Excursions in Computing Science: Book 9c. Heat: Histograms and Gases Part I. Histograms, etc.

T. H. Merrett*
McGill University, Montreal, Canada

August 1, 2021

I. Prefatory Notes

1. Histograms. For two months of a remarkable Montreal winter the following temperatures were recorded.

January

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	−1	0	1	1	1	0	−1	−1	−1	1	−1	1	−1	1	0

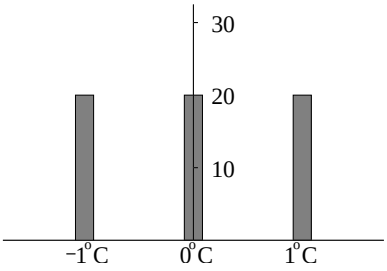
17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
−1	−1	0	0	0	1	0	0	1	−1	1	1	0	0	−1

February

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
−1	0	1	1	−1	0	0	−1	0	−1	1	−1	0	−1	0

16	17	18	19	20	21	22	23	24	25	26	27	28	29
1	0	1	1	0	−1	−1	1	−1	1	0	1	−1	0

It is probably hard to see the pattern in this, so we abstract to a histogram.



That is, over the 60 days, the temperature -1^o Celcius was noted 20 times, the temperature 0^o Celcius was noted 20 times, and so was the temperature 1^o Celcius.

A *histogram* records a set of values and the number of times that each of the values occurs.

All abstractions throw away irrelevant data and considerations and so lose information through a process of judgement as to what is pertinent. Histograms throw away so much data that I am tempted to be redundant and call them "lossy abstractions".

We revealed a pattern in the data by taking the histogram but we thereby ditched all the sequence information as irrelevant. Other abstractions might be made, retaining the sequencing, but much can be done with histograms alone and our purpose in Book 9c is to see how far we can get.

Note that we would have made the same histogram if the data had been "ordered" with the first 20 days at -1^oC, the next 20 at 0^oC and the last 20 days at 1^oC. It doesn't matter.

But what do we mean by "order"? Here is the shuffling trick. A card player shows you a perfectly sorted deck of cards—ace, 2, 3, .. king of hearts, then ace, 2, 3, .. king of spades, then diamonds in order then finally clubs. Hey shuffles the deck thoroughly, and then, saying "watch this", appears to shuffle it once more. When hey hands you the final deck in triumph, you expect to find it miraculously back in sorted order. But it is not. It appears just as random as it was after the thorough shuffle. The cardplayer's point is this: it is almost impossible for anyone to duplicate what hey did and produce a deck in the *exact same order* as hey just produced.

A similar story concerns a golfer who has just driven the ball off the 4th tee onto a tuft of grass 7/8 of the way to the 4th hole. The probability that the ball landed on that particular tuft of grass is minutely small, so small that it is almost impossible for another golfer to land another ball on the *exact same* tuft of grass.

The point of these two stories is that what is meant by "order" must be decided in advance. Maybe it is alphabetical order or a strict ascending sequence of temperatures or putting the ball into the 4th hole instead of on some tuft of grass which we did not notice until after the ball had landed. If an objective is not set, what happens happens. Nature does not know from alphabetical order.

A suggestive characterization of order is the more ordered a system was the easier it is later to tell if it has been messed up. Disturbing a rack of six billiard balls in a triangle, by hitting it with the cue ball, is much more conspicuous than subsequent disturbances by the cue ball.

2. Histogram arithmetic. Histograms can also be represented as numbers only.

occurrences	20	20	20
value	-1^oC	0^oC	1^oC

This is a data aggregate which might remind us of matrices. As with matrices, we can do arithmetic on histograms.

For instance, if we had two histograms, we could add them.

Let's change examples and suppose that Joe and Sue are doing business with the same three other people. The histogram for Joe

# people	1	1	1	total 3
owes Joe	-1\$	0\$	1\$	

means that one person owes Joe -1\$ (i.e., *Joe* owes the dollar), one person owes Joe 1\$, and Joe and the third person are all settled up (0\$ entry).

In this example, the abstraction that got rid of the sequencing in the winter temperature example becomes the abstraction that we do not name the people Joe and Sue are dealing with or care which person owes what.

The histogram for Sue

# people	1	1	1	total 3
owes Sue	−1$	0$	1$	

has the corresponding meaning.

As with the temperature example, we have not recorded, and don't know, who these people are, except that they are the same three in both cases: we don't know which one owes or is owed how much in each case. The same person could owe both Joe and Sue, or could owe Joe but be owed by Sue or vice versa, and so on for all possible combinations.

What are the combinations?

Sue	−1$	0$	1$	
Joe				
−1$	−2$	−1$	0$	owed
0$	−1$	0$	1$	to Joe
1$	0$	1$	2$	or Sue

How many ways could we arrange people's debts and credits, given that we do not know how these are actually distributed?

Clearly there is only one way one person could owe 2$: hey must owe both Joe and Sue. There is only one 2$ entry in the table. Similarly there is only one −2$ entry, and there is only one way a person can be owed 2$.

But a person can owe 1$ in two ways, either to Joe or to Sue. Similarly a person can be owed 1$ in two ways.

Finally, a person can be free of both debt and credit (in total) in three ways: by owing and being owed nothing to and by either Sue or Joe; by owing Sue and being owed by Joe; or by owing Joe and being owed by Sue.

We can summarize all this in a new histogram.

# ways	1	2	3	2	1	total 9
owes Joe or Sue	−2$	−1$	0$	1$	2$	

This is the sum of the Joe and Sue histograms because the values (lower row) are the sums of the values of the two input histograms.

If Jan were also to join Sue and Joe as an entrepreneur involving the same three (other) people and with histogram

# people	1	1	1	total 3
owes Jan	−1$	0$	1$	

then the sum of all three would come from

		1	2	3	2	1
	Joe or Sue	−2$	−1$	0$	1$	2$
	Jan					
1	−1$	−3$	−2$	−1$	0$	1$
1	0$	−2$	−1$	0$	1$	2$
1	1$	−1$	0$	1$	2$	3$

with the same meaning.

This time, I've written the # ways/# people rows in addition to the value rows, because they must be used in finding the new histogram.

For instance, since there are 2 ways Sue or Joe can be owed 1$, the number of ways Sue, Joe or Jan can be owed 2$ is

a) Jan is owed 1$ (1 way) and Sue or Joe are owed 1$ (2 ways)—2 ways together; or
b) Jan is owed 0$ (1 way) and Sue or Joe are owed 2$ (1 way)—1 way together.

This is a total of 3 ways. Doing this for each possible debt/credit we get the histogram summing Joe, Sue and Jan.

# ways	1	3	6	7	6	3	1	total 27
owes Joe, Sue, Jan	−3$	−2$	−1$	0$	1$	2$	3$	

Do you see how this histogram can quickly be built up from the table combining Jan with Joe-and-Sue?

We'll add a fourth entrepreneur, Pat, in two different ways. The second way is important because it shows that we *multiply* the # ways entries.

a)

		1	3	6	7	6	3	1
Joe, Sue, Jan		−3$	−2$	−1$	0$	1$	2$	3$
	Pat							
1	−1$	−4$	−3$	−2$	−1$	0$	1$	2$
1	0$	−3$	−2$	−1$	0$	1$	2$	3$
1	1$	−2$	−1$	0$	1$	2$	3$	4$

giving (check it out!) for a total of 81

# ways	1	4	10	16	19	16	10	4	1
owes Joe, Sue, Jan, Pat	−4$	−3$	−2$	−1$	0$	1$	2$	3$	4$

b)

		1	2	3	2	1
Joe or Sue		−2$	−1$	0$	1$	2$
Jan or Pat						
1	−2$	−4$	−3$	−2$	−1$	0$
2	−1$	−3$	−2$	−1$	0$	1$
3	0$	−2$	−1$	0$	1$	2$
2	1$	−1$	0$	1$	2$	3$
1	2$	0$	1$	2$	3$	4$

*	1	2	3	2	1
1	1	2	3	2	1
2	2	4	6	4	2
3	3	6	9	6	3
2	2	4	6	4	2
1	1	2	3	2	1

giving exactly the same histogram (check it out!). Note that we *multiply* the # ways entries.

