = \lambda^k w^{k-1} e^{-\lambda w}/(k-1)!$ for $w > 0$, zero otherwise.

It is easier to remember the distribution if we rescale the process, defining $T_k = \lambda W_k$. The new T_k has a continuous distribution with a ***gamma*(k) *density***,

$$f_k(t) = \frac{t^{k-1}e^{-t}}{(k-1)!}\mathbb{I}\{t > 0\}$$

Remark. Notice that $g_k = f_k$ when $\lambda = 1$. That is, T_k is the waiting time to the kth point for a Poisson process with rate 1. Put another way, we can generate a Poisson process with rate λ by taking the points appearing at times $0 < T_1 < T_2 < T_3 < \ldots$ from a Poisson process with rate 1, then rescaling to produce a new process with points at

$$0 < \frac{T_1}{\lambda} < \frac{T_2}{\lambda} < \frac{T_3}{\lambda} < \ldots$$

You could verify this assertion by checking the two defining properties for a Poisson process with rate λ. Doesn't it makes sense that, as λ gets bigger, the points appear more rapidly?

For $k = 1$, Example <10.1> shows that the waiting time, W_1, to the first point has a continuous distribution with density $\lambda e^{-\lambda w}\mathbf{1}\{w > 0\}$, which is called the ***exponential distribution with expected value*** $\boldsymbol{1/\lambda}$. (You should check that $\mathbb{E}W_1 = 1/\lambda$.) The random variable λW_1 has a ***standard exponential distribution***, with density $f_1(t) = e^{-t}\mathbf{1}\{t > 0\}$ and expected value 1.

> **Remark.** I write the exponential distribution symbolically as "exp, mean $1/\lambda$". Do you see why the name $\exp(1/\lambda)$ would be ambiguous? Don't confuse the exponential density (or the exponential distribution that it defines) with the exponential function.

Just as for coin tossing, the independence properties of the Poisson process ensures that the times $W_1, W_2 - W_1, W_3 - W_2, \dots$ are independent, each with the same distribution. You can see why this happens by noting that the future evolution of the process after the occurence of the first point at time W_1 is just a Poisson process that is independent of everything that happened up to time W_1. In particular, the standardized time $T_k = \lambda W_k$, which has a gamma(k) distribution, is a sum of independent random variables $Z_1 = \lambda W_1$, $Z_2 = \lambda(W_2 - W_1)$, …each with a standard exponential distribution.

The gamma density can also be defined for fractional values $\alpha > 0$:

$$f_\alpha(t) = \frac{t^{\alpha-1}e^{-t}}{\Gamma(\alpha)}\mathbf{1}\{t > 0\}$$

is called the ***gamma(α) density***. The scaling constant, $\Gamma(\alpha)$, which ensures that the density integrates to one, is given by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx \qquad \text{for each } \alpha > 0.$$

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 <10.2> Facts about the gamma function: $\Gamma(k) = (k-1)!$ for $k = 1, 2, \dots,$ and $\Gamma(1/2) = \sqrt{\pi}$.

The change of variable used in Example <10.2> to prove $\Gamma(1/2) = \sqrt{\pi}$ is essentially the same piece of mathematics as the calculation to find the density for the distribution of $Y = Z^2/2$ when $Z \sim N(0,1)$. The random variable Y has a gamma(1/2) distribution.

Example <10.3> Moments of the gamma distribution

Poisson processes are often used as the simplest model for stochastic processes that involve arrivals at random times.

Example <10.4> A process with random arrivals

Poisson Processes can also be defined for sets other than the half-line.

Example <10.5> A Poisson Process in two dimensions.

10.2 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 heads	↔	Poisson process with rate λ
Bin(n, p) $X =$ #heads in n tosses	↔	Poisson(λt) $X = \#$ points in $[a, a+t]$
$\mathbb{P}\{X = i\} = \binom{n}{i} p^i q^{n-i}$ for $i = 0, 1, \dots, n$		$\mathbb{P}\{X = i\} = e^{-\lambda t}(\lambda t)^i / i!$ for $i = 0, 1, 2 \dots$
geometric(p) $N_1 = \#$ tosses to first head;	↔	(standard) exponential $T_1/\lambda =$ time to first point;
$\mathbb{P}\{N_1 = 1 + i\} = q^i p$ for $i = 0, 1, 2, \dots$		T_1 has density $f_1(t) = e^{-t}$ for $t > 0$
negative binomial	↔	gamma
See HW10		T_k has density $f_k(t) = t^{k-1}e^{-t}/k!$ for $t > 0$
negative binomial as sum of independent geometrics		gamma(k) as sum of independent exponentials

10.3 Examples for Chapter 10

<10.1> **Example.** Let W_k denote the waiting time to the kth point in a Poisson process on $[0, \infty)$ with rate λ. It has a continuous distribution, whose density g_k we can find by an argument similar to the one used in Chapter 7 to find the distribution of an order statistic for a sample from the Uniform$(0, 1)$.

For a given $w > 0$ and small $\delta > 0$, write M for the number of points landing in the interval $[0, w)$, 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

$$M \sim \text{Poisson}(\lambda w) \qquad \text{AND} \qquad N \sim \text{Poisson}(\lambda\delta).$$

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 $[0, w)$. Thus

$$\mathbb{P}\{w \leq W_k \leq w+\delta\} = \mathbb{P}\{M = k-1,\ N = 1\} + \mathbb{P}\{w \leq W_k \leq w+\delta,\ N \geq 2\}.$$

The second term on the right-hand side is of order $o(\delta)$. Independence of M and N lets us factorize the contribution from $N = 1$ into

$$\begin{aligned}\mathbb{P}\{M = k-1\}\mathbb{P}\{N = 1\} &= \frac{e^{-\lambda w}(\lambda w)^{k-1}}{(k-1)!}\frac{e^{-\lambda\delta}(\lambda\delta)^1}{1!} \\ &= \frac{e^{-\lambda w}\lambda^{k-1}w^{k-1}}{(k-1)!}\left(\lambda\delta + o(\delta)\right),\end{aligned}$$

Thus

$$\mathbb{P}\{w \leq W_k \leq w + \delta\} = \frac{e^{-\lambda w}\lambda^k w^{k-1}}{(k-1)!}\delta + o(\delta),$$

which makes

$$g_k(w) = \frac{e^{-\lambda w}\lambda^k w^{k-1}}{(k-1)!}\mathbf{1}\{w > 0\}$$

the density function for W_k. □

<10.2> **Example.** 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

$$\begin{aligned}\Gamma(1/2) &= \int_0^\infty x^{-1/2}e^{-x}dx \\ &= \int_0^\infty \sqrt{2}e^{-y^2/2}dy \\ &= \frac{\sqrt{2}}{2}\frac{\sqrt{2\pi}}{\sqrt{2\pi}}\int_{-\infty}^\infty e^{-y^2/2}dy \\ &= \sqrt{\pi} \qquad \text{cf. integral of } N(0,1) \text{ density.}\end{aligned}$$

For each $\alpha > 0$, an integration by parts gives

$$\begin{aligned}\Gamma(\alpha+1) &= \int_0^\infty x^\alpha e^{-x} dx \\ &= \left[-x^\alpha e^{-x}\right]_0^\infty + \alpha \int_0^\infty x^{\alpha-1} e^{-x} dx \\ &= \alpha\Gamma(\alpha).\end{aligned}$$

Repeated appeals to the same formula, for $\alpha > 0$ and each positive integer m less than α, give

$$\Gamma(\alpha+m) = (\alpha+m-1)(\alpha+m-2)\dots(\alpha)\Gamma(\alpha).$$

In particular,

$$\Gamma(k) = (k-1)(k-2)(k-3)\dots(2)(1)\Gamma(1) = (k-1)! \qquad \text{for } k = 1, 2, \dots.$$

□

<10.3> **Example.** For parameter value $\alpha > 0$, the gamma(α) 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(α) distribution then, for each positive integer m,

$$\begin{aligned}\mathbb{E}T^m &= \int_0^\infty t^m f_\alpha(t)\, dt \\ &= \int_0^\infty \frac{t^m t^{\alpha-1} e^{-t}}{\Gamma(\alpha)} dt \\ &= \frac{\Gamma(\alpha+m)}{\Gamma(\alpha)} \\ &= (\alpha+m-1)(\alpha+m-2)\dots(\alpha) \qquad \text{by Example <10.2>.}\end{aligned}$$

In particular, $\mathbb{E}T = \alpha$ and

$$\text{var}(T) = \mathbb{E}\left(T^2\right) - (\mathbb{E}T)^2 = (\alpha+1)\alpha - \alpha^2 = \alpha.$$

□

<10.4> **Example.** 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 λ_w for the walk-ins, and rate λ_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 $f(t) = \lambda_w e^{-\lambda_w t}\mathbf{1}\{t>0\}$. We need to calculate $\mathbb{P}\{N=i\}$ for $i = 0,1,2,\dots$. Condition on T:

$$\mathbb{P}\{N=i\} = \int_0^\infty \mathbb{P}\{N=i \mid T=t\} f(t)\,dt.$$

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 Poisson($\lambda_c t$) conditional distribution. Thus

$$\begin{aligned}\mathbb{P}\{N=i\} &= \int_0^\infty \frac{e^{-\lambda_c t}(\lambda_c t)^i}{i!}\lambda_w e^{-\lambda_w t}\,dt \\ &= \lambda_w \frac{\lambda_c^i}{i!}\int_0^\infty \left(\frac{x}{\lambda_c+\lambda_w}\right)^i e^{-x}\frac{dx}{\lambda_c+\lambda_w} \qquad \text{putting } x=(\lambda_c+\lambda_w)t \\ &= \frac{\lambda_w}{\lambda_c+\lambda_w}\left(\frac{\lambda_c}{\lambda_c+\lambda_w}\right)^i \frac{1}{i!}\int_0^\infty x^i e^{-x}dx\end{aligned}$$

The $1/i!$ and the last integral cancel. (Compare with $\Gamma(i+1)$.) Writing p for $\lambda_w/(\lambda_c+\lambda_w)$ we have

$$\mathbb{P}\{N=i\} = p(1-p)^i \qquad \text{for } i=0,1,2,\dots$$

That is, $1+N$ has a 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.

