Read M&M §2.6. Read M&M Chapter 4 only if you want to. Probability and randomness. Conditional probabilites. Bayes's rule. Random variables. Means and variances.

In these notes I have included more of the mathematical reasoning than you really need to know for this course. Sections leaders will tell you which parts, if any, they want you to understand. Some Sections might use M&M, and ignore these notes altogether.

1. Probability rules

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

I will refer to any situation where outcomes are random as an ***experiment***, for the sake of a concise description. Please do not confuse the term with the special case of designed experiments, as described in Chapter 3 of M&M.

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

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

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

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

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

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

If the observed outcome of an experiment lies in the set defining some particular event, one says that the event has occurred. For example, with the outcome hhh each of the events {no tails}, {at least one head}, {more heads than tails} occurs, but the event {even number of heads} does not.

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

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

That is, probabilities are essentially proportions of times that events occur in many repetitions of whatever experiment was generating the random outcomes.

As many authors have pointed out, there is something slightly fishy about this interpretation. For example, it is difficult to make precise the meaning of "independent

repetitions" without resorting to explanations that degenerate into circular discussions about the meaning of probability and independence. This fact does not seem to trouble most supporters of the frequency theory. The interpretation is regarded as a justification for the adoption of a set of mathematical rules, or axioms.

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

Rules for probabilities

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

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

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

(P4) : If an event E is broken into disjoint pieces $E_1, E_2, \ldots$ then $\mathbb{P}E = \sum_i \mathbb{P}E_i$.

For rule (P4), the event E could also be called a (disjoint) ***union*** of the events $E_1, E_2, \ldots$, and written as

$$E = E_1 \cup E_2 \cup E_3 \cup \ldots$$

or (in M&M notation)

$$E = E_1 \text{ OR } E_2 \text{ OR } E_3 \text{ OR } \ldots$$

The notation A OR B means the event where either A or B, or both, occur. The notation A AND B means that both A and B occur. Sometimes A AND B is written as $A \cap B$, or just AB, the ***intersection*** of the two events. For A and B to be disjoint events, the intersection A AND B must be the empty subset of the sample space. That is, two events that are disjoint can never occur together.

<1> **Example.** Find $\mathbb{P}${at least two heads} for the tossing of three coins. Use the sample space from the previous page. If we *assume* that each coin is fair and that the outcomes from the coins don't affect each other ("independence"), then we must conclude by symmetry ("equally likely") that

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

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

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

□

In general, if we have a sample space consisting of N outcomes, say $s_1, s_2, \ldots, s_N$, then rules (P3) and (P4) imply

$$\mathbb{P}\{s_1\} + \mathbb{P}\{s_2\} + \ldots + \mathbb{P}\{s_N\} = 1$$

In particular, if each of the events $\{s_i\}$ is equally likely, then each must have probability $1/N$. For that special case, the calculation of probabilities reduces to counting: if an event A consists of k outcomes from the sample space then $\mathbb{P}A = k/N$. Typically appeals to symmetry lead to sample spaces with equal probability attached to each outcome.

Probability theory would be very boring if all problems were solved as in Example <1>: break the event into pieces whose probabilities you know, then add. Thing become much more interesting when we recognize that the assignment of probabilities depends upon what we know or have learnt (or assume) about the random situation. For example, in the last problem we could have written

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

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

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

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

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

or

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

or

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

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

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

Rule for conditional probability

(P5) : if A and B are events then

$$\mathbb{P}\big(A \text{ AND } B \mid \text{info}\big) = \mathbb{P}\big(A \mid \text{info}\big) \times \mathbb{P}\big(B \mid A \text{ AND info}\big).$$

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

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

Then, for big N,

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

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_{AB} of which B occurs. Thus $\mathbb{P}\big(B \mid A, \text{info}\big) \approx N_{AB}/N_A$. The rest is multiplication.

<2> **Example.** M&M (§2.6, Example 2.32) discuss "Simpson's paradox" by means of an (artificial, I believe) example using survival rates of patients undergoing surgery at two hospitals. The "data" come in the form of a three-way cross classification of patients: by hospital (A or B), condition (good or poor) before operation, and by fate (died or survived 6 weeks).