So we know how to add histograms. What happens when we subtract histograms is surprising. Let's subtract the Jan histogram from the Joe, Sue histogram.

		1	2	3	2	1
Joe or Sue		−2$	−1$	0$	1$	2$
	Jan					
1	−1$	−1$	0$	1$	2$	3$
1	0$	−2$	−1$	0$	1$	2$
1	1$	−3$	−2$	−1$	0$	1$

*	1	2	3	2	1
1	1	2	3	2	1
1	1	2	3	2	1
1	1	2	3	2	1

and the *differences* histogram is

# ways	1	3	6	7	6	3	1	total 27
owes Joe+Sue−Jan	−3$	−2$	−1$	0$	1$	2$	3$	

The meaning of "owes Joe+Sue−Jan" is not clear, so we must consider this example formally rather than try to interpret it.

Stretching meaning even further, we can find product histograms.

		1	1	1
Sue		−1$	0$	1$
	Joe			
1	−1$	1$	0$	−1$
1	0$	0$	0$	0$
1	1$	−1$	0$	1$

# ways	2	5	2	total 9
"owes Joe×Sue"	−1$	0$	1$	

Quotient histograms, quite apart from what they may mean, often cannot be calculated. Input histograms with values $-1, 1$ give an example which will work.

Here is the "maximum" histogram, giving the occurrences of the maximum amount owed Joe or Sue.

		1	1	1
Sue		−1$	0$	1$
	Joe			
1	−1$	−1$	0$	1$
1	0$	0$	0$	1$
1	1$	1$	1$	1$

# ways	1	3	5	total 9
"owes Joe **max** Sue"	−1$	0$	1$	

3. Distributions and densities. A *distribution* is a histogram normalized so that the sum of its # ways is 1.

Returning to the temperature example

occurrences	20	20	20
value	−1°C	0°C	1°C

the corresponding distribution is

frequencies	1/3	1/3	1/3
value	−1°C	0°C	1°C

The reason for normalizing is that it saves some arithmetical hassle. For example, for that same Montreal 60-day winter, someone with a more accurate thermometer might have recorded

occurrences	12	12	12	12	12
value	−1°C	−0.5°C	0°C	0.5°C	1°C

i.e.,

frequencies	1/5	1/5	1/5	1/5	1/5
value	−1°C	−0.5°C	0°C	0.5°C	1°C

And someone really fussy might have found the 60 days divided as follows

occurrences	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
value	−1	−.86	−.71	−.57	−.43	−.29	−.14	0	.14	.29	.43	.57	.71	.86	1

which gives a frequency of 1/15 at each value of the temperature.

Now we see that all these readings are saying that the temperature was uniformly distributed between -1°C and 1°C during those 60 days. Since there are an integer number of measurements, 60, the temperature resolution must be some divisor of 60 to show the uniformity. But this restriction is arbitrary and has nothing to do with the uniformity of the distribution.

For instance, someone with a thermometer which can resolve only 9 different temperatures in this range cannot get the same number of occurrences for each temperature value from the 60 readings. Hey could get, instead, say

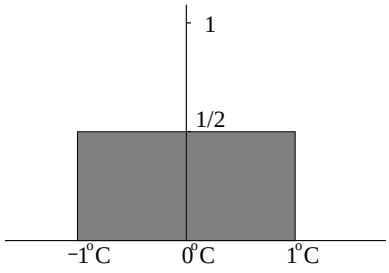
occurrences	7	7	6	7	7	6	7	7	6
value	-1	$-.75$	$-.5$	$-.25$	0	.25	.5	.75	1

These might indeed be the readings, but if that observer wished to describe the distribution as uniform, it is much easier to say that each of the 9 measurements had a frequency of 1/9.

So histograms are frequently normalized into distributions.

A further trick to make calculating easier, even though it departs still more from the measuremnts actually made, is to use continuous math,

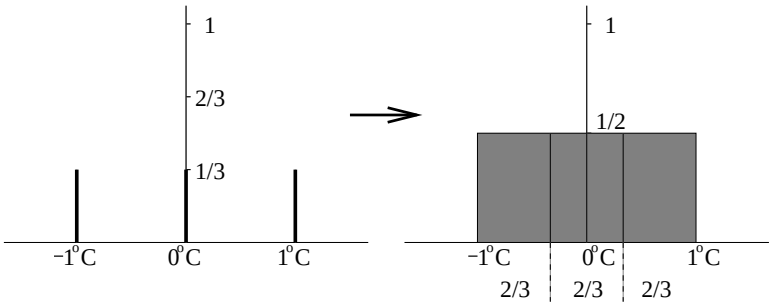
The uniform distribution becomes a uniform *density*



whose *area* over the range is kept = 1.

The only tricky aspect is converting from the sum of the discrete distribution to the area of the continuous density. Two examples show how to do this.

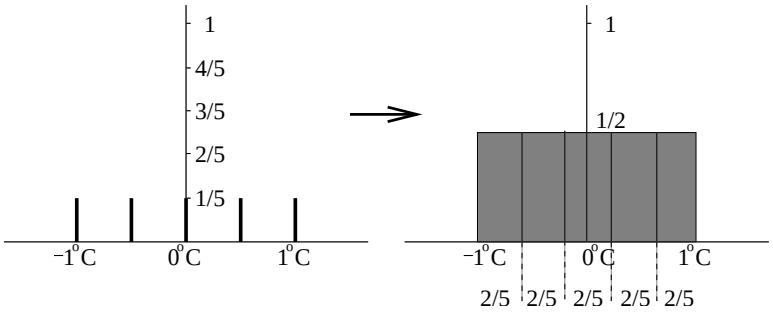
Converting from three measurements to continuous density



the discrete sum is $1/3 + 1/3 + 1/3 = 1$.

The area is $1/2 \times 2/3 + 1/2 \times 2/3 + 1/2 \times 2/3 = 1$, where 2/3 is the width of each of the three equal divisions of the range from -1 to 1 of the values. We can call this breadth Δv and here $\Delta v = 2/3$. So the area is $h\Delta v + h\Delta v + h\Delta v$ where the height $h = 1/2$.

Converting from five measurements to continuous density



the discrete sum is $1/5 + 1/5 + 1/5 + 1/5 + 1/5 = 1$.

The area is $1/2 \times 2/5 + 1/2 \times 2/5 + 1/2 \times 2/5 + 1/2 \times 2/5 + 1/2 \times 2/5 = 1$ where $\Delta v = 2/5$ this time.

Densities and continuous math are not particularly helpful for uniform distributions, which we have been considering in this Note. But a nonuniform distribution is sometimes more easily described as a continuous density, $d(v)$, whose area is approximated better and better as the sum

$$\sum_{v_i=-1}^{1} \Delta v$$

is refined by taking ever smaller Δv and ever more frequent v_i. (The step size is Δv: I haven't shown all the details in the way I've written the sum.)

We will sometimes not use the word "density" but let "distribution" stand for both the discrete and the continuous versions, when no confusion is likely to arise.

4. Aggregates: the moments of distributions. A whole distribution is a complicated thing to remember or to work with (unless it has a simple representation as a mathematical expression, such as the density $d(v) = 1/2$ for a uniform distribution on -1 to 1).

We can extract some especially significant *aggregate* values from a distribution, such as the *average* value, giving its centre, or the *variance* in its values, giving its spread, etc.

In fact, we can extract any number of aggregates we wish, giving us an alternative representation of the distribution, just as the Fourier transform in Week 9 gave us an alternative representation of a function $f(x)$.

These aggregates are based on the *moments* of the distribution. The first few should give the idea.

