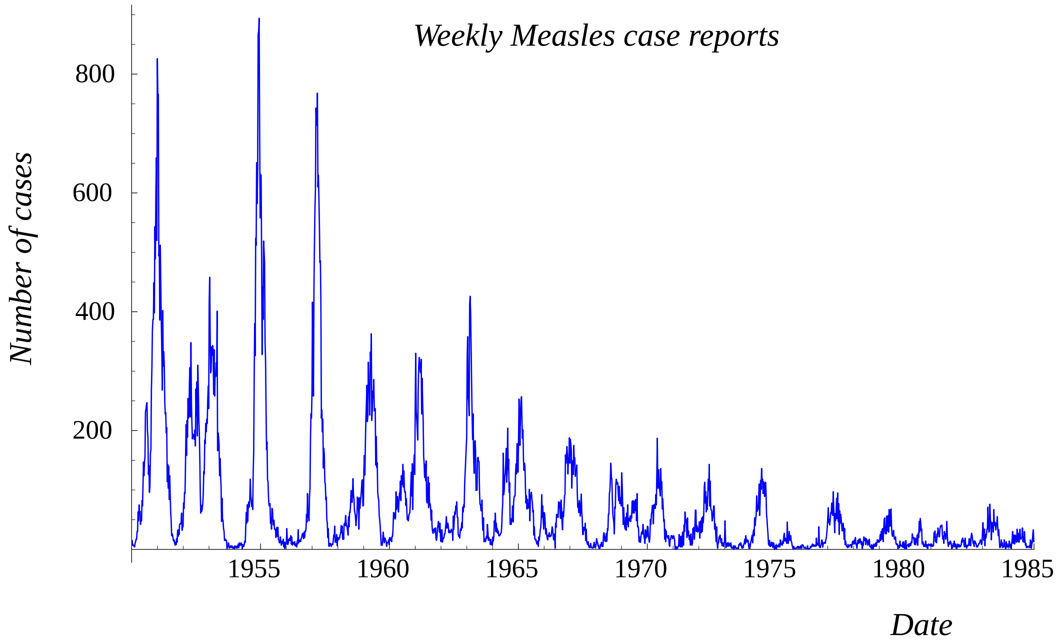


Figure 1.1: Four decades of measles data for Manchester.

1 Describing Data

The aim of the first part of this lecture is to present certain standard ways of exploring, summarizing and presenting data. I'll mainly discuss the definitions and uses of numerical quantities such as the mean, median, mode, range, variance and standard deviation, but I'll also introduce some graphical techniques.

The examples in this lesson will be of two sorts: fictitious data sets made up to illustrate certain pedagogical points and one real sample drawn from a large data set assembled by Dr. Ben Bolker. The latter is a weekly record of reported cases of measles in seven British cities: London, Bristol, Liverpool, Manchester, Newcastle, Birmingham and Sheffield. The data for Manchester are pictured in Figure 1.1.

Example 1.1 (Mancunian Measles around New Year) *Here, in chronological order, are the number of measles cases reported in Manchester during the first week of January for each of the years 1949-1987.*

> *339, 15, 767, 78, 413, 6, 631, 35, 416, 16, 133, 20, 330, 32, 322, 21, 177, 62, 185, 9, 129, 39, 22, 46, 48, 12, 0, 8, 70, 9, 15, 7, 5, 1, 45, 4, 15, 22, 2*

In what follows it will prove helpful to have these numbers sorted into ascending order.

> 0, 1, 2, 4, 5, 6, 7, 8, 9, 9, 12, 15, 15, 15, 16, 20, 21, 22, 22, 32, 35, 39, 45, 46, 48, 62, 70, 78, 129, 133, 177, 185, 322, 330, 339, 413, 416, 631, 767

But even after sorting, these data are not terribly instructive so we might try to pick out a few numbers that help characterise the data.