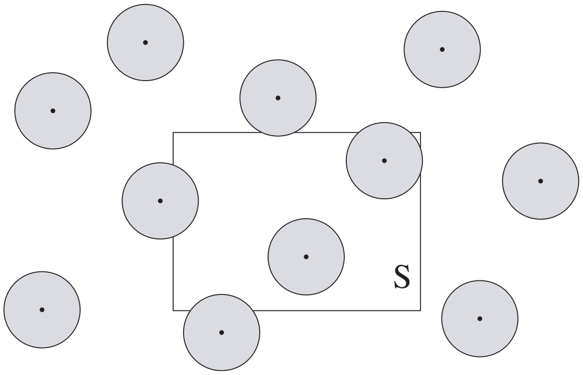
Such a clean result couldn't happen just by accident. HW10 will give you a neater way to explain how the geometric got into the Poisson process. □

<10.5> **Example.** A Poisson process with rate λ on $\mathbb{R}^2$ is a random mechanism that generates "points" in the plane in such a way that

(i) the number of points landing in any region of area A is a random variable with a Poisson(λA) distribution

(ii) the numbers of points landing in disjoint regions are independent random variables.

Suppose mold spores are distributed across the plane as a Poisson process with intensity λ. Around each spore, a circular moldy patch of radius r forms. Let S be some bounded region. Find the expected proportion of the area of S that is covered by mold.



Write $\mathbf{x} = (x, y)$ for the typical point of $\mathbb{R}^2$. If B is a subset of $\mathbb{R}^2$,

$$\text{area of } S \cap B = \iint_{\mathbf{x}\in S} \mathbb{I}\{\mathbf{x} \in B\}\, d\mathbf{x}$$

If B is a random set then

$$\mathbb{E}\big(\text{area of } S \cap B\big) = \iint_{\mathbf{x}\in S} \mathbb{E}\mathbb{I}\{\mathbf{x} \in B\}\, d\mathbf{x} = \iint_{\mathbf{x}\in S} \mathbb{P}\{\mathbf{x} \in B\}\, d\mathbf{x}.$$

If B denotes the moldy region of the plane,

$$\begin{aligned} 1 - \mathbb{P}\{\mathbf{x} \in B\} &= \mathbb{P}\{\text{ no spores land within a distance } r \text{ of } \mathbf{x} \text{ }\} \\ &= \mathbb{P}\{\text{ no spores in circle of radius } r \text{ around } \mathbf{x} \text{ }\} \\ &= \exp\big(-\lambda\pi r^2\big). \end{aligned}$$

Notice that the probability does not depend on $\mathbf{x}$. Consequently,

$$\begin{aligned} \mathbb{E}\big(\text{area of } S \cap B\big) &= \iint_{\mathbf{x}\in S} 1 - \exp\big(-\lambda\pi r^2\big)\, d\mathbf{x} \\ &= \big(1 - \exp\big(-\lambda\pi r^2\big)\big) \times \text{ area of } S \end{aligned}$$

The expected value of the proportion of the area of S that is covered by mold is $1 - \exp\big(-\lambda\pi r^2\big)$. □

Chapter 11

Joint densities

11.1 Overview

Consider the general problem of describing probabilities involving two random variables, X and Y. If both have discrete distributions, with X taking values $x_1, x_2, \ldots$ and Y taking values $y_1, y_2, \ldots$, then everything about the joint behavior of X and Y can be deduced from the set of probabilities

$$\mathbb{P}\{X = x_i, Y = y_j\} \qquad \text{for } i = 1, 2, \ldots \text{ and } j = 1, 2, \ldots$$

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 $\mathbb{P}\{X^2 + Y^2 \le 9\}$ or $\mathbb{P}\{X + Y \le 7\}$?

Definition. Say that random variables X and Y have a jointly continuous distribution with ***joint density*** function $f(\cdot, \cdot)$ if

$$\mathbb{P}\{(X, Y) \in B\} = \iint_B f(x, y)\, dx\, dy.$$

for each subset B of $\mathbb{R}^2$.

Remark. To avoid messy expressions in subscripts, I will sometimes write $\iint \mathbf{1}\{(x, y) \in B\} \ldots$ instead of $\iint_B \ldots$.

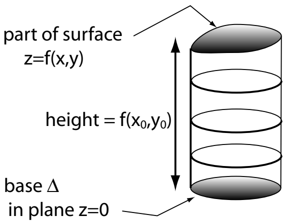
To ensure that $\mathbb{P}\{(X,Y) \in B\}$ is nonnegative and that it equals one when B is the whole of $\mathbb{R}^2$, we must require

$$f \geq 0 \qquad \text{and} \qquad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1.$$

The density function defines a surface, via the equation $z = f(x,y)$. The probability that the random point (X,Y) lands in B is equal to the volume of the "cylinder"

$$\{(x,y,z) \in \mathbb{R}^3 : 0 \leq z \leq f(x,y) \quad \text{and} \quad (x,y) \in B\}.$$

In particular, if Δ is small region in $\mathbb{R}^2$ around a point (x_0, y_0) at which f is continuous, the cylinder is close to a thin column with cross-section Δ and height $f(x_0,y_0)$, so that



$$\mathbb{P}\{(X,Y) \in \Delta\} = (\text{area of } \Delta) f(x_0,y_0) + \text{ smaller order terms.}$$

More formally,

$$\lim_{\Delta \downarrow \{x_0,y_0\}} \frac{\mathbb{P}\{(X,Y) \in \Delta\}}{\text{area of } \Delta} = f(x_0,y_0).$$

The limit is taken as Δ shrinks to the point (x_0,y_0).

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 7.

Example <11.1> Expectations of functions of random variable with jointly continuous distributions: $\mathbb{E}H(X,Y) = \iint_{\mathbb{R}^2} H(x,y)f(x,y)\,dx\,dy$.

The joint density for (X,Y) includes information about the ***marginal distributions*** of the random variables. To see why, write $A \times \mathbb{R}$ for the subset $\{(x,y) \in \mathbb{R}^2 : x \in A, y \in \mathbb{R}\}$ for a subset A of the real line. Then

$$\begin{aligned}
&\mathbb{P}\{X \in A\} \\
&\quad = \mathbb{P}\{(X,Y) \in A \times \mathbb{R}\} \\
&\quad = \iint \mathbf{1}\{x \in A, y \in \mathbb{R}\} f(x,y)\,dx\,dy \\
&\quad = \int_{-\infty}^{+\infty} \mathbf{1}\{x \in A\} \left(\int_{-\infty}^{+\infty} \mathbf{1}\{y \in \mathbb{R}\} f(x,y)\,dy \right) dx \\
&\quad = \int_{-\infty}^{+\infty} \mathbf{1}\{x \in A\} h(x)\,dx \qquad \text{where } h(x) = \int_{-\infty}^{+\infty} f(x,y)\,dy.
\end{aligned}$$

It follows that X has a continuous distribution with ***(marginal) density*** h. Similarly, Y has a continuous distribution with (marginal) density $g(y) = \int_{-\infty}^{+\infty} f(x,y)\,dx$.

> **Remark.** The word *marginal* is used here to distinguish the joint density for (X,Y) from the individual densities g and h.

When we wish to calculate a density, the small region Δ 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 <11.2> (Joint densities for independent random variables) Suppose X has a continuous distribution with density g and Y has a continuous distribution with density h. Then X and Y are independent if and only if they have a jointly continuous distribution with joint density $f(x,y) = g(x)h(y)$ for all $(x,y) \in \mathbb{R}^2$.

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 <11.3> Suppose X and Y have a jointly continuous distribution with density function f. Define $S = X + Y$ and $T = X - Y$. Show that (S,T) has a jointly continuous distribution with density $\psi(s,t) = \frac{1}{2} f\left(\frac{s+t}{2}, \frac{s-t}{2}\right)$.

For instance, suppose the X and Y from Example <11.3> are independent and each is $N(0,1)$ distributed. From Example <11.2>, the joint density for (X,Y) is

$$f(x,y) = \frac{1}{2\pi} \exp\left(-\tfrac{1}{2}(x^2+y^2)\right).$$

The joint density for $S = X + Y$ and $T = X - Y$ is

$$\begin{aligned}\psi(s,t) &= \frac{1}{4\pi} \exp\left(-\tfrac{1}{8}((s+t)^2 + (s-t)^2)\right)\\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{s^2}{2\sigma^2}\right) \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{t^2}{2\sigma^2}\right) \qquad \text{where } \sigma^2 = 2.\end{aligned}$$

It follows that S and T are independent, each with a $N(0,2)$ distribution.

Example <11.3> also implies the convolution formula from Chapter 8. For if X and Y are independent, with densities g and h, then their joint density is $f(x,y) = g(x)h(y)$ and the joint density for $S = X+Y$ and $T = X-Y$ is

$$\psi(s,t) = \tfrac{1}{2} g\left(\frac{s+t}{2}\right) h\left(\frac{s-t}{2}\right)$$

Integrate over t to get the marginal density for S:

$$\begin{aligned}\int_{-\infty}^{+\infty} \psi(s,t)\,dt &= \int_{-\infty}^{+\infty} \tfrac{1}{2} g\left(\frac{s+t}{2}\right) h\left(\frac{s-t}{2}\right)\,dt \\ &= \int_{-\infty}^{+\infty} g(x)h(s-x)\,dx \qquad \text{putting } x = (s+t)/2.\end{aligned}$$

The argument for general linear combinations is slightly more complicated, unless you already know about Jacobians. You could skip the next Example if you don't know about matrices.