GOOD CONDITION

	Hospital A	Hospital B	total
Died	6	8	14
Survived	594	592	1186
total	600	600	1200

POOR CONDITION

	Hospital A	Hospital B	total
Died	57	8	65
Survived	1443	192	1635
total	1500	200	1700

The paradox can also be understood in terms of conditional probabilities, if we consider the characteristics of a single patient chosen at random from the 2900 individuals counted in the tables.

The sample space would consist of 2900 items (patient identifiers) each with probability 1/2900. Consider the events

$$A = \text{patient entered hospital A}$$
$$G = \text{patient in good condition before operation}$$
$$D = \text{patient died}$$

Write $B = A^c$ for the complement of the event A, that is,

$$B = A^c = \text{patient entered hospital B}$$

and so on.

Under the assumption that each patient has the same probability (1/2900) of being chosen, all probabilities and conditional probabilities reduce to proportions. If we calculate probabilities conditional on some event, then we should take the proportion only amongst patients included in that event. For example,

$$\mathbb{P}A = \frac{600+1500}{2900} \approx 0.724 \qquad \text{cf. marginal totals for hospital A}$$
$$\mathbb{P}(A \mid G) = \frac{600}{1200} = 0.500 \qquad \text{only 1200 in good condition}$$
$$\mathbb{P}(G \mid A) = \frac{600}{600+1500} \approx .286$$

From the point of view of a new patient who is about to choose a hospital at which to have surgery, the relevant conditional probabilities are

$$\mathbb{P}(D \mid A \text{ AND } G) = \frac{6}{600} = 0.010$$
$$\mathbb{P}(D \mid B \text{ AND } G) = \frac{8}{600} \approx 0.013$$
$$\mathbb{P}(D \mid A \text{ AND } G^c) = \frac{57}{1500} \approx 0.038$$
$$\mathbb{P}(D \mid B \text{ AND } G^c) = \frac{8}{200} = 0.040$$

Notice that

$$\mathbb{P}(D \mid A \text{ AND } G) < \mathbb{P}(D \mid B \text{ AND } G)$$
$$\mathbb{P}(D \mid A \text{ AND } G^c) < \mathbb{P}(D \mid B \text{ AND } G^c)$$

No matter whether the new patient is in good or poor condition, it would seem that hospital A is preferrable—a smaller probability of death for either condition.

The apparent paradox comes if we ignore the information about the condition of patients before the operation:

$$\mathbb{P}(D \mid A) = \frac{6+57}{600+1500} = 0.030$$
$$\mathbb{P}(D \mid B) = \frac{8+8}{600+200} = 0.020$$

That is, the overall chances of death were lower for hospital B.

The apparent paradox is resolved when we notice that many more patients who entered hospital A were in poor condition to begin with. The calculation of conditional

probabilities exposes the source of the paradox.

$$\begin{aligned}
\mathbb{P}(D \mid A) &= \frac{\mathbb{P}(D \text{ AND } A)}{\mathbb{P}(A)} && \text{rule (P5)} \\
&= \frac{\mathbb{P}(D \text{ AND } A \text{ AND } G) + \mathbb{P}(D \text{ AND } A \text{ AND } G^c)}{\mathbb{P}(A)} && \text{rule (P4) for numerator} \\
&= \frac{\mathbb{P}(D \mid A \text{ AND } G)\mathbb{P}(A \text{ AND } G) + \mathbb{P}(D \mid A \text{ AND } G^c)\mathbb{P}(A \text{ AND } G^c)}{\mathbb{P}(A)} && \text{rule (P5)} \\
(*) \qquad &= \mathbb{P}(D \mid A \text{ AND } G)\mathbb{P}(G \mid A) + \mathbb{P}(D \mid A \text{ AND } G^c)\mathbb{P}(G^c \mid A) && \text{rule (P5)}
\end{aligned}$$