1.1 Measures of "central tendency"

The first thing one wants to know about a set of data is: "What is a typical value?". So, for example, one might ask "How many cases of measles are reported in the first week of a typical year?" or "How much of the page area of a typical copy of *The Guardian* is devoted to advertising?", "What fraction of people pass their driving tests on the first go?" or "How much does a typical fresher drink in her first week at Uni?" The most common (statistical) answers to these sorts of questions are sketched below and illustrated with the measles data:

Arithmetic Mean: This is probably the most common answer to the question: "What's a typical value?" One computes the

(sum of all values)/(number of values).

Here the result is 115.5 cases. This quantity is often called the *average*, in everyday speech, but formal statistical discussions usually prefer *mean.* There are two other associated conventions: when one is writing about the mean of a whole population (not always a thing that is accessible experimentally) one uses the greek letter 'mu': μ; but when one is writing about the mean of a sample, say, of a list of N observations $\{x_1, x_2, \ldots, x_N\}$, then the mean is written $\bar{x}$ so that the standard formula is:

$$\begin{aligned} \bar{x} &= \frac{\sum_{j=1}^{N} x_j}{N} \\ &= \frac{x_1 + x_2 + \ldots + x_N}{N} \end{aligned}$$

Median: The median is that value which divides the data in half in the sense that 50% of the values are less than or equal to the median and 50% are greater. When, as in the measles data, there is an odd number of measurements the median is the value that falls in the middle of the sorted list; 32 cases in the data above. When there is an even number of measurements the median is the average of the middle two measurements in the sorted list. Thus if the data are {3, 6, 7, 2} the sorted list is {2, 3, 6, 7} and the median is (3 + 6)/2 = 4.5.

Mode: the value that occurs most frequently: here it is 15 cases. Suppose, for example, that once in the grip of incredible boredom you decided to count the number of each colour of M&M in a package; the results might look like:

Colour	Count
Blue	4
Green	5
Brown	3
Red	5
Orange	4

Here the most frequently occurring colours are green and red, so the data set has two modes.

1.2 Displaying and measuring variability

The measures of central tendency sketched above do a reasonable job of capturing the notion of a typical measurement or value, but they give no information about the the "spread" of the data—the way it is distributed around its central tendency. It's not at all difficult to make up two data sets that have the same mean despite being very different: here is a pair invented by Dr. Kamlesh Chauhan from the Optometry department for a course about similar topics. It lists the intraocular pressures (IOPs in mm. Hg above atmospheric pressure) for two groups.

Optometrist IOP	**Engineer IOP**
14	16
15	17
15	15
15	13
15	13
15	10
15	20
16	16
15	10
15	8
14	15
15	19
14	17
16	13
15	23
15	13
15	18
15	12
15	17
16	15

Both columns of this table have the same sum, 300, and the same number of entries, 20, so both have the same arithmetic mean, $\bar{x} = (300/20) = 15$. But the two sets of measurements are really very different. To illustrate this, I will introduce a couple of very useful kinds of plots, the *frequency histogram* and the *cumulative frequency plot.*

1.2.1 Drawing histograms

These plots essentially count the number of times that a particular value occurs in the data set: this number is sometimes called the *frequency* with which the value is observed. For the Optometrist's data above such counting provides a very concise summary: the value 14 mm. Hg occurs 3 times; the value 15 mm. Hg occurs 14 times and the value 16 mm. Hg occurs 3 times. But the data from the Engineers are more spread out and a straightforward count of frequencies-of-measurements isn't really all that helpful. Instead, we might tabulate the number of values that fall in certain small intervals, as is done in the table below.