Example <11.4> Suppose X and Y have a jointly continuous distribution with joint density $f(x,y)$. For constants a,b,c,d, define $U = aX + bY$ and $V = cX + dY$. Find the joint density function $\psi(u,v)$ for (U,V), under the assumption that the quantity $\kappa = ad - bc$ is nonzero.

The method used in Example <11.4>, for linear transformations, extends to give 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 a good approximation to the density for smooth transformations in small regions.

Example <11.5> Suppose X and Y are independent random variables, with $X \sim \text{gamma}(\alpha)$ and $Y \sim \text{gamma}(\beta)$. Show that the random variables $U = X/(X+Y)$ and $V = X+Y$ are independent, with $U \sim \text{beta}(\alpha,\beta)$ and $V \sim \text{gamma}(\alpha+\beta)$.

The conclusion about $X+Y$ from Example <11.5> 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 <11.6> Sums of independent gamma random variables.

And finally, a polar coordinates way to generate independent normals:

Example <11.7> Building independent normals

11.2 Examples for Chapter 11

<11.1> **Example.** *Expectations of functions of a random variable with jointly continuous distributions*

Suppose X and Y have a jointly continuous distribution with joint density function $f(x,y)$. Let $Y = H(X,Y)$ be a new random variable, defined as a function of X and Y. An approximation argument similar to the one used in Chapter 7 will show that

$$\mathbb{E}H(X,Y) = \iint_{\mathbb{R}^2} H(x,y)f(x,y)\,dx\,dy.$$

For simplicity suppose H is nonnegative. (For the general case split H into positive and negtive parts.) For a small $\delta > 0$ define

$$A_n = \{(x,y) \in \mathbb{R}^2 : n\delta \le H(x,y) < (n+1)\delta\} \qquad \text{for } n = 0,1,\dots$$

The function $H_\delta(x,y) = \sum_{n\ge 0} n\delta\mathbf{1}\{(x,y) \in A_n\}$ approximates H:

$$H_\delta(x,y) \le H(x,y) \le H_\delta(x,y) + \delta \qquad \text{for all } (x,y) \in \mathbb{R}^2.$$

In particular,

$$\mathbb{E}H_\delta(X,Y) \le \mathbb{E}H(X,Y) \le \delta + \mathbb{E}H_\delta(X,Y).$$

and

$$\begin{aligned}&\iint_{\mathbb{R}^2} H_\delta(x,y)f(x,y)\,dx\,dy \\ &\qquad \le \iint_{\mathbb{R}^2} H(x,y)f(x,y)\,dx\,dy \le \delta + \iint_{\mathbb{R}^2} H_\delta(x,y)f(x,y)\,dx\,dy\end{aligned}$$

The random variable $H_\delta(X,Y)$ has a discrete distribution, with expected value

$$\begin{aligned}\mathbb{E}H_\delta(X,Y) &= \mathbb{E}\sum\nolimits_{n\ge 0} n\delta\mathbf{1}\{(X,Y)\in A_n\}\\ &= \sum\nolimits_{n\ge 0} n\delta\,\mathbb{P}\{(X,Y)\in A_n\}\\ &= \sum\nolimits_{n} n\delta \iint_{\mathbb{R}^2}\mathbf{1}\{(x,y)\in A_n\}f(x,y)\,dx\,dy\\ &= \iint_{\mathbb{R}^2}\sum\nolimits_{n} n\delta\mathbf{1}\{(x,y)\in A_n\}f(x,y)\,dx\,dy\\ &= \iint_{\mathbb{R}^2} H_\delta(x,y)f(x,y)\,dx\,dy.\end{aligned}$$

Deduce that

$$\begin{aligned}&\iint_{\mathbb{R}^2} H(x,y)f(x,y)\,dx\,dy-\delta\\ &\qquad\le \mathbb{E}H(X,Y)\\ &\qquad\le \delta+\iint_{\mathbb{R}^2} H(x,y)f(x,y)\,dx\,dy\end{aligned}$$

for every $\delta > 0$. □

<11.2> **Example.** *(Joint densities for independent random variables) Suppose X has a continuous distribution with density g and Y has a continuous distribution with density h. Then X and Y are independent if and only if they have a jointly continuous distribution with joint density $f(x,y)=g(x)h(y)$ for all $(x,y)\in\mathbb{R}^2$.*

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 Δ be a small rectangle with one corner at (x_0,y_0) and small sides of length $\delta>0$ and $\epsilon>0$,

$$\Delta=\{(x,y)\in\mathbb{R}^2 : x_0\le x\le x_0+\delta, y_0\le y\le y_0+\epsilon\}.$$

By independence,

$$\begin{aligned}\mathbb{P}\{(X,Y)\in\Delta\} &= \mathbb{P}\{x_0\le X\le x_0+\delta\}\mathbb{P}\{y_0\le Y\le y_0+\epsilon\}\\ &\approx \delta g(x_0)\epsilon h(y_0) = \big(\text{area of }\Delta\big)\times g(x_0)h(y_0).\end{aligned}$$

Thus X and Y have a jointly continuous distribution with joint density that takes the value $f(x_0,y_0)=g(x_0)h(y_0)$ at (x_0,y_0).

Conversely, if X and Y have a joint density f that factorizes, $f(x,y) = g(x)h(y)$, then for each pair of subsets C, D of the real line,

$$\begin{aligned}\mathbb{P}\{X \in C, Y \in D\} &= \iint \mathbf{1}\{x \in C, y \in D\} f(x,y)\,dx\,dy \\ &= \iint \mathbf{1}\{x \in C\}\mathbf{1}\{y \in D\} g(x)h(y) dx\,dy \\ &= \left(\int \mathbf{1}\{x \in C\} g(x)\,dx\right)\left(\int \mathbf{1}\{y \in D\} h(y)\,dy\right)\end{aligned}$$

Define $K := \int_{-\infty}^{+\infty} g(x)\,dx$. The choice $C = D = \mathbb{R}$ in the previous display then shows that $\int_{-\infty}^{+\infty} h(y)\,dy = 1/K$.

If we take only $D = \mathbb{R}$ we get

$$\mathbb{P}\{X \in C\} = \mathbb{P}\{X \in C, Y \in \mathbb{R}\} = \int_C g(x)/K\,dx$$

from which it follows that $g(x)/K$ is the marginal density for X. Similarly, $Kh(y)$ is the marginal density for Y, so that

$$\mathbb{P}\{X \in C, Y \in D\} = \int_C \frac{g(x)}{K}\,dx \times \int_D Kh(y)\,dy = \mathbb{P}\{X \in C\} \times \mathbb{P}\{Y \in D\}.$$

Put another way,

$$\mathbb{P}\{X \in C \mid Y \in D\} = \mathbb{P}\{X \in C\} \qquad \text{provided } \mathbb{P}\{Y \in D\} \neq 0.$$

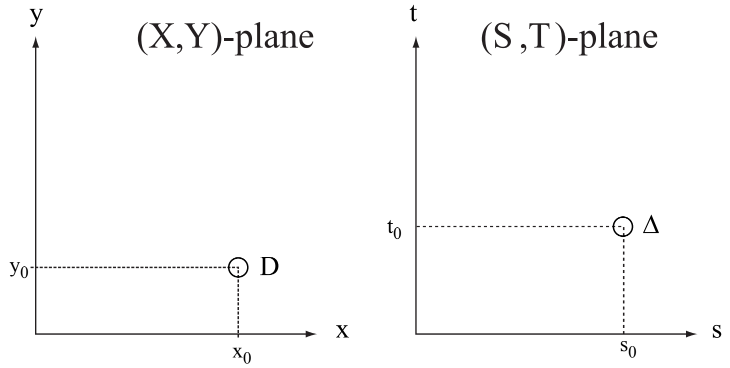
The random variables X and Y are independent.

Of course, if we know that g and h are the marginal densities then we have $K = 1$. The argument in the previous paragraph actually shows that any factorization $f(x,y) = g(x)h(y)$ of a joint density (even if we do not know that the factors are the marginal densities) implies independence. □

<11.3> **Example.** *Suppose X and Y have a jointly continuous distribution with density function f. Define $S = X+Y$ and $T = X-Y$. Show that (S,T) has a jointly continuous distribution with density $g(s,t) = \frac{1}{2}f\left(\frac{s+t}{2}, \frac{s-t}{2}\right)$.*

Consider a small ball Δ of radius ϵ centered at a point (s_0, t_0) in the plane. The area of Δ equals $\pi\epsilon^2$. The point (s_0, t_0) in the (S,T)-plane (the region where (S,T) takes its values) corresponds to the point (x_0, y_0) in the (X,Y)-plane for which $s_0 = x_0 + y + 0$ and $t_0 = x_0 - y_0$. That is, $x_0 = (s_0 + t_0)/2$ and $y_0 = (s_0 - t_0)/2$.

We need to identify $\{(S,T) \in \Delta\}$ with some set $\{(X,Y) \in D\}$.