The *0th moment* is just the sum. For a distribution, this is always 1, by definition, and that's an important justification for normalizing the histogram. For a density, the 0th moment is the area, which is also 1 by definition.

moment	*histogram*	*distribution*	*density*
0	$\sum h_i$	$\sum d_i = 1$	area $d(v) = 1$

The *1st moment* is the average (or *mean*) and is defined

moment	*histogram*	*distribution*	*density*
1	$\frac{\sum v_i h_i}{\sum h_i}$	$\frac{\sum v_i d_i}{\sum d_i} = \sum v_i d_i$	$\frac{\text{area } vd(v)}{\text{area } d(v)} = \text{area } vd(v)$

Let's calculate it for two uniform distributions.
a) The first moment is zero for any distribution symmetric about the origin. Obviously the origin

is thus the average.

$$\begin{array}{ccc} 20 & 20 & 20 \\ \hline -1 & 0 & 1 \end{array} \qquad \frac{\sum v_i h_i}{\sum h_i} = \frac{-1 \times 20 + 0 \times 20 + 1 \times 20}{20 + 20 + 20} = 0$$

$$\begin{array}{ccc} 1/3 & 1/3 & 1/3 \\ \hline -1 & 0 & 1 \end{array} \qquad \sum v_i d_i = -1 \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = 0$$

$$d(v) = 1/2 \qquad \text{area } vd(v) \,|_{-1}^{1} = \text{antislope} \frac{v}{2} \,|_{-1}^{1} = \frac{v^2}{4} \,|_{-1}^{1} = 0$$

b) For the uniform distribution from 1 to 3 the average is 2 in all versions.

$$\begin{array}{ccc} 20 & 20 & 20 \\ \hline 1 & 2 & 3 \end{array} \qquad \frac{\sum v_i h_i}{\sum h_i} = \frac{1 \times 20 + 2 \times 20 + 3 \times 20}{20 + 20 + 20} = 2$$

$$\begin{array}{ccc} 1/3 & 1/3 & 1/3 \\ \hline 1 & 2 & 3 \end{array} \qquad \sum v_i d_i = 1 \times \frac{1}{3} + 2 \times \frac{1}{3} + 3 \times \frac{1}{3} = 2$$

$$d(v) = 1/2 \qquad \text{area } vd(v) \,|_{1}^{3} = \text{antislope} \frac{v}{2} \,|_{1}^{3} = \frac{v^2}{4} \,|_{1}^{3} = 2$$

Note that we have used the Fundamental Theorem of Calculus (Excursion "Calculus" of Week 12) to transfer from the area under a curve to the antislope of the curve evaluated at the endpoints. We will be using quite a lot of calculus in discussing densities, so a review of that Excursion is suggested.

The *2nd moment* is the obvious next step.

moment	*histogram*	*distribution*	*density*
2	$\frac{\sum v_i^2 h_i}{\sum h_i}$	$\sum v_i^2 d_i$	area $v^2 d(v)$

We calculate it for the same two uniform distributions.

a) Symmetric

$$\begin{array}{ccc} 20 & 20 & 20 \\ \hline -1 & 0 & 1 \end{array} \qquad \frac{\sum v_i^2 h_i}{\sum h_i} = \frac{1 \times 20 + 0 \times 20 + 1 \times 20}{20 + 20 + 20} = \frac{2}{3}$$

$$\begin{array}{ccc} 1/3 & 1/3 & 1/3 \\ \hline -1 & 0 & 1 \end{array} \qquad \sum v_i^2 d_i = 1 \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{2}{3}$$

$$d(v) = 1/2 \qquad \text{area } v^2 d(v) \,|_{-1}^{1} = \text{antislope} \frac{v^2}{2} \,|_{-1}^{1} = \frac{v^3}{6} \,|_{-1}^{1} = \frac{1}{3}$$

b) Unsymmetric

$$\begin{array}{ccc} 20 & 20 & 20 \\ \hline 1 & 2 & 3 \end{array} \qquad \frac{\sum v_i^2 h_i}{\sum h_i} = \frac{1 \times 20 + 4 \times 20 + 9 \times 20}{20 + 20 + 20} = \frac{14}{3}$$

$$\begin{array}{ccc} 1/3 & 1/3 & 1/3 \\ \hline 1 & 2 & 3 \end{array} \qquad \sum v_i^2 d_i = 1 \times \frac{1}{3} + 4 \times \frac{1}{3} + 9 \times \frac{1}{3} = \frac{14}{3}$$

$$d(v) = 1/2 \qquad \text{area } v^2 d(v) \,|_{1}^{3} = \text{antislope} \frac{v^2}{2} \,|_{1}^{3} = \frac{v^3}{6} \,|_{1}^{3} = \frac{13}{3}$$

The discrepancies arise in the continuous calculation because the discrete sum is only approximated by the area under $v^2 d(v)$, and vice versa. If we were to take 5 values between -1 and 1 in the sum

in (a), we get 1/2. Nine values gives 5/12, getting closer to 1/3. And so on.

The second moment gives the variance only when the mean is zero, as in (a). In general the *variance* is

$$\begin{aligned}\mu_2 = \sum (v_i - \mu)^2 d_i &= \sum v_i^2 d_i - \sum 2\mu v_i d_i + \sum \mu^2 d_i \\ &= \sum v_i^2 d_i - \mu^2\end{aligned}$$

where μ is the mean (1st moment).

So for (b) the variances are $14/3 - 4 = 2/3$ and $13/3 - 4 = 1/3$ as for (a).

The *standard deviation* is the square root of the variance, and we shall see its significance when we discuss the normal distribution. The standard deviation is usually called σ, so the variance is $\sigma^2 = \mu_2$.

The *3rd moment* gives the *skewness* of the distribution.

moment	*histogram*	*distribution*	*density*
3	$\frac{\sum v_i^3 h_i}{\sum h_i}$	$\sum v_i^3 d_i$	area $v^3 d(v)$

a) Symmetric

20	20	20
−1	0	1

1/3	1/3	1/3
−1	0	1

$$d(v) = 1/2$$

$$\frac{\sum v_i^3 h_i}{\sum h_i} = \frac{-1 \times 20 + 0 \times 20 + 1 \times 20}{20 + 20 + 20} = 0$$

$$\sum v_i^3 d_i = -1 \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = 0$$

$$\text{area } v^3 d(v) \;|_{-1}^{1} = \text{antislope} \frac{v^3}{2} \;|_{-1}^{1} = \frac{v^4}{8} \;|_{-1}^{1} = 0$$

b) Unsymmetric

20	20	20
1	2	3

1/3	1/3	1/3
1	2	3

$$d(v) = 1/2$$

$$\frac{\sum v_i^3 h_i}{\sum h_i} = \frac{1 \times 20 + 8 \times 20 + 27 \times 20}{20 + 20 + 20} = 12$$

$$\sum v_i^3 d_i = 1 \times \frac{1}{3} + 8 \times \frac{1}{3} + 27 \times \frac{1}{3} = 12$$

$$\text{area } v^3 d(v) \;|_{1}^{3} = \text{antislope} \frac{v^3}{2} \;|_{1}^{3} = \frac{v^4}{8} \;|_{1}^{3} = 10$$

We see that third moments are zero for symmetric distributions. We recall the first moment was zero for these also. In fact all odd moments of these distributions will be zero.

Once again, if the mean, μ, is not zero, the result of interest is the moment *about the mean.*

$$\begin{aligned}\mu_3 = \sum (v_i - \mu)^3 d_i &= \sum v_i^3 d_i - 3\mu \sum v_i^2 d_i + 3\mu^2 \sum v_i d_i - \mu^3 \sum d_i \\ &= \sum v_i^3 d_i - 3\mu(\sigma^2 + \mu^2) + 3\mu^3 - \mu^3 \\ &= \sum v_i^3 d_i - 3\mu\sigma^2 - \mu^3\end{aligned}$$

and similarly for the density.

For case (b) these are all 0, as for case (a).

The *coefficient of skewness* is μ_3/σ^3.