Compare with the decomposition $\mathbb{P}(D) = \mathbb{P}(D \mid G)\mathbb{P}(G) + \mathbb{P}(D \mid G^c)\mathbb{P}(G^c)$, or

$$\mathbb{P}(D \mid \text{info}) = \mathbb{P}(D \mid G \text{ AND info})\mathbb{P}(G \mid \text{info}) + \mathbb{P}(D \mid G^c \text{ AND info})\mathbb{P}(G^c \mid \text{info}),$$

with the event A playing the role of the "info". Plug the values for the conditional probabilities into (*):

$$\mathbb{P}(D \mid A) \approx 0.010 \times \frac{600}{2100} + 0.038 \times \frac{1500}{2100}$$

Similarly,

$$\mathbb{P}(D \mid B) \approx 0.013 \times \frac{600}{800} + 0.040 \times \frac{200}{800}$$

The calculation for $\mathbb{P}(D \mid A)$ puts more weight (1500/2100) on the poor condition than the calculation for $\mathbb{P}(D \mid B)$, withweight 200/800. □

You can safely skip the next example if you are allergic to arithmetic. I include it merely to make the point that you can solve quite complicated problems by breaking them into smaller, more manageable pieces, then reassembling the pieces by means of the rules for probabilities.

<4> **Example.** What is the probability that a hand of 5 cards contains four of a kind?

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

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

$$\{\text{four of a kind}\} = F_1 \text{ OR } F_2 \text{ OR } \ldots \text{ OR } F_{13}$$

where

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

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

$$\mathbb{P}\{\text{four of a kind}\} = \mathbb{P}F_1 + \mathbb{P}F_2 + \ldots + \mathbb{P}F_{13} = 13\mathbb{P}F_1.$$

Now break F_1 into simpler pieces,

$$F_1 = F_{1,1} \text{ OR } F_{1,2} \text{ OR } \ldots \text{ OR } F_{1,5}$$

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",

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

where $N_1 = \{\text{first card is not an ace}\}$, $A_1 = \{\text{first card is an ace}\}$, and so on. To save on space, I will omit the AND, writing $N_1A_2A_3A_4$ instead of N_1 AND A_2 AND A_3 AND A_4, and so on. By repeated application of rule P5,

$$\begin{aligned}
\mathbb{P}F_{1,1} &= \mathbb{P}N_1\,\mathbb{P}(A_2A_3A_4A_5 \mid N_1) \\
&= \mathbb{P}N_1\,\mathbb{P}(A_2 \mid N_1)\mathbb{P}(A_3A_4A_5 \mid N_1A_2) \\
&= \ldots \\
&= \mathbb{P}N_1\,\mathbb{P}(A_2 \mid N_1)\,\mathbb{P}(A_3 \mid N_1A_2)\,\ldots\,\mathbb{P}(A5 \mid N_1A_2A_3A_4) \\
&= \frac{48}{52} \times \frac{4}{51} \times \frac{3}{50} \times \frac{2}{49} \times \frac{1}{48}.
\end{aligned}$$

Thus

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

□ Can you see any hidden assumptions in this analysis?

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

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

<5> **Example.**

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

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

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

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

There are three events to be considered

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

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

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

from which we deduce (via rule P5) that

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

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

Notice that

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

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

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

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

he would calculate a conditional probability

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

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

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

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

In Example <5> we had a situation where a particular piece of information could be ignored in the calculation of another conditional probability, namely $\mathbb{P}\big(\mathcal{A} \mid B^*\big) = \mathbb{P}(\mathcal{A})$. Such a situation is an instances of a property called ***independence***.

<6> **Definition.** *Call events E and F conditionally independent given a particular piece of information if*

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

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

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

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

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

As an example, you might assume that the event {president gets impeached} is independent of the event {I eat a yoghurt for breakfast tomorrow}, but that it is not independent of the event {time to next election is less than one year}. Would you judge it independent of an event like {national newspaper breaks story of sexual shenanigans on part of some member of a House committee}? What other conditioning information would you be assuming?

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. But be careful: convenient assumptions need not even crudely approximate reality. See the famous case of *People v. Collins*† for an example where independence is a dubious assumption.

2. Bayes's rule

The Bayesian argument consists of little more than a routine application of the rules of probability, with conditioning playing the leading role. I see no reason to memorize the formula behind the Bayesian calculation, unless you intend to convert to Bayesianism.