By great luck (or by a clever choice for Δ) the region D in the (X,Y)-plane turns out to be another ball:

$$\begin{aligned}\{(S,T) \in \Delta\} &= \{(S-s_0)^2 + (T-t_0)^2 \le \epsilon^2\}\\ &= \{(X+Y-x_0-y_0)^2 + (X-Y-x_0+y_0)^2 \le \epsilon^2\}\\ &= \{2(X-x_0)^2 + 2(Y-y_0)^2 \le \epsilon^2\}.\end{aligned}$$

(Notice the cancellation of $(X-x_0)(Y-y_0)$ terms.) That is D is a ball of radius $\epsilon/\sqrt{2}$ centered at (x_0,y_0), with area $\pi\epsilon^2/2$, which is half the area of Δ. Now we can calculate.

$$\begin{aligned}\mathbb{P}\{(S,T) \in \Delta\} &= \mathbb{P}\{(X,Y) \in D\}\\ &\approx (\text{area of } D) \times f(x_0,y_0)\\ &= \tfrac{1}{2}(\text{area of } \Delta) \times f\left(\frac{s_0+t_0}{2}, \frac{s_0-t_0}{2}\right)\end{aligned}$$

It follows that (S,T) has joint density $g(s,t) = \frac{1}{2} f\left(\frac{s+t}{2}, \frac{s-t}{2}\right)$. □

<11.4> **Example.** *Suppose X and Y have a jointly continuous distribution with joint density $f(x,y)$. For constants a,b,c,d, define $U = aX + bY$ and $V = cX + dY$. Find the joint density function $\psi(u,v)$ for (U,V), under the assumption that the quantity $\kappa = ad - bc$ is nonzero.*

In matrix notation,

$$(U,\, V) = (X,\, Y)A \qquad \text{where } A = \begin{pmatrix} a & c \\ b & d \end{pmatrix}.$$

Notice that $\det A = ad - bc = \kappa$. The assumption that $\kappa \neq 0$ ensures that A has an inverse:

$$A^{-1} = \frac{1}{\kappa}\begin{pmatrix} d & -c \\ -b & a \end{pmatrix}$$

That is, if $(u, v) = (x, y)A$ then

$$\frac{du - bv}{\kappa} = x \qquad \text{and} \qquad \frac{-cu + av}{\kappa} = y.$$

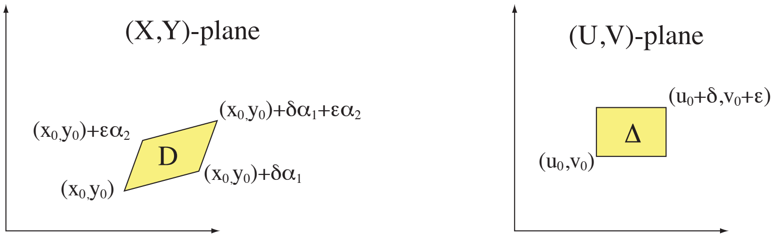
Notice that $\det\left(A^{-1}\right) = 1/\kappa = 1/(\det A)$.

Consider a small rectangle $\Delta = \{u_0 \leq u \leq u_0 + \delta,\ v_0 \leq v \leq v_0 + \epsilon\}$, for (u_0, v_0) in the (U, V)-plane and small, positive δ and ϵ. The joint density function $\psi(u, v)$ is characterized by the property that

$$\mathbb{P}\{(U, V) \in \Delta\} \approx \psi(u_0, v_0)\delta\epsilon$$

The event $\{(U, V) \in \Delta\}$ is equal to some event $\{(X, Y) \in D\}$. 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 D must be a parallelogram. We have only to determine its vertices, which correspond to the four vertices of the rectangle Δ. Define vectors $\alpha_1 = (d, -c)/\kappa$ and $\alpha_2 = (-b, a)/\kappa$, which correspond to the two rows of the matrix A^{-1}. Then D has vertices:

$$\begin{aligned}
(x_0, y_0) &= (u_0, v_0)A^{-1} = u_0\alpha_1 + v_0\alpha_2 \\
(x_0, y_0) + \delta\alpha_1 &= (u_0 + \delta, v_0)A^{-1} = (u_0 + \delta)\alpha_1 + v_0\alpha_2 \\
(x_0, y_0) + \epsilon\alpha_2 &= (u_0, v_0 + \epsilon)A^{-1} = u_0\alpha_1 + (v_0 + \epsilon)\alpha_2 \\
(x_0, y_0) + \delta\alpha_1 + \epsilon\alpha_2 &= (u_0 + \delta, v_0 + \epsilon)A^{-1} = (u_0 + \delta)\alpha_1 + (v_0 + \epsilon)\alpha_2
\end{aligned}$$



From the formula in the Appendix to this Chapter, the parallelogram D has area equal to $\delta\epsilon$ times the absolute value of the determinant of the matrix with rows α_1 and α_2. That is,

$$\text{area of } D = \delta\epsilon|\det(A^{-1})| = \frac{\delta\epsilon}{|\det A|}.$$

In summary: for small $\delta > 0$ and $\epsilon > 0$,

$$\begin{aligned}\psi(u_0, v_0)\delta\epsilon &\approx \mathbb{P}\{(U,V) \in \Delta\} \\ &= \mathbb{P}\{(X,Y) \in D\} \\ &\approx (\text{area of } D)f(x_0, y_0) \\ &\approx \delta\epsilon f(x_0, y_0)/|\det(A)|.\end{aligned}$$

It follows that (U,V) have joint density

$$\psi(u,v) = \frac{1}{|\det A|}f(x,y) \qquad \text{where } (x,y) = (u,v)A^{-1}.$$

On the right-hand side you should substitute $(du - bv)/\kappa$ for x and $(-cu + av)/\kappa$ for y, in order to get an expresion involving only u and v. □

Remark. In effect, I have calculated a Jacobian by first principles.

<11.5> **Example.** *Suppose X and Y are independent random variables, with $X \sim$ gamma(α) and $Y \sim$ gamma(β). Show that the random variables $U = X/(X+Y)$ and $V = X+Y$ are independent, with $U \sim$ beta(α,β) and $V \sim$ gamma$(\alpha+\beta)$.*

The random variables X and Y have marginal densities

$$g(x) = x^{\alpha-1}e^{-x}\mathbf{1}\{x > 0\}/\Gamma(\alpha) \qquad \text{and} \qquad h(y) = y^{\beta-1}e^{-y}\mathbf{1}\{y > 0\}/\Gamma(\beta)$$

From Example <11.2>, they have a jointly continuous distribution with joint density

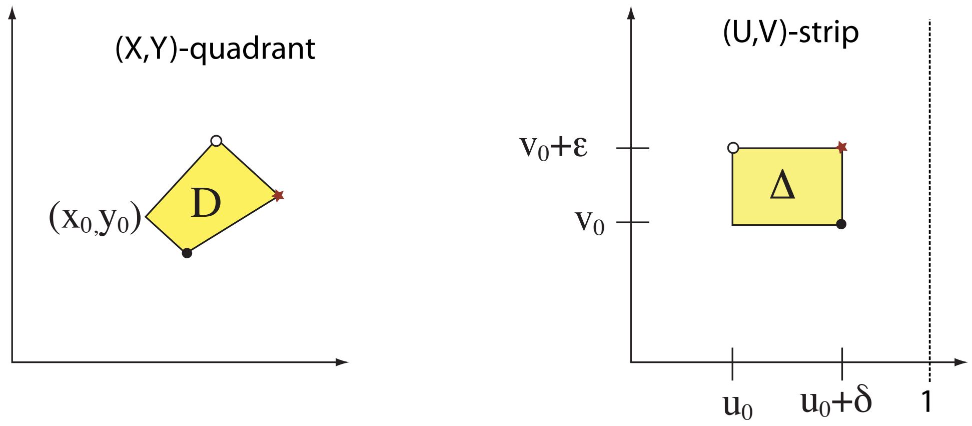
$$f(x,y) = g(x)h(y) = \frac{x^{\alpha-1}e^{-x}y^{\beta-1}e^{-y}}{\Gamma(\alpha)\Gamma(\beta)}\mathbf{1}\{x > 0, y > 0\}.$$

We need to find the joint density function $\psi(u,v)$ for the random variables $U = X/(X+Y)$ and $V = X+Y$. The pair (U,V) takes values in the strip defined by $\{(u,v) \in \mathbb{R}^2 : 0 < u < 1, 0 < v < \infty\}$. The joint density function ψ 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 for which

$$u_0 = x_0/(x_0 + y_0) \qquad \text{AND} \qquad v_0 = x_0 + y_0,$$

that is,

$$x_0 = u_0 v_0 \qquad \text{AND} \qquad y_0 = (1-u_0)v_0.$$



When (U,V) lies near (u_0,v_0) then (X,Y) lies near the point $(x_0,y_0) = (u_0v_0, v_0(1-u_0))$. More precisely, for small positive δ and ϵ, there is a small region D in the (X,Y)-quadrant corresponding to the small rectangle

$$\Delta = \{(u,v) : u_0 \le u \le u_0+\delta, v_0 \le v \le v_0+\epsilon\}$$

in the (U,V)-strip. That is, $\{(U,V) \in \Delta\} = \{(X,Y) \in D\}$. The set D is not a parallelogram but it is well approximated by one. For small perturbations, the map from (u,v) to (x,y) is approximately linear. First locate the points corresponding to the corners of Δ, under the maps $x = uv$ and $y = v(1-u)$:

$$\begin{aligned}(u_0,v_0) &\mapsto (x_0,y_0)\\ (u_0+\delta,v_0) &\mapsto (x_0,y_0)+\delta(v_0,-v_0)\\ (u_0,v_0+\epsilon) &\mapsto (x_0,y_0)+\epsilon(u_0,1-u_0).\end{aligned}$$

The fourth vertex, $(u_0+\delta,v_0+\epsilon)$ corresponds to the point (x,y) with

$$\begin{aligned}x &= (u_0+\delta)(v_0+\epsilon) = u_0v_0+\delta v_0+\epsilon u_0+\delta\epsilon\\ y &= (v_0+\epsilon)(1-u_0-\delta) = v_0u_0+\epsilon(1-u_0)-\delta v_0-\delta\epsilon\end{aligned}$$

Put another way,

$$\begin{aligned}(u_0, v_0) &\mapsto (x_0, y_0)\\ (u_0, v_0) + (\delta, 0) &\mapsto (x_0, y_0) + (\delta, 0)J\\ (u_0, v_0) + (0, \epsilon) &\mapsto (x_0, y_0) + (0, \epsilon)J\\ (u_0, v_0) + (\delta, \epsilon) &\mapsto (x_0, y_0) + (\delta, \epsilon)J + \text{ smaller order terms}\end{aligned}$$

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

$$\begin{pmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial y}{\partial u} \\[2ex] \dfrac{\partial x}{\partial v} & \dfrac{\partial y}{\partial v} \end{pmatrix}$$

evaluated at (u_0, v_0).