Interval From	To	Freq. f	Relative Freq. (f/N)	Percentage Freq. $100\ (f/N)$	Cumulative Freq. (%)
7.5	9.5	1	1/20 (0.05)	5.0	1 (5%)
9.5	11.5	2	2/20 (0.1)	10.0	3 (15%)
11.5	13.5	5	5/20 (0.25)	25.0	8 (40%)
13.5	15.5	3	3/20 (0.15)	15.0	11 (55%)
15.5	17.5	5	5/20 (0.25)	25.0	16 (80%)
17.5	19.5	2	2/20 (0.1)	10.0	18 (90%)
19.5	21.5	1	1/20 (0.05)	5.0	19 (95%)
21.5	23.5	1	1/20 (0.05)	5.0	20 (100%)
Totals		$N = 20$	1 (1)	100	

Table 1.1: *A summary of the IOP data from the Engineers. The intervals used here are all the same width and their boundaries have been chosen so that all the data fall* inside *intervals.*

For the moment, concentrate on the first three columns, which are summarised in Figure 1.2 below. The horizontal axis is divided up into the same intervals as described in Table 1.1 and the heights of the bars show the number of measurements that fall into each intervals' range. Such a plot is called a *frequency histogram* and the intervals are often called *bins.*

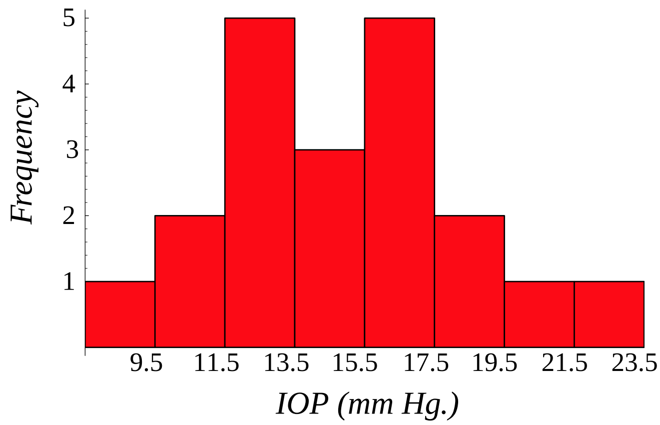


Figure 1.2: *A histogram of the IOP data from the Engineers.*

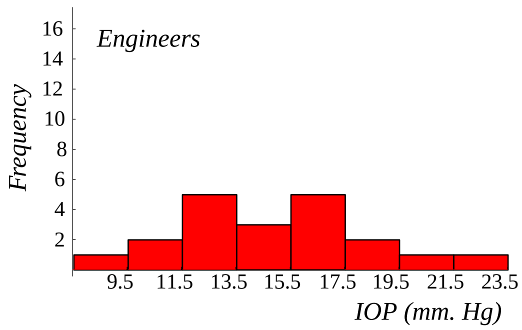
When plotting a histogram there are a few guidelines that make for better-looking and clearer figures:

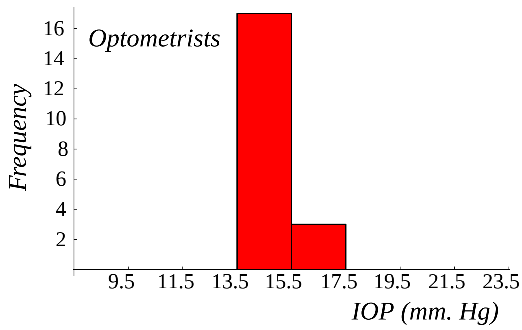
a) The number of intervals should be of order 10-20: the idea is to strike a balance between too many small intervals (most of which will contain just a few measurements) and too few, from which one gets only a poor sense of the spread of the data.

b) Intervals of equal width are convenient both graphically and for calculations.

c) The endpoints of the intervals should have about the same accuracy as the measurements themselves: if one's data consists of whole numbers it makes little sense to define edges of intervals that require three significant digits to express.

The histogram is, all by itself, a powerful tool for exploring and displaying the variation in a data set. Here are histograms for both the Engineers and the Optometrists, both using the same bins as in Table 1.1 and plotted on the same scale. It is immediately clear that although both groups have the same mean IOP, they differ radically from each other.