<7> **Example.** The great Sherlock is trying to solve the case of the battered cod. The fishmonger has been murdered. Sherlock knows that 30% of the murders in town are usually committed by Freddie the Fish, 50% by the Evil Pilchard, and only 20% by old Mrs. Smith. He also knows something about the modus operandi of each villain: half the time Freddie dispatches his victim by a blow to the head with whatever seafood is handy, and the rest of the time he uses his trusty whale harpoon; Pilchard also uses the seafood method 10% of the time, but he prefers other techniques (toxic hamburgers, mad cow virus, and other methods too terrible to describe), which he employs 90% of the time; Mrs. Smith carries an old fish of some description, which she invariably uses as her murder weapon, in her handbag.

The fishmonger was found slumped over his counter with a fish-shaped indentation in his skull. The coroner declares death to have been caused by repeated assault with a blunt, smelly instrument—most likely a cod. What can the great Sherlock deduce?

Sherlock denotes by F the event that Freddie was the murderer, and denotes by E and S the other two possibilities. He writes C for the event that the murder is committed by fishy means (such as cod). His ***prior*** knowledge he writes down as

$$\mathbb{P}(F) = 0.3, \qquad \mathbb{P}(E) = 0.5 \qquad \mathbb{P}(S) = 0.2$$

† discussed, for example, in "Statistics and Public Policy" by Fairley and Mosteller (Addison-Wesley 1997) and also in the Freedman, Pisani, Purves, Adhikari text

His vast knowledge of criminal behaviour he distills into

$$\mathbb{P}(C \mid F) = 0.5 \qquad \mathbb{P}(C \mid E) = 0.1 \qquad \mathbb{P}(C \mid S) = 1$$

As the awed Dr. Watson looks on, Sherlock employs the Calculus of Probability (on which he has written a small treatise) to calculate his ***posterior*** opinion in the form of probabilities conditional on the coroner's verdict.

$$\mathbb{P}(F \mid C) = \frac{\mathbb{P}(F \text{ AND } C)}{\mathbb{P}(C)}$$

He breaks the denominator into a sum of terms

$$\mathbb{P}(C \text{ AND } F) + \mathbb{P}(C \text{ AND } E) + \mathbb{P}(C \text{ AND } S)$$

like the numerator, and then he invokes rule (P5) for each term, reducing the ratio to

$$\frac{\mathbb{P}(C \mid F)\mathbb{P}(F)}{\mathbb{P}(C \mid F)\mathbb{P}(F) + \mathbb{P}(C \mid E)\mathbb{P}(E) + \mathbb{P}(C \mid S)\mathbb{P}(S)}$$

His object throughout has been the reexpression of the unknown conditional probability in terms of what already knows.

With a cry of triumph, Sherlock sees that all factors are now within his grasp. A quick mental calculation yields

$$\mathbb{P}(F \mid C) = \frac{0.5 \times 0.3}{(0.5 \times 0.3) + (0.1 \times 0.5) + (1 \times 0.1)} = 0.500$$

His mind races as he calculates the posterior probabilities for Pilchard in similar fashion:

$$\begin{aligned}\mathbb{P}(E \mid C) &= \frac{\mathbb{P}(C \mid E)\mathbb{P}(E)}{\mathbb{P}(C \mid F)\mathbb{P}(F) + \mathbb{P}(C \mid E)\mathbb{P}(E) + \mathbb{P}(C \mid S)\mathbb{P}(S)} \\ &= \frac{0.1 \times 0.5}{(0.5 \times 0.3) + (0.1 \times 0.5) + (1 \times 0.1)} \approx 0.167\end{aligned}$$

With smug satisfaction he notes, just as he had expected, the same denominator as before.

Knowing that probabilities conditional on any information must still add to one, Sherlock realizes that $\mathbb{P}(S \mid C)$ must account for the missing fraction $0.333 = 1 - 0.500 - 0.167$.