The *4th moment* tells us how peaked or flat the distribution is around its centre.

moment	*histogram*	*distribution*	*density*
4	$\frac{\sum v_i^4 h_i}{\sum h_i}$	$\sum v_i^4 d_i$	area $v^4 d(v)$

a) Symmetric

20	20	20
−1	0	1

1/3	1/3	1/3
−1	0	1

$$\frac{\sum v_i^4 h_i}{\sum h_i} = \frac{1 \times 20 + 0 \times 20 + 1 \times 20}{20 + 20 + 20} = \frac{2}{3}$$

$$\sum v_i^4 d_i = 1 \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{2}{3}$$

$$d(v) = 1/2$$

$$\text{area } v^4 d(v) \; |_{-1}^{1} = \text{antislope} \frac{v^4}{2} \; |_{-1}^{1} = \frac{v^5}{10} \; |_{-1}^{1} = \frac{1}{5}$$

b) Unsymmetric

20	20	20
1	2	3

1/3	1/3	1/3
1	2	3

$$\frac{\sum v_i^4 h_i}{\sum h_i} = \frac{1 \times 20 + 4 \times 20 + 9 \times 20}{20 + 20 + 20} = \frac{98}{3} = 32.\overline{6}$$

$$\sum v_i^4 d_i = 1 \times \frac{1}{3} + 4 \times \frac{1}{3} + 9 \times \frac{1}{3} = \frac{98}{3} = 32.\overline{6}$$

$$d(v) = 1/2$$

$$\text{area } v^4 d(v) \; |_{1}^{3} = \text{antislope} \frac{v^4}{2} \; |_{1}^{3} = \frac{v^5}{10} \; |_{1}^{3} = \frac{242}{10} = 24.2$$

$$\begin{aligned} \mu_4 = \sum (v_i - \mu)^4 d_i &= \sum v_i^4 d_i - 4\mu \sum v_i^3 d_i + 6\mu 2 \sum v_i^2 d_i - 4\mu^3 \sum v_i d_i + \mu^4 \sum d_i \\ &= \sum v_i^4 d_i - 4\mu(\mu_3 + 3\mu\sigma^2 + \mu^3) + 6\mu^2(\sigma^2 + \mu^2) - 4\mu^4 + \mu^4 \\ &= \sum v_i^4 d_i - 4\mu(\mu_3 - 6\mu^2\sigma^2 - \mu^4) \end{aligned}$$

and we see that it is much easier to subtract the mean before taking the power and doing the sum.

The *kurtosis* (or coefficient of excess) is $\mu_4/\sigma^4 - 3$, which if negative indicates that the distribution is more peaked at the centre than normal ("normal" is precisely defined in the Note after next), and if positive indicates that the distribution is flatter than normal.

For the discrete uniform distributions above, the kurtosis is $-3/2$ (peaked) while for the uniform density above it is 2 (flat). Once again increasing the number of discrete steps will bring the kurtosis closer to that for the density.

Although the individual meanings of higher moments than $\mu_0, .., \mu_4$ become obscure, we can go on. For theoretical reasons we must go on, because it takes all the moments fully to determine the distribution (in many cases: it is not *always* possible to find the distribution even given all the moments).

Finding all these moments requires a bit of calculus which is worth reviewing for future reference. Here it is for a uniform distribution from a to b.

$$\text{area } v^k d(v) \; |_{a}^{b} = \text{area} \frac{v^k}{b-a} \Big|_a^b = \text{antislope} \frac{v^k}{b-a} \Big|_a^b = \frac{v^{k+1}}{k(b-a)} \Big|_a^b$$

For a symmetrical distribution (of mean $\mu = 0$), $a = -b$ and we can immediately see that $\mu_{2k-1} = 0$ for all odd subscripts $2k - 1$:

$$\mu_{2k-1} = \frac{v^{(2k-1)+1}}{2kb} \Big|_{-b}^{b} = \frac{b^{2k}}{2kb} - \frac{(-b)^{2k}}{2kb} = 0$$

Since we cannot find an infinity of moments, $\mu_k = \text{area}(v-\mu)^k d(v)$ one at a time, we need a shortcut.

The set $1, v, v^2, v^3, ..$ reminds us of the infinite series we saw in Note 7 of Week ii.

So our shortcut consists of finding a *moment generating function* which is some infinite sum of $1, v, v^2, v^3, ..$

The handiest such sum is

$$e^v = 1 + v + \frac{v^2}{2!} + \frac{v^3}{3!} + ..$$

because

$$\text{slope}_v e^v = e^v = \text{antislope}_v e^v$$

so we can do the calculus easily.

Well, what we actually do is replace each of $1, v, v^2, v^3, ..$ by e^{tv} in the area/antislope calculation. Let's see what comes out for the uniform density $d(v) = 1/2$ on -1 to 1.

$$\begin{aligned}
\text{antislope}_v e^{tv} d(v) \,|_{-1}^{1} &= \text{antislope}_v \left.\frac{e^{tv}}{2}\right|_{-1}^{1} \\
&= \left.\frac{e^{tv}}{2t}\right|_{-1}^{1} \\
&= \frac{e^t - e^{-t}}{2t} \\
&= \frac{1}{2t}(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \frac{t^5}{5!} + .. - (1 - t + \frac{t^2}{2!} - \frac{t^3}{3!} + \frac{t^4}{4!} - \frac{t^5}{5!} + ..)) \\
&= \frac{1}{2t}(2t + 2\frac{t^3}{3!} + 2\frac{t^5}{5!} + ..) \\
&= 1 + 0t + \frac{1}{3}\frac{t^2}{2!} + 0\frac{t^3}{3!} + \frac{1}{5}\frac{t^4}{4!} + .. \\
&= \mu_0 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \mu_4 \frac{t^4}{4!} + ..
\end{aligned}$$

This gives all the moments by just reading off the coefficients of $\frac{t^k}{k!}$:

$$\begin{aligned}
\mu_0 &= 1 \\
\mu_1 &= 0 = \mu_3 = \mu_{2k-1} \\
\mu_2 &= \frac{1}{3} \\
\mu_4 &= \frac{1}{5} \\
&\vdots \\
\mu_{2k} &= \frac{1}{2k+1}
\end{aligned}$$

(Remember that the continuous math is only an approximation to the actual discrete distributions—but an approximation which gets better and better as the discrete distributions get more and more densely populated.)

5. Quantum distributions: the density matrix. [1] An interesting way to write a distribution is

[1] This Note may be skipped without breaking continuity.

in terms of a set of orthonormal vectors. Thus, the three-element distribution of Note 3 becomes a three-dimensional space.

If we take the usual basis vectors we can add them to the distribution.

values	-1	0	1
frequencies	$1/3$	$1/3$	$1/3$
basis vectors	$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

Let's consider the matrix built up of weighted outer products (Week 2 Note 5).

$$\begin{aligned} \rho &= \frac{1}{3}\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}(0\ 0\ 1) + \frac{1}{3}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}(0\ 1\ 0) + \frac{1}{3}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}(1\ 0\ 0) \\ &= \frac{1}{3}\begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} \end{aligned}$$

This is called the *density matrix*. (To be self-consistent I should call it the distribution matrix, but I'll be conventional this time.)

Note that its trace (Book 8c Note 8 (Part I)) is 1:

$$\mathrm{Tr}\frac{1}{3}\begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix} = 1$$

Before investigating the significance of this. let's do it again with different basis vectors.

$$\begin{aligned} \rho' &= \frac{1}{3}\frac{1}{\sqrt{6}}\begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}\frac{1}{\sqrt{6}}(1\ 1\ -2) + \frac{1}{3}\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}\frac{1}{\sqrt{2}}(1\ -1\ 0) + \frac{1}{3}\frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\frac{1}{\sqrt{3}}(1\ 1\ 1) \\ &= \frac{1}{18}\begin{pmatrix} 6 & & \\ & 6 & \\ & & 6 \end{pmatrix} \end{aligned}$$

It's the same, so of course the trace is the same.

The trace will always be 1 but the matrix itself usually depends on how it was built. Let's try different weights, 1/4, 1/2, 1/4.

$$\begin{aligned} \rho &= \frac{1}{4}\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}(0\ 0\ 1) + \frac{1}{2}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}(0\ 1\ 0) + \frac{1}{4}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}(1\ 0\ 0) \\ &= \frac{1}{4}\begin{pmatrix} 1 & & \\ & 2 & \\ & & 1 \end{pmatrix} \\ \rho' &= \frac{1}{4}\frac{1}{\sqrt{6}}\begin{pmatrix} 1 \\ 1 \\ -2 \end{pmatrix}\frac{1}{\sqrt{6}}(1\ 1\ -2) + \frac{1}{2}\frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}\frac{1}{\sqrt{2}}(1\ -1\ 0) + \frac{1}{4}\frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\frac{1}{\sqrt{3}}(1\ 1\ 1) \\ &= \frac{1}{24}\begin{pmatrix} 9 & -3 & \\ -3 & 9 & \\ & & 6 \end{pmatrix} \end{aligned}$$

Are these traces the same?