1.2.2 Cumulative frequency plots

The information in the remaining columns of Table 1.1 is related simply to the frequencies used to plot the histogram: the fourth column, the one titled Relative Frequency, gives the fraction of the total number of data points that fall in each bin; the fifth column just represents the relative frequency as a percentage. For example, the bin that stretches from 11.5 to 13.5 (and so includes all the data with values 12 and 13) accounts for 5 out of 20, or 25% of the observations. The leftmost column in the table is the most useful and it will serve as the foundation for the other plot I want to introduce in this lesson. It shows the cumulative frequency. That is, it shows a running total of the amount of data that has been accounted for as we move down the table. Thus the first bin includes 1 point, so the cumulative frequency for the first bin is 1. The second bin contains 2 points, so its cumulative total of points-accounted-for is $1 + 2 = 3 \ldots$ and so on down the table until, with the last bin, we have accounted for all 20 observations.

But there is another useful way to think of the cumulative frequency: it is the amount of data that lies in a bin *or any of the bins to its left.* This means we can use the cumulative frequency column to characterise values that are "large" or "small" as compared with the bulk of the data. To see how, look at Figure 1.3. The horizontal (x) axis has roughly the same range as the IOP data from the Engineers while the vertical (y) axis is marked in units of percentage. The points on the curve have as their x-coordinate the positions of the centres of the bins from Table 1.1 and have, as their y-coordinate the corresponding cumulative frequency. The curve swoops up from zero (for IOPs lower that the smallest observed value) to 100% for values greater than the largest observed value. The dashed line near the top of the plot shows how to use the cumulative frequency diagram to characterise an improbably large value:

it stretches from the 95% mark over to the cumulative frequency curve; from there it descends to that value of the IOP (20.5 mm Hg., reading from the graph) which is greater than or equal to 95% of all observed values. But if 95% of all values are *smaller* than this value, then only 5% can be greater: we have characterised an improbably large value.

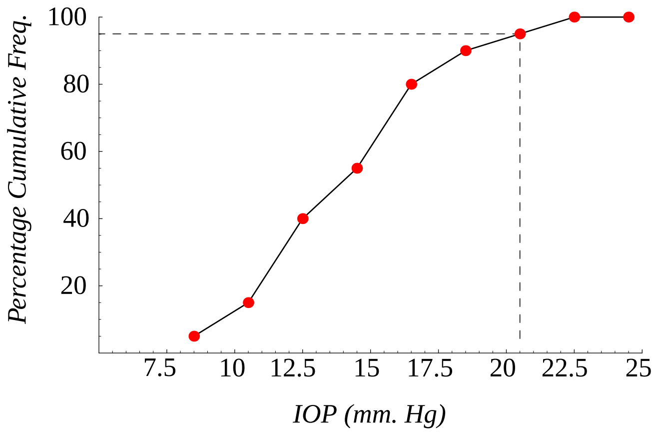


Figure 1.3: *A cumulative frequency plot of the IOP data from the Engineers. The dashed line near the top show to use the cumulative frequency plot to characterise improbably large values.*

1.3 Numerical measures of variation

In the last two sections of the lecture we examined graphical means to explore the distribution of data, but there are a also number of widely-used numerical measures of variability and I will review them briefly below.

Range: the difference between the minimum and the maximum of the data. In the measles data this is 767-0 = 767 cases.

Variance: in words, this is the sum of the squares of all the deviations of the values from their mean divided by (number of values - 1). The variance, around 33293 for the measles data, is often written s^2 and in symbols is given by:

$$\begin{aligned} s^2 &= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots (x_N - \bar{x})^2}{N-1} \\ &= \frac{\sum_{j=1}^{N} (x_j - \bar{x})^2}{N-1} \end{aligned}$$

The reason one has to add up sums-of-squares is that the sum of the ordinary, non-squared deviations from the mean is, because of the definition of the mean, automatically zero.