The region D is approximately a parallelogram, with the edges oblique to the coordinate axes. To a good approximation, the area of D is equal to $\delta\epsilon$ times the area of the parallelogram with corners at

$$(0,0), \qquad \mathbf{a} = (v_0, -v_0), \qquad \mathbf{b} = (u_0, 1 - u_0), \qquad \mathbf{a} + \mathbf{b},$$

which, from the Appendix to this Chapter, equals $|\det(J)| = v_0$.

The rest of the calculation of the joint density ψ for (U, V) is easy:

$$\begin{aligned}\delta\epsilon\psi(u_0, v_0) &\approx \mathbb{P}\{(U, V) \in \Delta\}\\ &= \mathbb{P}\{(X, Y) \in R\}\\ &\approx f(x_0, y_0)(\text{area of } D) \approx \frac{x_0^{\alpha-1} e^{-x_0}}{\Gamma(\alpha)} \frac{y_0^{\beta-1} e^{-y_0}}{\Gamma(\beta)} \,\delta\,\epsilon\, v_0\end{aligned}$$

Substitute $x_0 = u_0 v_0$ and $y_0 = (1 - u_0)v_0$ to get the joint density at (u_0, v_0):

$$\begin{aligned}\psi(u_0, v_0) &= \frac{u_0^{\alpha-1} v_0^{\alpha-1} e^{-u_0 v_0}}{\Gamma(\alpha)} \frac{(1 - u_0)^{\beta-1} v_0^{\beta-1} e^{-v_0 + u_0 v_0}}{\Gamma(\beta)} v_0\\ &= \frac{u_0^{\alpha-1}(1 - u_0)^{\beta-1}}{B(\alpha, \beta)} \times \frac{v_0^{\alpha+\beta-1} e^{-v_0}}{\Gamma(\alpha + \beta)} \times \frac{\Gamma(\alpha + \beta)B(\alpha, \beta)}{\Gamma(\alpha)\Gamma(\beta)}.\end{aligned}$$

Once again the final constant must be equal to 1, which gives the identity

$$B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The joint density factorizes; the random variables U and V are independent, with $U \sim \text{Beta}(\alpha,\beta)$ and $V \sim \text{gamma}(\alpha+\beta)$. □

Remark. The fact that $\Gamma(1/2) = \sqrt{\pi}$ also follows from the equality

$$\begin{aligned}\frac{\Gamma(1/2)\Gamma(1/2)}{\Gamma(1)} &= B(1/2,1/2)\\ &= \int_0^1 t^{-1/2}(1-t)^{-1/2}\,dt \qquad \text{put } t = \sin^2(\theta)\\ &= \int_0^{\pi/2} \frac{1}{\sin(\theta)\cos(\theta)} 2\sin(\theta)\cos(\theta)\,d\theta = \pi.\end{aligned}$$

<11.6> **Example.** If $X_1, X_2, \dots, X_k$ are independent random variables, with X_i distributed gamma(α_i) for $i = 1,\dots,k$, then

$$\begin{aligned}&X_1 + X_2 \sim \text{gamma}(\alpha_1+\alpha_2),\\ &X_1+X_2+X_3 = (X_1+X_2)+X_3 \sim \text{gamma}(\alpha_1+\alpha_2+\alpha_3)\\ &X_1+X_2+X_3+X_4 = (X_1+X_2+X_3)+X_4 \sim \text{gamma}(\alpha_1+\alpha_2+\alpha_3+\alpha_4)\\ &\dots\\ &X_1+X_2+\dots+X_k \sim \text{gamma}(\alpha_1+\alpha_2+\dots+\alpha_k)\end{aligned}$$

A particular case has great significance for Statistics. Suppose $Z_1, \dots Z_k$ are independent random variables, each distributed N(0,1). You know that the random variables $Z_1^2/2, \dots, Z_k^2/2$ are independent gamma(1/2) distributed random variables. The sum

$$(Z_1^2 + \dots + Z_k^2)/2$$

must have a gamma($k/2$) distribution with density $t^{k/2-1}e^{-t}\mathbf{1}\{0<t\}/\Gamma(k/2)$. It follows that the sum $Z_1^2 + \dots + Z_k^2$ has density

$$\frac{(t/2)^{k/2-1}e^{-t/2}\mathbf{1}\{0<t\}}{2\Gamma(k/2)}.$$

This distribution is called the ***chi-squared*** on k degrees of freedom, usually denoted by χ_k^2. The letter χ is a lowercase Greek chi. □

<11.7> **Example.** Here are the bare bones of the polar coordinates way of manufacturing two independent $N(0,1)$'s. Start with independent random variables $U \sim \text{Uniform}(0, 2\pi)$ and $W \sim \text{gamma}(1)$ (a.k.a. standard exponential). Define $R = \sqrt{2W}$ and $X = R\cos(U)$ and $Y = R\sin(U)$. Calculate the density for R as

$$g(r) = r\exp(-r^2/2)\mathbf{1}\{r > 0\}.$$

For $0 < \theta_0 < 1$ and $r_0 > 0$, and very small $\delta > 0$ and $\epsilon > 0$, check that the region

$$D = \{(u, r) \in (0,1) \times (0, \infty) : \theta_0 \le U \le \theta_0 + \delta,\ r_0 \le r \le r_0 + \epsilon\}$$

corresponds to the region Δ in the (X, Y)-plane that is bounded by circles of radius r_0 and $r_0 + \epsilon$ and by radial lines from the origin at angles θ_0 and $\theta_0 + \delta$ to the horizontal axis. The area of Δ is approximately $2\pi r_0 \epsilon\delta$.

Deduce that the joint density f for (X, Y) satisfies

$$2\pi r_0 \epsilon \delta f(x_0, y_0) \approx \epsilon g(r_0)\frac{\delta}{2\pi} \qquad \text{where } x_0 = r_0\cos(\theta_0), \quad y_0 = r_0 \sin(\theta_0)$$

That is,

$$\begin{aligned} f(x,y) &= \frac{g(r)}{2\pi r} \qquad \text{where } x = r\cos(\theta), \quad y = r\sin(\theta) \\ &= \frac{1}{2\pi}\exp\left(-\frac{1}{2}(x^2 + y^2)\right). \end{aligned}$$

The random variables X and Y are independent, with each distributed $N(0,1)$. □

11.3 Appendix: area of a parallelogram

Let R be a parallelogram in the plane with corners at $\mathbf{0} = (0,0)$, and $\mathbf{a} = (a_1, a_2)$, and $\mathbf{b} = (b_1, b_2)$, and $\mathbf{a} + \mathbf{b}$. The area of R is equal to the absolute value of the determinant of the matrix

$$J = \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix} = \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix}.$$

That is, the area of R equals $|a_1 b_2 - a_2 b_1|$.

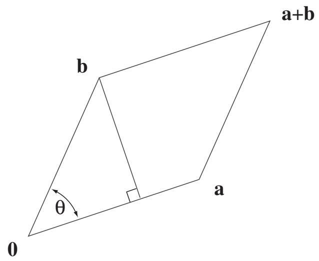
Proof Let θ denotes the angle between $\mathbf{a}$ and $\mathbf{b}$. Remember that

$$\|\mathbf{a}\| \times \|\mathbf{b}\| \times \cos(\theta) = \mathbf{a} \cdot \mathbf{b}$$

The area of R is twice the area of the triangle with vertices at $\mathbf{0}$, $\mathbf{a}$, and $\mathbf{b}$. The triangle has area

$$\tfrac{1}{2}(\text{base length}) \ \times (\text{height}) = \tfrac{1}{2}\,\|\mathbf{a}\| \times (\|\mathbf{b}\| \times |\sin\theta|)$$

The square of the area of R equals



$$\begin{aligned}\|\mathbf{a}\|^2\,\|\mathbf{b}\|^2\sin^2(\theta) &= \|\mathbf{a}\|^2\,\|\mathbf{b}\|^2 - \|\mathbf{a}\|^2\,\|\mathbf{b}\|^2\cos^2(\theta)\\ &= (\mathbf{a}\cdot\mathbf{a})(\mathbf{b}\cdot\mathbf{b}) - (\mathbf{a}\cdot\mathbf{b})^2\\ &= \det\begin{pmatrix}\mathbf{a}\cdot\mathbf{a} & \mathbf{a}\cdot\mathbf{b}\\ \mathbf{a}\cdot\mathbf{b} & \mathbf{b}\cdot\mathbf{b}\end{pmatrix}\\ &= \det\left(JJ'\right)\\ &= (\det J)^2\,.\end{aligned}$$

If you are not sure about the properties of determinants used in the last two lines, you should check directly that

$$(a_1^2 + a_2^2)(b_1^2 + b_2^2) - (a_1b_1 + a_2b_2)^2 = (a_1b_2 - a_2b_1)^2.$$

□

Chapter 12

Conditional densities

12.1 Overview

Density functions determine continuous distributions. If a continuous distribution is calculated conditionally on some information, then the density is called a ***conditional density***. When the conditioning information involves another random variable with a continuous distribution, the conditional density can be calculated from the joint density for the two random variables.