Instead of keeping the weights separate, let's integrate them into the basis vectors. Because each basis vector appears twice in forming the density matrix, we must use the square root of the weights.

values	-1	0	1
weighted basis vectors	$\frac{1}{\sqrt{3}}\begin{pmatrix}0\\0\\1\end{pmatrix}$	$\frac{1}{\sqrt{3}}\begin{pmatrix}0\\1\\0\end{pmatrix}$	$\frac{1}{\sqrt{3}}\begin{pmatrix}1\\0\\0\end{pmatrix}$

$$\begin{aligned}\rho &= \frac{1}{\sqrt{3}}\begin{pmatrix}0\\0\\1\end{pmatrix}\frac{1}{\sqrt{3}}(0\ 0\ 1)+\frac{1}{\sqrt{3}}\begin{pmatrix}0\\1\\0\end{pmatrix}\frac{1}{\sqrt{3}}(0\ 1\ 0)+\frac{1}{\sqrt{3}}\begin{pmatrix}1\\0\\0\end{pmatrix}\frac{1}{\sqrt{3}}(1\ 0\ 0)\\ &= \frac{1}{3}\begin{pmatrix}1&&\\&1&\\&&1\end{pmatrix}\end{aligned}$$

or

values	-1	0	1
weighted basis vectors	$\frac{1}{\sqrt{2}}\begin{pmatrix}1\\1\\-2\end{pmatrix}$	$\frac{1}{\sqrt{4}}\begin{pmatrix}1\\-1\\0\end{pmatrix}$	$\frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\\0\end{pmatrix}$

$$\begin{aligned}\rho' &= \frac{1}{\sqrt{2}}\frac{1}{\sqrt{6}}\begin{pmatrix}1\\1\\-2\end{pmatrix}\frac{1}{\sqrt{2}}\frac{1}{\sqrt{6}}(0\ 0\ 1)+\frac{1}{\sqrt{4}}\frac{1}{\sqrt{6}}\begin{pmatrix}0\\1\\0\end{pmatrix}\frac{1}{\sqrt{2}}\frac{1}{\sqrt{6}}(1\ -1\ 0)+\frac{1}{\sqrt{2}}\frac{1}{\sqrt{6}}\begin{pmatrix}1\\1\\1\end{pmatrix}\frac{1}{\sqrt{2}}\frac{1}{\sqrt{6}}(1\ 1\ 1)\\ &= \frac{1}{24}\begin{pmatrix}9&-3&\\-3&9&\\&&6\end{pmatrix}\end{aligned}$$

We can speed this up (for the first example).

Suppose the weights are generally p^2, q^2 and r^2. (I've used squares for a reason which will become clear next.)

values	-1	0	1
frequencies	p^2	q^2	r^2

$$\begin{aligned}\rho &= \begin{pmatrix}p\\q\\r\end{pmatrix}(p\ q\ r)=\begin{pmatrix}p^2&pq&pr\\pq&q^2&qr\\pr&qr&r^2\end{pmatrix}\\ \mathrm{Tr}\rho &= p^2+q^2+r^2=1\end{aligned}$$

This is suggestive: p, q and r are *amplitudes* and their squares are *probabilities* in quantum mechanics. The amplitudes may even be 2-numbers, e.g., $p = 1/\sqrt{2}, q = 0, r = i/\sqrt{2}$.

$$\begin{aligned}\rho &= \frac{1}{\sqrt{2}}\begin{pmatrix}1\\0\\i\end{pmatrix}\frac{1}{\sqrt{2}}(1\ 0\ -i)=\frac{1}{2}\begin{pmatrix}1&0&-i\\0&0&0\\i&0&1\end{pmatrix}\\ \mathrm{Tr}\rho &= 1\end{aligned}$$

Since quantum mechanical operators are matrices, formulating a distribution as a matrix is useful. Let's find the expected value of an operator such as the z-rotation operator from Week 6 Note 9.

$$\begin{pmatrix} e^{i\beta} & & \\ & 1 & \\ & & e^{-i\beta} \end{pmatrix}$$

For state (p^*, q^*, r^*) (the 2-number example above shows that we must use conjugates (Week 4 Excursion "conjugates") for the horizontal vector) the expected value of this operator is

$$(p^*\ q^*\ r^*)\begin{pmatrix} e^{i\beta} & & \\ & 1 & \\ & & e^{-i\beta} \end{pmatrix}\begin{pmatrix} p \\ q \\ r \end{pmatrix} = p^2e^{i\beta} + q^2 + r^2e^{-i\beta}$$

But this is the same as the trace of

$$\begin{pmatrix} p \\ q \\ r \end{pmatrix}(p^*\ q^*\ r^*)\begin{pmatrix} e^{i\beta} & & \\ & 1 & \\ & & e^{-i\beta} \end{pmatrix} = \begin{pmatrix} p^2 & pq^* & pr^* \\ p^*q & q^2 & qr^* \\ p^*r & q^*rq & r^2 \end{pmatrix}\begin{pmatrix} e^{i\beta} & & \\ & 1 & \\ & & e^{-i\beta} \end{pmatrix}$$
$$= \begin{pmatrix} p^2e^{i\beta} & pq^* & pr^*e^{-i\beta} \\ p^*qe^{i\beta} & q^2 & qr^*e^{-i\beta} \\ p^*re^{i\beta} & q^*rq & r^2e^{-i\beta} \end{pmatrix}$$

So the density matrix, ρ, for any distribution, helps us find the average of matrix $\mathcal{A}$, $< \mathcal{A} >= \mathrm{Tr}\rho\mathcal{A}$. In particular, for quantum mechanics, the distribution comes from the vector of amplitudes, observable quantities come from matrix operations, and it is appropriate to speak of "the expected value".

What does the matrix "operator" look like for an ordinary distribution? Let's find the average value for

values	-1	0	1
frequencies	1/3	1/3	1/3

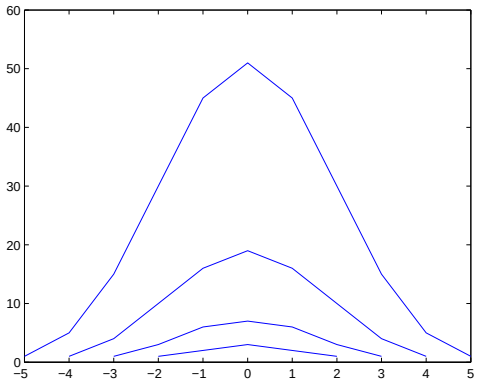
$$\begin{aligned} \text{average value} &= \mathrm{Tr}\rho\mathcal{A} \\ &= \mathrm{Tr}\frac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}\frac{1}{\sqrt{3}}(1\ 1\ 1)\mathcal{A} \\ &= \mathrm{Tr}\frac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} -1 & & \\ & 0 & \\ & & 1 \end{pmatrix} \\ &= \mathrm{Tr}\frac{1}{3}\begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} \\ &= 0 \end{aligned}$$

The matrix $\mathcal{A}$ is zero everywhere but the diagonal, which contains the values to be averaged.

6. The normal distribution.[2] The MATLAB program developed in Excursion "histogArith" for Note 2 enables us to see the result of repeatedly adding the uniform histogram

[2]This Note is lengthy and covers some calculus which is important for Normal distributions as well as more generally. You can skim the calculus on first reading.

1	1	1
−1	0	1



What is the eventual shape of the histogram?

We can find the moment generating function of the sum of densities if we know the moment generating functions of the input densities. This will determine the resulting density, although going backwards from moment generating function to density is not always easy.

First, for two densities with moment generating functions $e^{v_1 t}$ $e^{v_2 t}$ the moment generating function for $v_1 + v_2$ will be

$$e^{(v_1+v_2)t} = (e^{v_1 t})(e^{v_2 t})$$

In general, adding values of histograms results in multiplying moment generating functions.