The suspicion of guilt is wafting more strongly towards Freddie and Mrs. Smith, and away from Pilchard. □

Notice the effect of the evidence on the ***odds ratios***:

$$\frac{\mathbb{P}(S \mid C)}{\mathbb{P}(F \mid C)} = \frac{\mathbb{P}(C \mid S) \times \mathbb{P}(S)/\mathbb{P}(C)}{\mathbb{P}(C \mid F) \times \mathbb{P}(F)/\mathbb{P}(C)} = \frac{1}{0.5} \times \frac{\mathbb{P}(S)}{\mathbb{P}(F)}$$

Mrs. Smith's greater propensity for fish battery is reflected in Sherlock's posterior judgements about the relative chances of guilt for Smith versus Freddie.

3. Random variables (with discrete distributions)

Events either happen or they don't. They are a rather all-or-nothing way of saying something about the outcome of some random experiment. ***Random variables*** give more detailed information about an experiment, in the form of a number that is determined by the outcome.

For example, suppose an experiment consisted of choosing a simple random sample of size 51 from the Registrar's list of all Yale undergraduates for the current academic year. The median age, M, of the students in the sample would be a random variable, as would the number of seniors in the sample, or the minimum of the SAT scores of all students in the sample. The August rainfall would not be a random variable for

this particular experiment, because there is no meaningful way of attaching a figure for rainfall to each subset of 51 undergraduates.

Formally, a random variable is just a function that attaches a number to each outcome listed in the sample space. We typically don't need to specify the sample space 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.

<8> **Example.** From a set of five students—Achilles (aged 19), Bacchus (aged 21), Cleo (aged 19), Dionysus (aged 21), and Epithelium (aged 18)—I take a simple random sample of size 2. What is the distribution of the mean age of the students in my sample? What is the probability of the event {mean age in sample < 20}?

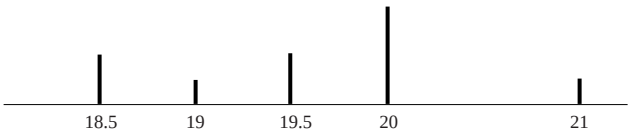
Write X for the mean age in the sample. Write [ab] for the outcome that the sample consists of Achilles and Bacchus, and so on. (There is no implied ordering in the sample.)

outcome	[ab]	[ac]	[ad]	[ae]	[bc]	[bd]	[be]	[cd]	[ce]	[de]
value of X	20	19	20	18.5	20	21	19.5	20	18.5	19.5

Each of the ten possible outcomes has the same probability, namely 1/10. For this sampling experiment, the random variable X takes values 18.5, 19, 19.5, 20, and 21, with the probabilities shown.

x	18.5	19	19.5	20	21
$\mathbb{P}\{X = x\}$	2/10	1/10	2/10	4/10	1/10

The distribution can also be represented by drawing lines of height p_i at each x_i:



The event $\{X < 20\}$ is a disjoint union of events,

$$\{X = 18.5\} \cup \{X = 19\} \cup \{X = 19.5\},$$

which has probability

$$\frac{2}{10} + \frac{1}{10} + \frac{2}{10} = \frac{1}{2}.$$

Alternatively,

$$\{X < 20\} = \{[ac], [ae], [be], [ce], [de]\}$$

Each of the five outcomes contributes probability 1/10 to $\mathbb{P}\{X < 20\}$.

□

4. Means and variances

If a random variable takes values $x_1, x_2, \ldots, x_k$ with probabilities $p_1, p_2, \ldots, p_k$, its ***mean*** is defined as the weighted average

$$\mu_X = p_1 x_1 + p_2 x_2 + \ldots + p_k x_k$$

The mean is also called the ***expected value*** or ***expectation*** of X, and is denoted by a symbol like $\mathbb{E}(X)$. As M&M (page 327) note, the term "expected" is slightly misleading, because one does not necessarily expect X to take the value μ_X. Remember

the old canard about statisticians expecting families to have 2.1 children if you feel yourself taking your expectations too literally.