Suppose X and Y have a jointly continuous distribution with joint density $f(x, y)$. From Chapter 11, you know that the marginal distribution of X is continuous with density

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

The conditional distribution for Y given $X = x$ has a (conditional) density, which I will denote by $h(y \mid X = x)$, or just $h(y \mid x)$ if the conditioning variable is unambiguous, for which

$$\mathbb{P}\{y \le Y \le y + \delta \mid X = x\} \approx \delta h(y \mid X = x), \qquad \text{for small } \delta > 0.$$

Conditioning on $X = x$ should be almost the same as conditioning on the event $\{x \le X \le x + \epsilon\}$ for a very small $\epsilon > 0$. That is, provided $g(x) > 0$,

$$\begin{aligned}\mathbb{P}\{y \le Y \le y + \delta \mid X = x\} &\approx \mathbb{P}\{y \le Y \le y + \delta \mid x \le X \le x + \epsilon\} \\ &= \frac{\mathbb{P}\{y \le Y \le y + \delta, x \le X \le x + \epsilon\}}{\mathbb{P}\{x \le X \le x + \epsilon\}} \\ &\approx \frac{\delta\epsilon f(x, y)}{\epsilon g(x)}.\end{aligned}$$

In the limit, as ϵ tends to zero, we are left with $\delta \approx \delta f(x,y)/g(x)$. That is,

$$h(y \mid X = x) = f(x,y)/g(x) \qquad \text{for each } x \text{ with } g(x) > 0.$$

Less formally, the conditional density is

$$\boxed{\quad h(y \mid X = x) = \frac{\text{joint } (X,Y) \text{ density at } (x,y)}{\text{marginal } X \text{ density at } x} \quad}$$

The first Example illustrates two ways to find a conditional density: first by calculation of a joint density followed by an appeal to the formula for the conditional density; and then by a sneakier method where all the random variables are built directly using polar coordinates.

Example <12.1> 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

$$h(x \mid R = r) = \frac{\mathbf{1}\{|x| < r\}}{\pi\sqrt{r^2 - x^2}} \qquad \text{for } r > 0.$$

The most famous example of a continuous condition distribution comes from pairs of random variables that have a bivariate normal distribution. For each constant $\rho \in (-1,+1)$, the ***standard bivariate normal with correlation ρ*** is defined as the joint distribution of a pair of random variables constructed from independent random variables X and Y, each distributed $N(0,1)$. Define $Z = \rho X + \sqrt{1-\rho^2}\,Y$. The pair X,Y has a jointly continuous distribution with density $f(x,y) = (2\pi)^{-1}\exp\big(-(x^2+y^2)/2\big)$. Apply the result from Example <11.4> with

$$(X,Z) = (X,Y)A \qquad \text{where } A = \begin{pmatrix} 1 & \rho \\ 0 & \sqrt{1-\rho^2} \end{pmatrix}$$

to deduce that X,Z have joint density

$$f_\rho(x,z) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 - 2\rho xz + z^2}{2(1-\rho^2)}\right).$$

Notice the symmetry in x and z. The X and Z marginals must be the same. Thus $Z \sim N(0,1)$. Also

$$\begin{aligned} \text{cov}(X,Z) &= \text{cov}(X, \rho X + \sqrt{1-\rho^2}\,Y) \\ &= \rho\,\text{cov}(X,X) + \sqrt{1-\rho^2}\,\text{cov}(X,Y) = \rho. \end{aligned}$$

Remark. The ***correlation*** between two random variables S and T is defined as

$$\text{corr}(S,T) = \frac{\text{cov}(S,T)}{\sqrt{\text{var}(S)\text{var}(T)}}.$$

If $\text{var}(S) = \text{var}(T) = 1$ the correlation reduces to the covariance.

By construction, the conditional distribution of Z given $X = x$ is just the conditional distribution of $\rho x + \sqrt{1-\rho^2}Y$ given $X = x$. Independence of X and Y then shows that

$$Z \mid X = x \sim N(\rho x, 1-\rho^2).$$

In particular, $\mathbb{E}(Z \mid X = x) = \rho x$. By symmetry of f_ρ, we also have $X \mid Z = z \sim N(\rho z, 1-\rho^2)$, a fact that you could check by dividing $f_\rho(x,z)$ by the standard normal density for Z.

Example <12.2> 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 the standardized variables $(S-\mu)/\sigma$ and $(T-\mu)/\sigma$ have a standard bivariate normal distribution with correlation $\rho = .3$.

If Sam has a height of $S = 74$ inches, what would one predict about the ultimate height T of his young son Tom?

For the standard bivariate normal, if the variables are uncorrelated (that is, if $\rho = 0$) then the joint density factorizes into the product of two $N(0,1)$ densities, which implies that the variables are independent. This situation is one of the few where a zero covariance (zero correlation) implies independence.

The final Example demonstrates yet another connection between Poisson processes and order statistics from a uniform distribution. The arguments make use of the obvious generalizations of joint densities and conditional densities to more than two dimensions.

Definition. Say that random variables X, Y, Z have a jointly continuous distribution with joint density $f(x,y,z)$ if

$$\mathbb{P}\{(X,Y,Z) \in A\} = \iiint_A f(x,y,z)\,dx\,dy\,dz \qquad \text{for each } A \subseteq \mathbb{R}^3.$$

As in one and two dimensions, joint densities are typically calculated by looking at small regions: for a small region Δ around (x_0, y_0, z_0)

$$\mathbb{P}\{(X,Y,Z) \in \Delta\} \approx (\text{volume of } \Delta) \times f(x_0, y_0, z_0).$$

Similarly, the joint density for (X, Y) conditional on $Z = z$ is defined as the function $h(x, y \mid Z = z)$ for which

$$\mathbb{P}\{(X,Y) \in B \mid Z = z\} = \iiint \mathbb{I}\{(x,y) \in B\} h(x, y \mid Z = z)\, dx\, dy$$

for each subset B of $\mathbb{R}^2$. It can be calculated, at z values where the marginal density for Z,

$$g(z) = \iint_{\mathbb{R}^2} f(x, y, z)\, dx\, dy,$$

is strictly positive, by yet another small-region calculation. If Δ is a small subset containing (x_0, y_0) then, for small $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}\{(X,Y) \in \Delta \mid Z = z_0\} &\approx \mathbb{P}\{(X,Y) \in \Delta \mid z_0 \le Z \le z_0 + \epsilon\} \\ &= \frac{\mathbb{P}\{(X,Y) \in \Delta,\ z_0 \le Z \le z_0 + \epsilon\}}{\mathbb{P}\{z_0 \le Z \le z_0 + \epsilon\}} \\ &\approx \frac{((\text{area of } \Delta)\ \times \epsilon)\, f(x_0, y_0, z_0)}{\epsilon g(z_0)} \\ &= (\text{area of } \Delta)\ \frac{f(x_0, y_0, z_0)}{g(z_0)}.\end{aligned}$$

> **Remark.** Notice the identification of the set of points (x, y, z) in $\mathbb{R}^3$ for which $(x, y) \in \Delta$ and $z_0 \le z \le z_0 + \epsilon$ as a small region with volume equal to (area of Δ) $\times\ \epsilon$.

That is, the conditional (joint) distribution of (X, Y) given $Z = z$ has density

$$h(x, y \mid Z = z) = \frac{f(x, y, z)}{g(z)} \qquad \text{provided } g(z) > 0.$$

> **Remark.** Many authors (including me) like to abbreviate $h(x, y \mid Z = z)$ to $h(x, y \mid z)$. Many others run out of symbols and write $f(x, y \mid z)$ for the conditional (joint) density of (X, Y) given $Z = z$. This notation is defensible if one can somehow tell which values are being conditioned on. In a problem with lots of conditioning it can get confusing to remember which f is the joint density and which is conditional on something. To avoid confusion, some authors write things like $f_{X,Y|Z}(x, y \mid z)$ for the conditional density and $f_X(x)$ for the X-marginal density, at the cost of more cumbersome notation.

Example <12.3> Let T_i denote the time to the ith point in a Poisson process with rate λ on $[0, \infty)$. Find the joint distribution of (T_1, T_2) conditional on T_3.

From the result in the previous Example, you should be able to deduce that, conditional on $T_3 = t_3$ for a given $t_3 > 0$, the random variables $(T_1/T_3, T_2/T_3)$ are uniformly distributed over the triangular region

$$\{(u_1 u_2) \in \mathbb{R}^2 : 0 < u_1 < u_2 < 1\}.$$

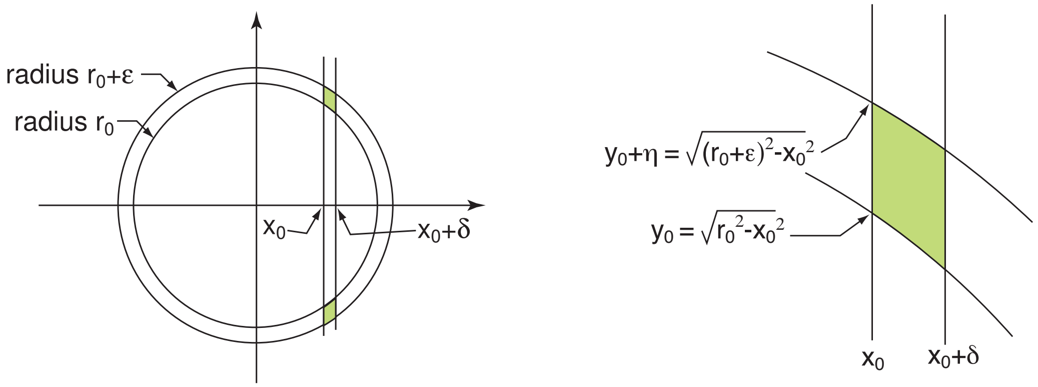
HW11 will step you through an analogous result for order statistics.