This expression above is appropriate for analysing a sample (a smaller set of data drawn from some larger population). If we were considering the whole of some finite population we would use slightly different symbols and formulae:

$$\sigma^2 = \frac{\sum_{j=1}^{N}(x_j - \mu)^2}{N}$$

Here are a pair of tables that summarise the computation of the variance for the Optometrists and Engineers above:

Optometrists

IOP mm. Hg	Deviation $(x-\bar{x})$ $(x-15)$	Deviation^2 $(x-\bar{x})^2$ $(x-15)^2$
14	-1	1
15	0	0
15	0	0
15	0	0
15	0	0
15	0	0
15	0	0
16	1	1
15	0	0
15	0	0
14	-1	1
15	0	0
14	-1	1
16	1	1
15	0	0
15	0	0
15	0	0
15	0	0
15	0	0
16	1	1
Totals	0.0	6.0
	$s^2 \approx 0.316$	

Engineers

IOP mm. Hg	Deviation $(x-\bar{x})$ $(x-15)$	Deviation^2 $(x-\bar{x})^2$ $(x-15)^2$
16	1	1
17	2	4
15	0	0
13	-2	4
13	-2	4
10	-5	25
20	5	25
16	1	1
10	-5	25
8	-7	49
15	0	0
19	4	16
17	2	4
13	-2	4
23	8	64
13	-2	4
18	3	9
12	-3	9
17	2	4
15	0	0
Totals	0.0	252.0
	$s^2 \approx 13.3$	

Standard Deviation: the square root of the variance. The main reason to take square roots is to recover a measure of variability that has the same units as the data. So, for example, the IOP data above had units of mm. Hg, as did its mean. The variance thus has units of (mm.)^2 Hg, which doesn't make sense physically. The standard deviation is usually written as s for a sample and one computes it with one of the formula:

$$s = \sqrt{\frac{\sum_{j=1}^{N}(x_j - \bar{x})^2}{N-1}}$$

It is the most common measure of the spread or breadth of a set of data. Many scientific results are reported as $\bar{x} \pm s$. Fore the data above we might write that we had measured IOPs of 15 ± 3.64 mm. Hg in the Engineers and 15 ± 0.562 mm. Hg in the Optometrists.

The mean, median and mode are all natural measures of the central tendency and are useful for describing data that clusters around some particular value. The variance and standard deviation are measures of the way the data are scattered around the central tendency. The mean, variance and standard deviation are the most commonly used measures and are particularly useful when discussing normally-distributed (more on this soon) data.

2 Discrete probability

The aim of this part of the lecture is to learn how to think about events whose outcome is influenced by chance. Ultimately we will be interested in questions such as "Given that I've just measured a difference between a sample of second-year Engineers students and the third years, what is the chance that the two groups really *are* different?" Answers to this sort of question come from the subject of *hypothesis testing*, which is the main theme of the statistical part of the course. Today we will begin with some simpler problems about the sort of probability needed to analyse coin-tossing and card games, then show how these simple rules lay the foundations for hypothesis testing.

2.1 Cards and coins

First we will consider various problems about coin tossing, card dealing and the rolling of dice[1]—perhaps the simplest sorts of random events. Let's consider the prototypical example, a *fair coin*. I'll imagine tossing it in the air and letting it fall flat so that it shows either Heads or Tails. I'll then say that each of these outcomes has probability 0.5 and will mean that if I tossed the coin a great many times, I would expect Heads to come up about half the time. Since Tails is the only other possible outcome, it too must have a long term average of 0.5. This example illustrates many of the basic notions about probability.

- Probabilities are numbers p with $0 \leq p \leq 1$. To say an event has probability zero means that we regard it as impossible; for example, rolling a 7 on an ordinary, six-sided die has probability zero. To say that an event has probability one means that it happens with certainty, *e.g.* rolling a number in the range 1-6 on a die.

- It is often helpful to make an exhaustive list of all the distinct possible outcomes of a probabilistic experiment: { Heads, Tails } in the example above. Since one of these two events must occur whenever we toss the coin, the probabilities of all distinct possible outcomes should add up to one.

- Following the remark above, suppose an event A has probability $P(A)$. Then the probability that A does *not* occur must then be $(1 - P(A))$, for the pair of events {A, (not A)} is the just sort of exhaustive list of outcomes mentioned above.