Moreover, the mean of the sum is the sum of the means and the variance of the sum is the sum of the variances. Here is the product of moment generating functions up to t^2.

$$\begin{aligned} & (1 + \mu_1 t + (\sigma_1^2 + \mu_1^2)\tfrac{t^2}{2!} + ..)(1 + \mu_2 t + (\sigma_2^2 + \mu_2^2)\tfrac{t^2}{2!} + ..) \\ = \quad & 1 + (\mu_1 + \mu_2)t + (\sigma_1^2 + \sigma_2^2 + (\mu_1 + \mu_2)^2)\tfrac{t^2}{2!} + .. \\ = \quad & 1 + \mu_{12} t + (\sigma_{12}^2 + \mu_{12}^2)\tfrac{t^2}{2!} + .. \end{aligned}$$

where μ_1 and μ_2 are the two means for densities 1 and 2 (not the first and second moments this time) and σ_1^2 and σ_2^2 are the respective two variances. The resulting density has mean $\mu_{12} = \mu_1 + \mu_2$ and variance $\sigma_{12}^2 = \sigma_1^2 + \sigma_2^2$.

This last tells us that if we add m copies of a central density ($\mu_1 = 0$) the resulting variance, $\sigma^2 = m\sigma_1^2$.

We can see this expansion of the variance in the MATLAB plot above. (Because the histograms plotted are central, the means remain 0.)

Second, focusing on central densities ($\mu = 0$), let's sum m of them all the same with variance σ_1^2. The moment generating function of the result is

$$(1 + \sigma_1^2 \frac{t^2}{2!})^m = (1 + \frac{\sigma^2}{m}\frac{t^2}{2!})^m \; m \longrightarrow \infty \; e^{\sigma^2 t^2/2}$$

where the last limit is the definition of Euler's number, e. (There is a serious omission in this argument: see Excursions.)

Thus $e^{\sigma^2 t^2/2}$ is the moment generating function for the limit of the repeated sum of any central density. The density of the uniform histograms plotted by MATLAB is a special case.

But what density *has* the moment generating function $e^{\sigma^2 t^2/2}$?

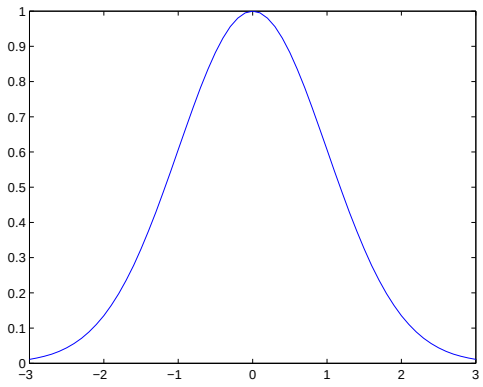
This, as I say, is tricky. It might be more straightforward to pull the answer out of a hat and then show that it *is* the answer. I think I can be less arbitrary—but I can't avoid arbitrariness altogether.

Let's start with a bit of subtle but useful calculus. This is motivated by the $e^{\sigma^2 t^2/2}$ we've just found. But it is still a little arbitrary. Hang on for the ride!

What is

$$I = \text{antislope } e^{-x^2/a} \mid_{-\infty}^{\infty}?$$

It is the area under $e^{-x^2/a}$ for all values of x. Let's let MATLAB (or a graphics calculator) show us what e^{-x^2} looks like.



Now a trick. We take two of these

$$\begin{aligned}
I^2 &= (\text{antislope } e^{-x^2/a} \mid_{x=-\infty}^{\infty})(\text{antislope } e^{-y^2/a} \mid_{y=-\infty}^{\infty}) \\
&\approx (\sum_{x=-\infty}^{\infty} e^{-x^2/a}\Delta x)(\sum_{y=-\infty}^{\infty} e^{-y^2/a}\Delta y) && 1 \\
&= \sum_{x=-\infty}^{\infty}\sum_{y=-\infty}^{\infty} e^{-x^2/a}\Delta x e^{-y^2/a}\Delta y && 2 \\
&= \sum_{x=-\infty}^{\infty}\sum_{y=-\infty}^{\infty} e^{(x^2+y^2)/a}\Delta x\Delta y \\
&= \sum_{r=0}^{\infty} e^{-r^2/a} 2\pi r\Delta r && 3 \\
&= \pi a \sum_{s=0}^{\infty} e^{-s}\Delta s && 4 \\
&\approx \pi a \text{ antislope } e^{-s} \mid_{s=0}^{\infty} && 5 \\
&= -\pi a e^{-s} \mid_{0}^{\infty} \\
&= \pi a
\end{aligned}$$

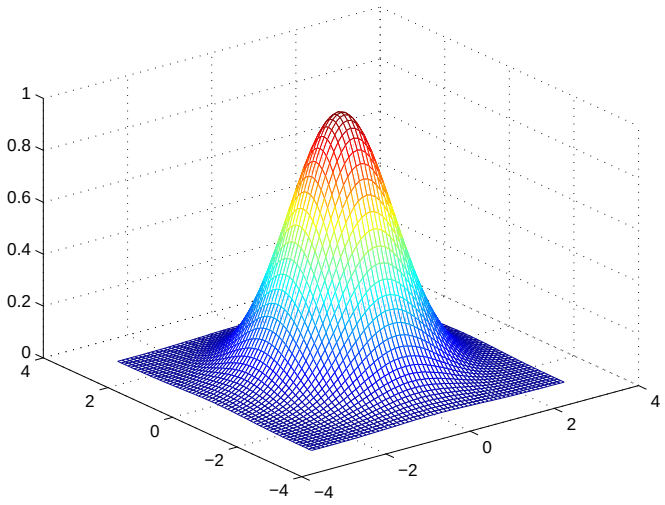
So $I = \sqrt{\pi a}$.

I should explain the steps in this chain of reasoning. First, in (1) we use the Fundamental Theorem of Calculus to convert the antislope $|_{-\infty}^{\infty}$ into an area, and in (5) we go back from area to antislope $|_{-\infty}^{\infty}$. The sums giving the areas have step sizes $\Delta x, \Delta y$ and Δs, respectively, and the equalities are approximate but are better the smaller the step sizes.

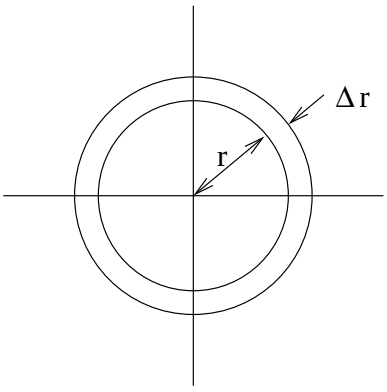
Second, in (2) we can swap the order of sums and products. You can see this in the example

$$\begin{aligned}\sum_j a_j \sum_k b_k = (a1 + a2)(b1 + b2) &= a_1 b_1 + a_1 b_2 \\ &+ a_2 b_1 + a_2 b_2 = \sum_j \sum_k a_j b_k\end{aligned}$$

The heart of the argument is in the change of variables in (3). Drawings will help. First, $e^{-(x^2+y^2)}$ is a 3D figure of revolution.



So we can write $r^2 = x^2 + y^2$. Second, $\Delta x \Delta y$ is a small increment of area ΔA. We could write this as $r\Delta r \Delta\theta$ in polar coordinates, but nothing in the area of the line above (3) depends on θ, so we took ΔA to be a thin ring of radius r: $\Delta A = 2\pi r \Delta r$



Note that r need only range from 0 to ∞ to cover the same total area.

Finally (4) does a formal change of variables $s = r^2/a$.

$$\begin{array}{lrcl} \text{So} & \Delta s = \Delta r \text{ slope}_r s & = & \Delta r 2r/a. \\ \text{So} & 2r\Delta r & = & a\Delta s \\ \text{and} & e^{-r^2/a} 2r\Delta r & = & e^{-s} a\Delta s \end{array}$$

The limits from 0 to ∞ do not change: $s = 0$ when $r = 0$ and $s = \infty$ when $r = \infty$.