12.2 Examples for Chapter 12

<12.1> **Example.** *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 $f(x, y) = (2\pi)^{-1} \exp\left(-(x^2 + y^2)/2\right)$. To find the conditional density for X given $R = r$, first I'll find the joint density ψ for X and R, then I'll calculate its X marginal, and then I'll divide to get the conditional density. A simpler method is described at the end of the Example.

We need to calculate $\mathbb{P}\{x_0 \le X \le x_0 + \delta, r_0 \le R \le r_0 + \epsilon\}$ for small, positive δ and ϵ. 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 $r_0 + \epsilon$.


radius $r_0+\varepsilon$
radius r_0
x_0
$x_0+\delta$
$y_0+\eta = \sqrt{(r_0+\varepsilon)^2-x_0^2}$
$y_0 = \sqrt{r_0^2-x_0^2}$
x_0
$x_0+\delta$

By symmetry, both regions contribute the same probability. Consider the upper region. For small δ and ϵ, 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 δ. We could expand the expression for η as a power series in ϵ by multiple applications of Taylor's theorem. It is easier to argue less directly, starting from the equalities

$$x_0^2 + (y_0+\eta)^2 = (r_0+\epsilon)^2 \qquad \text{AND} \qquad x_0^2 + y_0^2 = r_0^2.$$

Take differences to deduce that $2y_0\eta + \eta^2 = 2r_0\epsilon + \epsilon^2$. Ignore the lower order terms η^2 and ϵ^2 to conclude that $\eta \approx (r_0\epsilon/y_0)$. The upper region has approximate area $r_0\epsilon\delta/y_0$, which implies

$$\begin{aligned}
&\mathbb{P}\{x_0 \le X \le x_0+\delta,\, r_0 \le R \le r_0+\epsilon\}\\
&\quad\approx 2\frac{r_0\epsilon\delta}{y_0} f(x_0,y_0)\\
&\quad\approx \frac{2r_0}{\sqrt{r_0^2 - x_0^2}} \frac{\exp(-r_0^2/2)}{2\pi}\epsilon\delta.
\end{aligned}$$

Thus the random variables X and R have joint density

$$\psi(x,r) = \frac{r\exp(-r^2/2)}{\pi\sqrt{r^2-x^2}}\mathbf{1}\{|x|<r,\, 0<r\}.$$

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.

For $r > 0$, the random variable R has marginal density

$$\begin{aligned}
g(r) &= \int_{-r}^{r} \psi(x,r)\,dx\\
&= \frac{r\exp(-r^2/2)}{\pi}\int_{-r}^{r}\frac{dx}{\sqrt{r^2-x^2}} \qquad \text{put } x = r\cos\theta\\
&= \frac{r\exp(-r^2/2)}{\pi}\int_{\pi}^{0}\frac{-r\sin\theta}{r\sin\theta}\,d\theta = r\exp(-r^2/2).
\end{aligned}$$

The conditional density for X given $R = r$ equals

$$h(x \mid R=r) = \frac{\psi(x,r)}{g(r)} = \frac{1}{\pi\sqrt{r^2-x^2}} \qquad \text{for } |x|<r \text{ and } r>0.$$

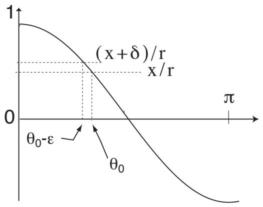
A goodly amount of work.

The calculation is easier when expressed in polar coordinates. From example <11.7> you know how to construct independent $N(0,1)$ distributed random variables by starting with independent random variables $\widetilde{R}$ with density

$$g(r) = r\exp(-r^2/2)\mathbf{1}\{r > 0\},$$

and $U \sim \text{Uniform}(0, 2\pi)$: define $X = \widetilde{R}\cos(U)$ and $Y = \widetilde{R}\sin(U)$.

If we start with X and Y constructed in this way then $R = \sqrt{X^2 + Y^2} = \widetilde{R}$ and the conditional density $h(x \mid R = r)$ is given, for $|x| < r$ by



$$\begin{aligned} &\delta h(x \mid R = r) \\ &\qquad \approx \mathbb{P}\{x \le R\cos(U) \le x + \delta \mid R = r\} \\ &\qquad = \mathbb{P}\{x \le r\cos(U) \le x + \delta\} \qquad \text{by independence of } R \text{ and } U \\ &\qquad = \mathbb{P}\{\theta_0 - \epsilon \le U \le \theta_0\} + \mathbb{P}\{\theta_0 - \epsilon + \pi \le U \le \theta_0 + \pi\} \end{aligned}$$

where θ_0 is the unique value in $[0, \pi]$ for which

$$x/r = \cos(\theta_0) \qquad \text{AND} \qquad (x + \delta)/r = \cos(\theta_0 - \epsilon) \approx \cos(\theta_0) + \epsilon\sin(\theta_0).$$

Solve (approximately) for ϵ then substitute into the expression for the conditional density:

$$\delta h(x \mid R = r) \approx \frac{2\epsilon}{2\pi} \approx \frac{\delta}{\pi r\sin(\theta_0)} = \frac{\delta}{\pi r\sqrt{1 - (x/r)^2}}, \qquad \text{for } |x| < r,$$

the same as before. □

<12.2> **Example.** *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 the standardized variables $(S - \mu)/\sigma$ and $(T - \mu)/\sigma$ have a standard bivariate normal distribution with correlation $\rho = .3$.*

If Sam has a height of $S = 74$ inches, what would one predict about the ultimate height T of his young son Tom?

In standardized units, Sam has height $X = (S - \mu)/\sigma$, which we are given to equal 2.5. Tom's ultimate standardized height is $Y = (T - \mu)/\sigma$. By assumption, before the value of X was known, the pair (X, Y) has a standard bivariate normal distribution with correlation ρ. The conditional distribution of Y given that $X = 2.5$ is

$$Y \mid X = 2.5 \sim N(2.5\rho, 1 - \rho^2)$$

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,

$$\text{Tom's ultimate height} \mid \text{Sam's height} = 74 \text{ inches} \sim N(70.5, 3.64)$$

If I had to make a guess, I would predict that Tom would ultimately reach a height of 70.5 inches. □

> **Remark.** Notice that Tom expected height (given that Sam 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.

<12.3> **Example.** *Let T_i denote the time to the ith point in a Poisson process with rate λ on $[0,\infty)$. Find the joint distribution of (T_1, T_2) conditional on T_3.*

For fixed $0 < t_1 < t_2 < t_3 < \infty$ and suitably small positive $\delta_1, \delta_2, \delta_3$ define disjoint intervals

$$\begin{aligned} &I_1 = [0, t_1) \quad I_2 = [t_1, t_1+\delta_1] \quad I_3 = (t_1+\delta_1, t_2), \\ &I_4 = [t_2, t_2+\delta_2], \quad I_5 = (t_2+\delta_2, t_3), \quad I_6 = [t_3, t_3+\delta_3]. \end{aligned}$$

Write N_j for the number of points landing in I_j, for $j = 1, \dots, 6$. The random variables $N_1, \dots, N_6$ are independent Poissons, with expected values

$$\lambda t_1, \quad \lambda\delta_1, \quad \lambda(t_2 - t_1 - \delta_1), \quad \lambda\delta_2, \quad \lambda(t_3 - t_2 - \delta_2), \quad \lambda\delta_3.$$

To calculate the joint density for (T_1, T_2, T_3) start from

$$\begin{aligned} &\mathbb{P}\{t_1 \le T_1 \le t_1+\delta_1,\ t_2 \le T_2 \le t_2+\delta_2, t_3 \le T_3 \le t_3+\delta_3\} \\ &\quad = \mathbb{P}\{N_1 = 0,\ N_2 = 1,\ N_3 = 0,\ N_4 = 1,\ N_5 = 0,\ N_6 = 1\} \\ &\quad + \text{smaller order terms.} \end{aligned}$$

Here the "smaller order terms" involve probabilities of subsets of events such as $\{N_2 \ge 2,\ N_4 \ge 1,\ N_6 \ge 1\}$, which has very small probability:

$$\mathbb{P}\{N_2 \ge 2\}\mathbb{P}\{N_4 \ge 1\}\mathbb{P}\{N_6 \ge 1\} = o(\delta_1\delta_2\delta_3).$$

Independence also gives a factorization of the main contribution:

$$\begin{aligned} &\mathbb{P}\{N_1 = 0,\ N_2 = 1,\ N_3 = 0,\ N_4 = 1,\ N_5 = 0,\ N_6 = 1\} \\ &\quad = \mathbb{P}\{N_1 = 0\}\mathbb{P}\{N_2 = 1\}\mathbb{P}\{N_3 = 0\}\mathbb{P}\{N_4 = 1\}\mathbb{P}\{N_5 = 0\}\mathbb{P}\{N_6 = 1\} \\ &\quad = e^{-\lambda t_1}[\lambda\delta_1 + o(\delta_1)]e^{-\lambda(t_2-t_1-\delta_1)} \times \\ &\qquad\quad \times [\lambda\delta_2 + o(\delta_2)]e^{-\lambda(t_3-t_2-\delta_2)}[\lambda\delta_3 + o(\delta_3)] \\ &\quad = \lambda^3\delta_1\delta_2\delta_3\, e^{-\lambda t_3} + o(\delta_1\delta_2\delta_3) \end{aligned}$$

If you think of Δ as a small shoebox (hyperrectangle) with sides δ_1, δ_2, and δ_3, with all three δ_j's of comparable magnitude (you could even take $\delta_1 = \delta_2 = \delta_3$), the preceding calculations reduce to

$$\mathbb{P}\{(T_1, T_2, T_3) \in \Delta\} = (\text{volume of } \Delta)\lambda^3 e^{-\lambda t_3} + \text{smaller order terms}$$

where the "smaller order terms" are small relative to the volume of Δ. Thus the joint density for (T_1, T_2, T_3) is

$$f(t_1, t_2, t_3) = \lambda^3 e^{-\lambda t_3} \mathbb{I}\{0 < t_1 < t_2 < t_3\}.$$

Remark. The indicator function is very important. Without it you would be unpleasantly surprised to find $\iiint_{\mathbb{R}^3} f = \infty$.