You don't have to calculate the distribution of a random variable to find its mean. For example, suppose X is defined on a sample space $S = \{s_1, s_2, \ldots, s_{10}\}$, with values:

$$\begin{aligned} X(s_1) = X(s_2) &= 16 \\ X(s_3) = X(s_4) = X(s_5) = X(s_6) &= 11 \\ X(s_7) &= 23 \\ X(s_8) = X(s_9) = X(s_{10}) &= 100 \end{aligned}$$

Then

$$\begin{aligned} \mu_X &= 16 \times \mathbb{P}\{s_1, s_2\} + 11 \times \mathbb{P}\{s_3, s_4, s_5, s_6\} + 23 \times \mathbb{P}\{s_7\} + 100 \times \mathbb{P}\{s_8, s_9, s_{10}\} \\ &= 16 \times (\mathbb{P}\{s_1\} + \mathbb{P}\{s_2\}) + 11 \times (\mathbb{P}\{s_3\} + \mathbb{P}\{s_4\} + \mathbb{P}\{s_5\} + \mathbb{P}\{s_6\}) \\ &\quad + 23 \times \mathbb{P}\{s_7\} + 100 \times (\mathbb{P}\{s_8\} + \mathbb{P}\{s_9\} + \mathbb{P}\{s_{10}\}) \qquad \text{by rule (P4)} \\ &= \sum_{i=1}^{10} X(s_i)\mathbb{P}\{s_i\} \end{aligned}$$

A similar argument works in the general case.

The name "mean" fits with the concept of a mean of a set of N numbers. If the value x_i occurs exactly N_i times amongst the numbers, for $i = 1, 2, \ldots, k$, then the mean $\bar{x}$ of the numbers equals

$$\bar{x} = \frac{N_1 x_1 + N_2 x_2 + \ldots + N_k x_k}{N} = p_1 x_1 + p_2 x_2 + \ldots + p_k x_k$$

where $p_i = N_i/N$ denotes the proportion of the numbers equal to x_i. By the same reasoning, after accounting for mutiplicities, we can write the variance of the set of numbers as

$$\frac{1}{N-1} \sum_i N_i (x_i - \bar{x})^2$$

If N is very large then $N_i/(N-1)$ is close to p_i, and the variance is practically the same as $\sum_i p_i (x_i - \bar{x})^2$.

Analogously, the ***variance*** of a random variable X is defined as

$$\sigma_X^2 = \sum_i p_i (x_i - \mu_X)^2 \qquad \text{where } p_i = \mathbb{P}\{X = x_i\},$$

the sum ranging over all values x_i that X can take. The variance is often also written as $\text{var}(X)$. The standard deviation σ_X is the square root of the variance.

If X is a random variable taking values $x_1, \ldots, x_k$ with probabilities $p_1, \ldots, p_k$, and α, β are constants, then the new random variable $Y = \alpha + \beta X$ takes values $y_i = \alpha + \beta x_i$ with probabilities $p_1, \ldots, p_k$. The new random variable has mean

$$\mu_Y = \sum_i p_i (\alpha + \beta x_i) = \alpha \sum_i p_i + \beta \sum_i p_i x_i = \alpha + \beta \mu_X$$

and variance

$$\begin{aligned} \sigma_Y^2 &= \sum_i p_i (y_i - \mu_Y)^2 \\ &= \sum_i (\alpha + \beta x_i - \alpha - \beta \mu_X)^2 \\ &= \beta^2 \sum_i (x_i - \mu_X)^2 \\ &= \beta^2 \sigma_X^2 \end{aligned}$$

That is, for constants α and β, and a random variable X,

$$\mathbb{E}(\alpha + \beta X) = \alpha + \beta \mathbb{E}(X)$$
$$\text{var}(\alpha + \beta X) = \beta^2 \text{var}(X)$$

or in M&M notation:

for constants α and β, and a random variable X,
$$\mu_{\alpha+\beta X} = \alpha + \beta \mu_X$$
$$\sigma^2_{\alpha+\beta X} = \beta^2 \sigma^2_X$$

As a particular case, notice that $\text{var}(-X) = \text{var}(X)$. Don't forget, variances cannot be negative.

5. Mean of a sum of random variables

If X and Y are random variables defined on a sample space $S = \{s_1, \ldots, s_N\}$, with $X(s_i) = x_i$ and $Y(s_i) = y_i$, then the new random variable $Z = X + Y$ takes the value $z_i = x_i + y_i$ at s_i and it has expectation

$$\mathbb{E}(Z) = \sum_i (x_i + y_i)\mathbb{P}\{s_i\} = \sum_i x_i \mathbb{P}\{s_i\} + \sum_i y_i \mathbb{P}\{s_i\} = \mathbb{E}(X) + \mathbb{E}(Y)$$

In M&M notation,

$$\boxed{\mu_{X+Y} = \mu_X + \mu_Y}$$

A similar formula works for sums of more than two random variables.