Now we know that

$$\text{antislope}_x e^{-x^2/a} \,|_{-\infty}^{\infty} = \sqrt{\pi a}$$

we can consider

$$\begin{aligned} 1 &= \frac{1}{\sqrt{2\pi\sigma^2}} \text{antislope}_y e^{-y^2/(2\sigma^2)} \,|_{-\infty}^{\infty} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \text{antislope}_x e^{-(x-\sigma^2 t)^2/(2\sigma^2)} \,|_{-\infty}^{\infty} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \text{antislope}_x e^{-(x^2 - 2\sigma^2 tx + \sigma^4 t^2)/(2\sigma^2)} \,|_{-\infty}^{\infty} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \text{antislope}_x e^{-x^2/(2\sigma^2)} e^{tx} e^{-\sigma^2 t^2)/2} \,|_{-\infty}^{\infty} \end{aligned}$$

So

$$e^{\sigma^2 t^2/2} = \text{antislope}_x e^{tx} \left. \frac{e^{-x^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} \right|_{-\infty}^{\infty}$$

and we've just found that

$$\frac{e^{-x^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$$

is the distribution whose moment generating function is $e^{\sigma^2 t^2/2}$.

That is why, in the first line above, we set $a = 2\sigma^2$ and why, in the next line, we replaced y by $x - \sigma^2 t$.

This new density,

$$\frac{e^{-x^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$$

is called the *normal density* or *normal distribution*.

This distribution is the famous "bell curve" that is central to all discussions of measurement errors, experimental design and statistics in general. We should explore it a little.

First, we have shown that the sum of enough identical central densities gives a central normal distribution. This is the gist of the "central limit theorem". (But look up a proper statement of the central limit theorem, e.g., [MGB74, p.234], before supposing you now really know what it is.) It is plausible that any randomness which is sufficiently complicated will be normally distributed.

We can now confirm by direct calculation that the mean (first moment) $\mu = 0$ and that the variance (second moment in this central case) is indeed σ^2. This gives us two more important calculus exercises as well as useful results.

The first moment (times $\sqrt{2\pi\sigma^2}$) is

$$\begin{aligned} \sqrt{2\pi\sigma^2} \,\text{antislope}_v d(v) \,|_{-\infty}^{\infty} &= \text{antislope}_v \, v e^{-v^2/(2\sigma^2)} \,|_{-\infty}^{\infty} \\ &= \text{antislope}_y \, e^{-y/\sigma^2} \,|_{-\infty}^{\infty} \end{aligned}$$

$$\begin{aligned} &= -\sigma^2 e^{-y/\sigma^2} \,|_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

using the same technique of changing variables for v to $y = v^2/2$ because

$$\Delta y \approx \text{slope}_v \; y\Delta v = \text{slope}_v \frac{v^2}{2}\Delta v = v\Delta v$$

and noting that $\text{slope}(_y - \sigma^2 e^{-y/\sigma^2}) = e^{-y/\sigma^2}$.

To find the second moment we "integrate by parts" using

$$\begin{aligned} \text{slope}_v \; ve^{-v^2/(2\sigma^2)} &= (\text{slope}_v v)e^{-v^2/(2\sigma^2)} + v \; \text{slope}_v e^{-v^2/(2\sigma^2)} \\ &= e^{-v^2/(2\sigma^2)} - \frac{v^2}{\sigma^2}e^{-v^2/(2\sigma^2)} \end{aligned}$$

So, apply antislope to both sides of the first line above:

$$ve^{-v^2/(2\sigma^2)} = \text{antislope}_v(e^{-v^2/(2\sigma^2)}) - \text{antislope}_v \frac{v^2}{\sigma^2}e^{-v^2/(2\sigma^2)}$$

hence

$$\begin{aligned} \text{antislope}_v \; v^2 e^{-v^2/(2\sigma^2)} \,|_{-\infty}^{\infty} &= \sigma^2 \text{antislope}_v(e^{-v^2/(2\sigma^2)}) \,|_{-\infty}^{\infty} - \sigma^2 ve^{-v^2/(2\sigma^2)} \,|_{-\infty}^{\infty} \\ &= \sigma^2\sqrt{2\pi\sigma^2} - 0 \end{aligned}$$

so the second moment (variance)

$$\text{antislope}_v \; v^2 \frac{e^{-v^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} \,|_{-\infty}^{\infty} = \sigma^2$$

as we wanted to show.

From these two results, we have confirmed that, like the uniform distribution, the normal distribution is fully described by μ and σ^2—assuming we know it is normal. Of course, the other moments must also work out right.

This integration by parts gives us all the other moments. By the same reasoning

$$\text{slope}_v \; v^k e^{-v^2/(2\sigma^2)} = kv^{k-1}e^{-v^2/(2\sigma^2)} - \frac{v^{k+1}}{\sigma^2}e^{-v^2/(2\sigma^2)}$$

so

$$\begin{aligned} \text{antislope}_v \; v^{k+1}e^{-v^2/(2\sigma^2)} \,|_{-\infty}^{\infty} &= \sigma^2 k \; \text{antislope}_v \; v^{k-1}e^{-v^2/(2\sigma^2)} \,|_{-\infty}^{\infty} - \sigma^2 v^k e^{-v^2/(2\sigma^2)} \,|_{-\infty}^{\infty} \\ &= \sigma^2 k \; \text{antislope}_v \; v^{k-1}e^{-v^2/(2\sigma^2)} \,|_{-\infty}^{\infty} - 0 \end{aligned}$$

since $e^{-v^2/(2\sigma^2)}$ gets small much faster than v^k gets big for any k as v gets arbitrarily large.

So the $k+1$st moment is $\sigma^2 k$ times the $k-1$st moment, skipping the kth. If k is even ($k+1, k-1$ odd) this can be taken all the way back to the 1st moment which we just showed is 0. So all odd moments are 0, as they were for the 0-centred uniform density.

If k is odd, the $k+1$st moment can be taken all the way back to the 2nd (or even 0th) moment. for example

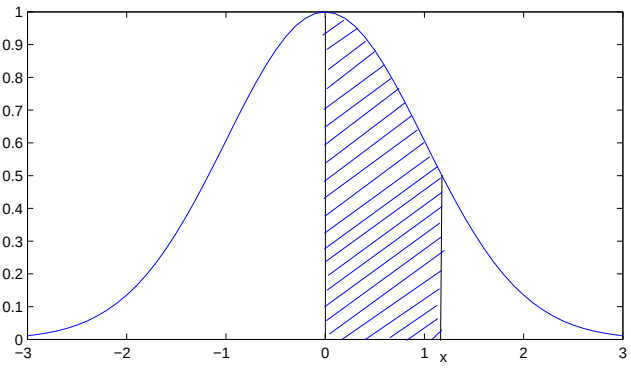
$$\begin{aligned} \mu_4 &= 3\sigma^2\mu_2 = 3\sigma^4 \\ \mu_6 &= 5\sigma^2\mu_4 = 15\sigma^6 \end{aligned}$$

and so on.

⋆ ⋆ ⋆

How much of the normal distribution lies within a given range, $-x$ to x, of the mean? This is a situation where calculus fails us and we must resort to calculation.

This quantity is important, however, and warrants a special name, the *error function*.



$$\begin{aligned}
\mathrm{erf}(x) &= \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{antislope}_v e^{-v^2/(2\sigma^2)} \;|_0^x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{antislope}_v (1 - \frac{v^2}{2\sigma^2} + (\frac{v^2}{2\sigma^2})^2\frac{1}{2!} - (\frac{v^2}{2\sigma^2})^3\frac{1}{3!} + ..) \;|_0^x \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}(x - \frac{x^3}{3\times 2\sigma^2} + \frac{x^5}{5\times 2!(2\sigma^2)^2} - \frac{x^7}{7\times 3!(2\sigma^2)^3} + ..) \\
&= \frac{1}{\sqrt{2\pi}}(y - \frac{y^3}{3\times 2} + \frac{y^5}{5\times 2!2^2} - \frac{y^7}{7\times 3!2^3} + ..)
\end{aligned}$$

where we have expanded $e^{-v^2/(2\sigma^2)}$ into the series, integrated the series term by term taking the upper limit (since the lower limit is 0), and finally changing variables $y = x/\sigma$ in order to use the standard deviation σ as the unit.