2.1.1 Mutually exclusive events

The last item above suggests that we should be more scrupulous about defining a probabilistic event. Let's consider a slightly richer example, drawing playing cards from a deck. We might ask: "What is the probability of drawing an ace?". As

[1]The strong flavour of gambling in these examples is not an accident: Blaise Pascal and Pierre de Fermat laid the foundations of mathematical probability during a five letter exchange over the summer of 1654. Much of this correspondence concerned games with dice.

there are 52 cards it seems natural to divide the possible outcomes into 52 distinct, mutually exclusive[2] events—one for each card.

But this division doesn't help us to answer the question about aces; we need a way to compute the probability of *composite events*, those made up by grouping together two or more of our basic events. In this case it's easy enough to see that there are 4 aces in the deck, so the probability of drawing one of them is:

$$\frac{\text{Number of aces}}{\text{Number of cards}} = \frac{4}{52} = \frac{1}{13}.$$

As this example suggests, the way to compute the probability of a collection of mutually exclusive events is to add them up:

If A & B are mutually exclusive events, $P(A \text{ or } B) = P(A) + P(B)$.

2.1.2 Independent events

Now consider flipping a pair of coins—say, a penny and a 50 pence coin—one after the other. If we use lowercase letters, h & t, to indicate the result for the penny and uppercase letters, H & T, for the 50p, then the complete list of possible outcomes is {hH, hT, tH, tT}. As we have no reason to think Heads or Tails more likely for either coin, we should imagine all four of these outcomes to be equally likely, so just by counting it is clear that

- P(Heads on penny) = (2/4) = 0.5
- P(Heads on 50p) = (2/4) = 0.5
- P(Heads on both) = (1/4) = 0.25

There is a new relationship between events here and a new rule for combining their probabilities. Getting Heads on the penny does not affect the chance of getting Heads on the 50p, so the two events are said to be *independent* of each other. If we consider the combined experiment of tossing both coins we end up with a new list of outcomes that is a kind of "product" of the outcomes of the two tosses. And the probabilities are products too. The rule is:

*If A and B are independent events, P(A & B) = P(A) * P(B).*

2.1.3 More about combining events

Finally, there is a rule for combining the probabilities of events that are not mutually exclusive (*i.e.* those for which $P(A \,\&\, B) \neq 0$).

Generally, P(A or B) = P(A) + P(B) - P(A & B).

To see how this last rule works, consider rolling two dice, one six-sided and one four-sided[3] Let us take as our events

[2] Two events are said to be *mutually exclusive* if they cannot both occur at the same time.

[3] The fantasy role-playing game community makes use of all sorts of weird dice: 4, 6, 8, 12 and 20 sided dice are fairly common and I have have heard of 100-sided dice that look a bit like a golf-balls, though I've never seen one.

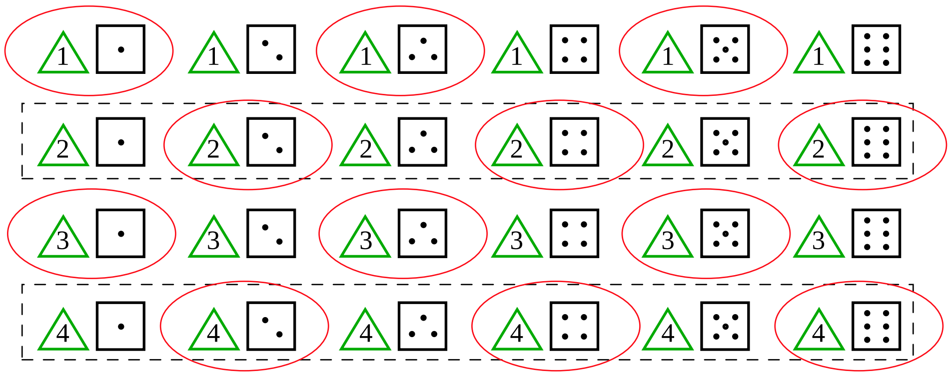


Figure 2.1: *The 24 possible outcomes of a game of chance in which the player rolls one four-sided die and one six-sided die.*

A The four-sided die comes up an even number;

B The sum of the two rolls is an even number.