Just as a check, calculate the marginal density for T_3 as

$$\begin{aligned} g(t_3) &= \iint_{\mathbb{R}^2} f(t_1, t_2, t_3)\, dt_1\, dt_2 \\ &= \lambda^3 e^{-\lambda t_3} \iint \mathbb{I}\{0 < t_1 < t_2 < t_3\}\, dt_1\, dt_2. \end{aligned}$$

The double integral equals

$$\int \mathbb{I}\{0 < t_2 < t_3\} \left(\int_0^{t_2} 1\, dt_1 \right) = \int_0^{t_3} t_2\, dt_2 = \tfrac{1}{2} t_3^2.$$

That is, T_3 has marginal density

$$g(t_3) = \tfrac{1}{2} \lambda^3 t_3^2 e^{-\lambda t_3} \mathbb{I}\{t_3 > 0\},$$

which agrees with the result calculated in Example <10.1>.

Calculate the conditional density for a given $t_3 > 0$ as

$$\begin{aligned} h(t_1, t_2 \mid T_3 = t_3) &= \frac{f(t_1, t_2, t_3)}{g(t_3)} \\ &= \frac{\lambda^3 e^{-\lambda t_3} \mathbb{I}\{0 < t_1 < t_2 < t_3\}}{\frac{1}{2}\lambda^3 t_3^2 e^{-\lambda t_3}} \\ &= \frac{2}{t_3^2} \mathbb{I}\{0 < t_1 < t_2 < t_3\}. \end{aligned}$$

That is, conditional on $T_3 = t_3$, the pair (T_1, T_2) is uniformly distributed in a triangular region of area $t_3^2/2$. □

Chapter 13

Moment generating functions

13.1 Basic facts

Formally the moment generating function is obtained by substituting $s = e^t$ in the probability generating function.

Definition. The moment generating function (m.g.f.) of a random variable X is the function M_X defined by $M_X(t) = \mathbb{E}(e^{Xt})$ for those real t at which the expectation is well defined.

Unfortunately, for some distributions the moment generating function is finite only at $t = 0$. The Cauchy distribution, with density

$$f(x) = \frac{1}{\pi(1+x^2)} \qquad \text{for all } x \in \mathbb{R},$$

is an example.

Remark. The problem with existence and finiteness is avoided if t is replaced by it, where t is real and $i = \sqrt{-1}$. In probability theory the function $\mathbb{E}e^{iXt}$ is usually called the *characteristic function*, even though the more standard term *Fourier transform* would cause less confusion.

When the m.g.f. is finite in a neighborhood of the origin it can be expanded in a power series, which gives us some information about the ***moments*** (the values of $\mathbb{E}X^k$ for $k = 1, 2, \dots$) of the distribution:

$$\mathbb{E}(e^{Xt}) = \sum_{k=0}^{\infty} \frac{\mathbb{E}(Xt)^k}{k!}$$

The coefficient of $t^k/k!$ in the series expansion of $M(t)$ equals the kth moment, $\mathbb{E}X^k$.

<13.1> **Example.** Suppose X has a standard normal distribution. Its moment generating function equals $\exp(t^2/2)$, for all real t, because

$$\begin{aligned}\int_{-\infty}^{\infty} e^{xt}\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\exp\left(-\frac{(x-t)^2}{2}+\frac{t^2}{2}\right)dx\\ &= \exp\left(\frac{t^2}{2}\right).\end{aligned}$$

For the last equality, compare with the fact that the $N(t,1)$ density integrates to 1.

The exponential in $M_X(t)$ expands to

$$\sum_{m=0}^{\infty}\frac{1}{m!}\left(\frac{t^2}{2}\right)^m = \sum_{m=0}^{\infty}\left(\frac{(2m)!}{m!2^m}\right)\frac{t^{2m}}{(2m)!}$$

Pick off coefficients.

$$\begin{aligned}\mathbb{E}X^2 &= \frac{2!}{1!2^1} = 1 \qquad \text{(you knew that)}\\ \mathbb{E}X^4 &= \frac{4!}{2!2^2} = 3\\ &\dots\\ \mathbb{E}(X^{2m}) &= \frac{(2m)!}{m!2^m} \qquad \text{for } m \text{ a positive integer.}\end{aligned}$$

The coefficient for each odd power of t equals zero, which reflects the fact that $\mathbb{E}X^k = 0$, by anti-symmetry, if k is odd.

□

<13.2> **Example.** If $X \sim \text{gamma}(\alpha)$, with $\alpha > 0$, then for $t < 1$

$$\begin{aligned}M_X(t) &= \frac{1}{\Gamma(\alpha)}\int_0^{\infty} e^{xt}x^{\alpha-1}e^{-x}dx\\ &= \frac{1}{(1-t)^{\alpha}\Gamma(\alpha)}\int_0^{\infty} y^{\alpha-1}e^{-y}dy \qquad \text{putting } y=(1-t)x\\ &= (1-t)^{-\alpha}.\end{aligned}$$

For $t \ge 1$ the integral diverges and $M_X(t) = \infty$. For $|t| < 1$,

$$\begin{aligned}M_X(t) &= \sum\nolimits_{k=0}^{\infty}\binom{-\alpha}{k}(-t)^k\\ &= \sum\nolimits_{k=0}^{\infty}(-1)^k\frac{(-\alpha)(-\alpha-1)\dots(-\alpha-k+1)}{k!}t^k.\end{aligned}$$

The kth moment, $\mathbb{E}(X^k)$, equals $(\alpha+k-1))(\alpha+k-2)\dots(\alpha)$, the coeeficient of $t^k/k!$. Compare with the direct calculation in Example <10.3>.

□

13.2 MGF's determine distributions

If two random variables X and Y have moment generating functions that are finite and equal in some neighborhood of 0 then they have the same distributions. This result is much harder to prove than its analog for probability generating functions.

For example, if $M_X(t) = e^{t^2/2}$, even just for t near 0, then X must have a $N(0,1)$ distribution.

13.3 Approximations via moment generating functions

If $X_n = \xi_1 + \dots + \xi_n$ with the ξ_i's independently Ber(p) distributed then

$$\begin{aligned} M_{X_n}(t) &= \mathbb{E}\left(e^{t\xi_1}e^{t\xi_2}\dots e^{t\xi_n}\right) \\ &= \left(\mathbb{E}e^{t\xi_1}\right)\left(\mathbb{E}e^{t\xi_2}\right)\dots\left(\mathbb{E}e^{t\xi_n}\right) \qquad \text{by independence} \\ &= \left(q+pe^t\right)^n. \end{aligned}$$

That is, the Bin(n,p) has m.g.f. $\left(q+pe^t\right)^n$.

Write q for $1-p$ and σ_n^2 for npq. You know that the standardized random variable $Z_n := (X_n - np)/\sigma_n$ is approximately $N(0,1)$ distributed. The moment generating function $M_{Z_n}(t)$ also suggests such an approximation. Then

$$\begin{aligned} M_{Z_n}(t) &= \mathbb{E}e^{t(X-np)/\sigma_n} \\ &= e^{-npt/\sigma}\mathbb{E}e^{X(t/\sigma_n)} = e^{-npt/\sigma}M_{X_n}(t/\sigma_n) \\ &= e^{-npt/\sigma_n}\left(q+pe^{t/\sigma_n}\right)^n \\ &= \left(qe^{-pt/\sigma_n}+pe^{qt/\sigma_n}\right)^n. \end{aligned}$$

The power series expansion for $qe^{-pt/\sigma}+pe^{qt/\sigma}$ simplifies:

$$\begin{aligned} &q\left(1-\frac{pt}{\sigma}+\frac{p^2t^2}{2!\sigma^2}-\frac{p^3t^3}{3!\sigma^3}+\dots\right)+p\left(1+\frac{qt}{\sigma}+\frac{q^2t^2}{2!\sigma^2}-\frac{q^3t^3}{3!\sigma^3}+\dots\right) \\ &= 1+\frac{pqt}{2\sigma^2}+\frac{pq(p-q)t^3}{6\sigma^3}+\dots \end{aligned}$$

For large n use the series expansion $\log(1+z)^n = n(z-z^2/2+\dots)$ to deduce that

$$\log M_{Z_n}(t) = \frac{t^2}{2} + \frac{(q-p)t^3}{6\sqrt{npq}} + \text{ terms of order } \frac{1}{n} \text{ or smaller}$$

The $t^2/2$ term agree with the logarithm of the moment generating function for the standard normal. As n tends to infinity, the remainder terms tend to zero.

The convergence of $M_{Z_n}(t)$ to $e^{t^2/2}$ can be used to prove rigorously that the distribution of the standardized Binomial "converges to the standard normal" as n tends to infinity. In fact the series expansion for $\log M_n(t)$ is the starting point for a more precise approximation result—but for that story you will have to take the more advanced probability course Statistics 330.