<9> **Example.** Suppose a coin has probability p of landing heads on any particular toss. Let X denote the number of heads obtained from n tosses. We can write X as a sum $X_1 + X_2 + \ldots + X_n$, where

$$X_i = \begin{cases} 1 & \text{if } i\text{th toss lands heads} \\ 0 & \text{if } i\text{th toss lands tails} \end{cases}$$

Each takes the value 1 with probability p and 0 with probability $1 - p$, giving a mean of $1 \times p + 0 \times (1 - p) = p$. Thus

$$\mu_X = \mu_{X_1} + \mu_{X_2} + \ldots + \mu_{X_n} = np$$

□ Sound reasonable?

6. Independent random variables

Two random variables X and Y are said to be ***independent*** if "knowledge of the value of X takes does not help us to predict the value Y takes", and vice versa. More formally, for each possible pair of values x_i and y_j,

$$\mathbb{P}\{Y = y_j \mid X = x_i\} = \mathbb{P}\{Y = y_j\},$$

that is,

$$\mathbb{P}\{Y = y_j \text{ AND } X = x_i\} = \mathbb{P}\{Y = y_j\} \times \mathbb{P}\{X = x_i\} \qquad \text{for all } x_i \text{ and } y_j,$$

and in general, events involving only X are independent of events involving only Y:

$$\mathbb{P}\{\text{something about } X \text{ AND something else about } Y\}$$
$$= \mathbb{P}\{\text{something about } X\} \times \mathbb{P}\{\text{something else about } Y\}$$

This factorization leads to other factorizations for independent random variables:

$$\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y) \qquad \text{if } X \text{ and } Y \text{ are independent}$$

or in M&M notation:

$$\boxed{\mu_{XY} = \mu_X \mu_Y \qquad \text{if } X \text{ and } Y \text{ are independent}}$$

You should skip the rest of this section if you don't like algebra.

To see why this result should be true, suppose X and Y are defined on a sample space $S = \{s_1, s_2, \ldots\}$. Then

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

Collect together all those s_i for which $X(s_i) = x_j$ and $Y(s_i) = y_k$, appealing to rule (P4), to consolidate the last sum into

$$\sum_{j,k} x_j y_k \mathbb{P}\{X = x_j \text{ AND } Y = y_k\}$$

Factorize each of the probabilities $\mathbb{P}\{X = x_j \text{ AND } Y = y_k\}$ into $\mathbb{P}\{X = x_j\} \times \mathbb{P}\{Y = y_k\}$, then recognize the result as

$$\left(\sum_j x_j \mathbb{P}\{X = x_j\}\right)\left(\sum_k y_k \mathbb{P}\{Y = y_k\}\right),$$

the product of the two expected values.

7. Variances of sums of independent random variables

Standard errors provide one measure of spread for the disribution of a random variable. If we add together several random variables the spread in the distribution increases, in general. For independent summands the increase in the spread is not as much as you might imagine: it is not just a matter of adding together standard deviations.

The key result is:

$$(*) \qquad \boxed{\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y \qquad \text{if } X \text{ and } Y \text{ are independent random variables}}$$

If $Y = -Z$, for another random variable Z, then we get

$$\sigma^2_{X-Z} = \sigma^2_X + \sigma^2_{-Z} = \sigma^2_X + \sigma^2_Z \qquad \text{if } X \text{ and } Z \text{ are independent}$$

Notice the plus sign on the right-hand side: subtracting an independent quantity from X cannot decrease the spread in its distribution.

A simlar result holds for sums of more than two random variables:

$$\sigma^2_{X_1+X_2+\ldots+X_n} = \sigma^2_{X_1} + \sigma^2_{X_2} + \ldots + \sigma^2_{X_n} \qquad \text{for independent } X_1, X_2, \ldots$$

In particular, if each X_i has the same variance, σ^2 then the variance of the sum increases as $n\sigma^2$, and the standard deviation increases as $\sqrt{n}\sigma$. It is this $\sqrt{n}$ rate of growth in the spread that makes a lot of statistical theory work.