Using the rule above, along with the facts that $P(A) = 1/2$, $P(B) = 1/2$ and $P(A\&B) = 1/4$, we find $P(A\text{or}B) = (1/2) + (1/2) - (1/4) = (3/4)$. To make this a little clearer, consider the diagram above. It shows an exhaustive list of the 24 possible ways that the rolls can turn out: by assumption, each of these is equally likely and so has probability (1/24). Those that correspond to event A (first roll is even) are enclosed in dashed horizontal boxes while those that correspond to event B are encircled by ovals. Counting things up, one finds half (12) of the possible outcomes in the A-boxes, half in the B-boxes and three quarters (18) in one or both of the A and the B boxes. There are 6 entries that appear in both the A and B boxes. Making the probability computation explicit

$$
\begin{aligned}
P(A\text{or}B) &= P(A) + P(B) - P(A \text{ and } B) \\
&= (12/24) + (12/24) - (6/24) \\
&= (1/2) + (1/2) - (1/4) \\
&= (3/4)
\end{aligned}
$$

2.1.4 Conditional probability

There is one more basic probabilistic concept that proves useful: *conditional probability.* The idea is to have a concise notation for the probability of one event, given that another has occurred. For example:

Example 2.1 *In rolling a six-sided die, what is the probability of getting a two, given that the result is an even number?*
There are three possible even numbers, { 2, 4, 6 } and only one of them is a 2, so, by direct counting, the probability is (1/3).

Generally one writes conditional probabilities as $P(A|B)$ and read them as "the probability of event A given event B" or, more briefly, "the probability of A given B". Less frivolous and more useful applications of the idea might include:

- $P($ It will rain tomorrow | it is raining now $)$
- $P($ It will rain tomorrow | one is in Manchester $)$
- $P($ A woman gets breast cancer | her mother and sister did $)$

The simplest rule about conditional probabilities lies behind such intuitively reasonable statements as:

$$P(\text{ rain } | \text{ Manchester }) + P(\text{ no rain } | \text{ Manchester }) = 1.$$

More formally, the rule is that if one has an exhaustive list of mutually exclusive events then their conditional probabilities add up to one.

Sometimes one needs to pass from conditional probabilities back to non-conditional ones. The main tool one needs is the formula:

$$P(A\&B) = P(A|B) \times P(B) \tag{2.1}$$

To see how to use it, suppose that we are studying the epidemiology of lung cancer and are interested in the influence of smoking. We might divide our subjects into three categories—say, heavy smokers (40 or more cigarettes daily), smokers (up to 39 cigarettes per day) and non-smokers—and then examine the risk for each, measuring, for example, $P($ subject develops lung cancer | smokes $\geq$ 40 cigs. daily $)$. But we might also be interested in $P($ subject develops lung cancer $)$ and we could compute it using:

$$\begin{aligned} &P(\text{ subject develops lung cancer }) \\ &\quad = P(\text{ [cancer \& heavy smoker] } \textit{or} \text{ [cancer \& smoker] } \textit{or} \text{ [cancer \& non-smoker] }) \\ &\quad = P(\text{ cancer \& heavy smoker }) + P(\text{ cancer \& smoker }) + \\ &\qquad P(\text{ cancer \& non-smoker }) \end{aligned}$$

To get this we have used only the rule about combining mutually exclusive events: our three categories of smoker are mutually exclusive and so the events "cancer & heavy smoker", "cancer & smoker" and "cancer & non-smoker" are mutually exclusive and so their probabilities just add, as described in Section 2.1.1. Now if we use Equation (2.1) above we obtain

$$\begin{aligned} &P(\text{ subject develops lung cancer }) = \\ &\qquad P(\text{ cancer} \mid \text{smokes} \geq 40 \text{ cigs. daily }) \times P(\text{ smokes} \geq 40 \text{ cigs. daily }) + \\ &\qquad P(\text{ cancer} \mid \text{smokes 1 to 39 daily }) \times P(\text{ smokes 1 to 39 daily }) + \\ &\qquad P(\text{ cancer} \mid \text{doesn't smoke }) \times P(\text{ doesn't smoke }) \end{aligned}$$

Look again at Equation (2.1). Notice that on the left hand side of the expression the events A and B play essentially the same role: $A\&B$ means the same thing as $B\&A$. But on the right side things appear to be different: $P(A|B)$ is not generally the same as $P(B|A)$ so[4] it seems that A and B aren't playing equivalent roles on the right. But they must be, as the following calculations show:

$$\begin{aligned} P(A|B) \times P(B) &= P(A\&B) \\ &= P(B\&A) \\ &= P(B|A) \times P(A) \\ \text{which means} \quad P(A|B) \times P(B) &= P(B|A) \times P(A) \end{aligned}$$

This final expression is sometimes known as *Bayes Theorem* and is one of the most useful formulae in the theory of probability.

Example 2.2
Out of a group of 50 patients being treated for a severe allergy, 10 are chosen at random to receive a new dietary treatment as opposed to the more usual drug therapy which the remaining 40 patients receive. Suppose it is known (from other studies) that the probability of a cure with the standard treatment is 0.6, while the probability of a cure from the new treatment is 0.9. Some time later, one of the 50 patients returns to thank the staff for her complete recovery. What is the probability that she was given the new treatment?
In the language of conditional probability, we are asked to find

$$P(\text{ new treatment} \mid \text{cure }).$$

The problem tells us that the patients receiving the new treatment were chosen at random, so we have

$$\begin{aligned} P(\text{ new treatment }) &= (10/50) = 0.2 \\ P(\text{ standard treatment }) &= (40/50) = 0.8 \end{aligned}$$

and we are also told

$$\begin{aligned} P(\text{ cure} \mid \text{new }) &= 0.9 \\ P(\text{ cure} \mid \text{std }) &= 0.6 \end{aligned}$$

[4] To see that they aren't the same, just substitute in some words: $P(\text{ rain} \mid \text{Manchester })$—the probability that it is raining given that one is in Manchester—is a biggish number and very different from $P(\text{ Manchester} \mid \text{rain })$, which is the probability that, given that it is raining, one can conclude that one is in Manchester.

where I have abbreviated the phrases "new treatment" and "standard treatment" as "new" and "std". The way to proceed now is to use Bayes Theorem, Equation (2.2). It says that

$$\begin{aligned} P(\text{ new } \mid \text{ cure }) \times P(\text{ cure }) &= P(\text{ cure } \mid \text{ new }) \times P(\text{ new }) \\ \text{or} \qquad P(\text{ new } \mid \text{ cure }) &= \frac{P(\text{ cure } \mid \text{ new }) \times P(\text{ new })}{P(\text{ cure })} \end{aligned}$$

The bottom line is something of an advance in that it has the expression we want on its left side and, mostly, has probabilities that we know on the right. The only exception is $P(\text{ cure })$. But we can evaluate that using conditional probabilities:

$$\begin{aligned} P(\text{ cure }) &= P(\text{ cure } \mid \text{new}) \times P(\text{ new }) \;+\; P(\text{ cure } \mid \text{ std }) \times P(\text{ std }) \\ &= (0.9) \times (0.2) \;+\; (0.6) \times (0.8) \\ &= 0.18 + 0.48 \\ &= 0.66 \end{aligned}$$

Putting this back into our expression above:

$$\begin{aligned} P(\text{ new } \mid \text{ cure }) &= \frac{P(\text{ cure } \mid \text{ new }) \times P(\text{ new })}{P(\text{ cure })} \\ &= ((0.9) \times (0.2))/0.66 \\ &= (0.18/0.66) \\ &\approx 0.273 \end{aligned}$$