You should skip the rest of this section if you don't like algebra.

To understand where $(*)$ comes from, simplify notation by writing $\widetilde{X}$ for $X - \mu_X$ and $\widetilde{Y}$ for $Y - \mu_Y$. Subtraction of a constant cannot create dependence when none existed before: if X and Y are independent then $\widetilde{X}$ and $\widetilde{Y}$ are independent.

Don't confuse $\widetilde{X}$ and $\widetilde{Y}$ with the $\widetilde{x}_i$ and $\widetilde{y}_i$ from Lecture 2. Those quantities were also scaled to have unit variance.

Note that $\mathbb{E}\widetilde{X} = \mathbb{E}\widetilde{Y} = 0$, and $\text{var}(X) = \text{var}(\widetilde{X}) = \mathbb{E}(\widetilde{X}^2)$, and $\text{var}(Y) = \text{var}(\widetilde{Y}) = \mathbb{E}(\widetilde{Y}^2)$. Thus

$$\begin{aligned}\text{var}(X+Y) &= \mathbb{E}\,(X+Y-\mu_X-\mu_Y)^2 \\ &= \mathbb{E}\left(\widetilde{X}+\widetilde{Y}\right)^2 \\ &= \mathbb{E}(\widetilde{X}^2) + 2\mathbb{E}(\widetilde{X}\widetilde{Y}) + \mathbb{E}(\widetilde{Y}^2)\end{aligned}$$

From Section 6, the middle term factorizes as

$$(\mathbb{E}\widetilde{X})(\mathbb{E}\widetilde{Y}) = 0 \times 0 = 0$$

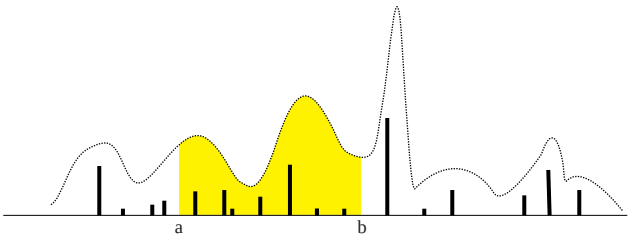
The remaining terms give the sum of the two variances.

8. Continuous distributions

Historically speaking, density functions were first invented only as convenient approximations for discrete distributions. One could approximate the probability that a random variable X would take values in the range $[a, b]$ by calculating a corresponding area under the density function:

$$\mathbb{P}\{a \le X \le b\} \approx \text{ area under the density curve between the lines at } a \text{ and } b$$

The approximation was supposed to hold for all choices of a and b:



Don't stare too carefully at the picture, trying to match up areas with sums of probabilities. It is just a rough sketch. The heights of the little lines are supposed to represent probabilities $\mathbb{P}\{X = x_i\}$ for the values x_i that X can take. The sum of these probabilities for all x_i between a and b gives $\mathbb{P}\{a \le X \le b\}$. The area of the shaded region represents the approximation to this probability.

If the values x_i that X can take are very close together, it is a natural step to ignore the discreteness of the distribution, making a conceptual leap to think of the probability distribution of X as smeared in a continuous fashion along the line. That is, we could treat some random variables as having ***continuous distributions***, with probabilities determined exactly by areas under the graph of a density function. For example, we could think of human heights as being distributed along a continuous range. Why restrict ourselves to any particular degree of discreteness?

In the real world, however, dimensions are always measured to some degree of accuracy determined by the measuring instrument. We could just as logically choose to think of human heights as having a discrete distribution, concentrating in a gritty fashion at values strung out, say, 1 mm apart. That is, we have a choice whether to think of heights as having a continuous distribution or a discrete distribution.

I would advise you to adopt whichever choice is more convenient. If you want to understand formulae for means and variances, think discrete, perhaps on a very fine scale. If you wish to use normal approximations, but you don't want to keep stumbling over caveats about approximation, just say that the random variable has a continuous distribution. In Statistics, you can have your cake and eat it too—sometimes.