Here is the MATLAB program to find $2\times\mathrm{erf}(\sigma_y)$ to `decpl` decimal places. (I have left in the scaffolding: even in this little program I made some silly mistakes. It's not the mistakes that matter but the ability to recover from them. But if one makes too many mistakes one might not keep up with the world leaders.)

```
% function erf =  erfNormal(y,decpl)            THM                    090907
% in file erfNormal.m
% calculate Gaussian error function for y standard deviations to decpl dec place
% erf = 2 intgrl_0^x exp(-v^2/2sigma^2) dv      NB double to get both sides
function erf =  erfNormal(y,decpl)
tol = 10^-(decpl+1)             % display this
k = 0;
term = y*sqrt(2/pi);    % NB 2/sqrt(2pi)
%step = y       % test only
erf = term
while abs(term) > tol
 k = k + 1;
 term = -term*(2*k-1)*y^2/(2*k+1)/(2*k);
 %step = step*(2*k-1)*y^2/(2*k+1)/(2*k) % test only
```

```
 erf = erf + term                % display this
end
```

`erfNormal(1,2)` gives `0.6827` after 6 steps
`erfNormal(2,2)` gives `0.9546` after 10 steps
`erfNormal(3,2)` gives `0.9972` after 16 steps

I doubled erf() in order to find the area from $-x$ to x. The results mean that 68% of the possibilities fall within ± 1 standard deviations of the mean and 95% fall within ± 2 standard deviations.

The 2- and 3-standard deviation results are what lead statisticians to talk about "95% confidence", "99% confidence" and so on.

7. Expectation, surprise and ignorance. "Expectation" is a technical term meaning "average". There are subtle differences: it is unusual to think of the "average value" of one thing. What is the average position of a single molecule in a box? We must reinterpret this question in terms of a time average. More generally, we can take an "ensemble" average. An *ensemble* is an imagined set of data covering all possible positions, in this case, of the molecule, or, since there are too many possible positions, of a reasonable sample of them. Imagine, say, 1000 virtual molecules, each representing a different position of our one molecule. This ensemble of positions is now many positions and we can speak unconfusingly of their average.

"Expectation" is often used in this situation so we do not need mentally to convert one into many. But the calculation process is still the same, via an ensemble.

"Expectation" can be a more general term than "average": the average is the expectation of the quantity itself. But the *variance* is the expectation of the *square* of the quantity. It can be confusing to use "average" for both mean and variance, and long-winded to spell each out.

"Expectation" meaning average is not really intuitive. In a uniform distribution there is no particular value that we "expect" to find. In a non-uniform distribution, the value we "expect" to find would be the *mode*, the most frequent value (see the excursion "Mean, median, mode").

The upshot is that I will try not to use "expectation" or "expected value". I will use "average" for the general sense so that, say, variance is the average of the squares, and "mean" for particular sense, namely the average of the values themselves.

But "expectation" is a common term, and the point of the above discussion is to explain how it is used. The idea of an ensemble is the more important idea to take away. I also do not want "expectation" confused with "surprise".

"Surprise" is an important aspect of knowing only the distribution (histogram), having thrown away the particular values (if they were ever known) giving rise to it.

If d_j for a certain value j is very small then we are very surprised when j happens. If d_j is large then we are not (or are less) surprised.

To quantify surprise into a number we'll call *surprisal* we work with normalized histograms, i.e., what we've called distributions. If d_j is 0 our surprise should be infinite. If d_j is 1 our surprise should be 0.

A function on d_j which produces such values is the logarithm—actually the negative of the logarithm since $\log d_j \leq 0$ because $d_j \leq 1$.

We can take the logarithm to any base and it will have these needed properties at 0 and at 1. So it is a convention to use base 2. (We will later find the math is easier if we use base e, the "natural" or "Naperian" logarithms.)

Thus *surprisal*

$$s_j = -\lg d_j = \lg(1/d_j)$$

where $\lg \equiv \log_2$ the logarithm to base 2.

For a uniform distribution, we are equally surprised by any result

value	-1^oC	0^oC	1^oC
frequencies	1/3	1/3	1/3
surprisal	lg 3	lg 3	lg 3

where lg 3 = 1.58..

For a non-uniform distribution, surprisal varies

value	-3^oC	-2^oC	-1^oC	0^oC	1^oC	2^oC	3^oC
frequencies	1/27	3/27	6/27	7/27	6/27	3/27	1/27
surprisal	3 lg 3	2 lg 3	2 lg 3 −1	3 lg 3 − lg 7	2 lg 3 −1	2 lg 3	3 lg 3
	4.75	3.42	2.42	1.94	2.42	3.42	4.75

Here we can see that the extremes have a surprisal of about 5 while the mode is closer to 2.

More useful than the surprisal of each value in a distribution is the average surprisal, which we will call our *ignorance* of the distribution.

$$\text{ignorance} = \sum d_j s_j = -\sum d_j \lg d_j$$

Thus, for the uniform distribution on three items

$$\text{ignorance} = \lg 3 = 1.58$$

For the nonuniform distribution on seven values, above

$$\text{ignorance} = (63 \lg 3 - 12 - 7 \lg 7)/27 = 2.53$$

This should be compared with the corresponding ignorance of the uniform distribution on 7 elements

$$\text{ignorance} = \lg 7 = 2.81$$

Ignorance is greatest for the uniform distribution over a given number of values. Let's check this for the special case of distributions over two values. We'll say $x = d_1$ and $y = d_2$. Then we want to maximize $-x \lg x - y \lg y$ subject to $x + y = 1$.

We can use a Lagrangian multiplier almost as we did in Week 8 Note 11 [3].

$$\begin{aligned} 0 &= \begin{pmatrix} \text{slope}_x \\ \text{slope}_y \\ \text{slope}_\lambda \end{pmatrix} (-x \lg x - y \lg y + \lambda(x + y - 1)) \\ &= \begin{pmatrix} \lambda - \lg x - \lg e \\ \lambda - \lg y - \lg e \\ x + y - 1 \end{pmatrix} \end{aligned}$$

[3] Because $\lg x = \ln x / \ln 2 = \lg e \ln x$ and slope ln $x = 1/x$,

$$\begin{aligned} \text{slope}_x(x \lg x) &= \lg x + x \, \text{slope}_x(\lg x) \\ &= \lg x + x \, \text{slope}_x(\lg e \ln x) \\ &= \lg x + x \lg e / x \\ &= \lg x + \lg e \end{aligned}$$

(Why might we consider using ln instead of lg to define surprisal?)

Thus $\lg x - \lg e = \lambda = \lg y - \lg e$ so $\lg x = \lg y$ so $x = y$.

This argument extends easily to any number n of distribution entries

$$0 = \begin{pmatrix} \text{slope}_{d_1} \\ \vdots \\ \text{slope}_{d_n} \\ \text{slope}_{\lambda} \end{pmatrix} (-\sum d_j \lg d_j + \lambda(\sum d_j - 1))$$

and eventually d_j are all equal to each other, i.e., the distribution is uniform.

Thus ignorance never exceeds $\lg n$ for n different possible values.

Note that ignorance increases with the number of possibilities (unless the distributions are 0 everywhere but for one possibility).

This may seem strange, since ignorance is an average and averages don't depend on the number of values. That is true for the average of the quantities themselves, but ignorance is not such an average.

Variance is also an average, but not such an average. It is entirely plausible that variance should increase with the number of possibilities, since variance measures the spread of the distribution and a distribution over many values spreads further than one over few values,

Maybe ignorance is also a measure of spread. Let's compare it with variance for three distributions of five entries each[4].

values	-2	-1	0	1	2
freq 1	1/5	1/5	1/5	1/5	1/5
freq 2	1/9	2/9	3/9	2/9	1/9
freq 3	2/9	2/9	1/9	2/9	2/9

	mean	variance	ignorance
freq 1	0	2	$\lg 5 = 2.32$
freq 2	0	$4/3 = 1.\overline{3}$	$\frac{5}{3}\lg 3 - 4/9 = 2.20$
freq 3	0	$20/9 = 2.\overline{2}$	$2\lg 3 - 8/9 = 2.28$

There are important differences. Understandably the variance for the uniform distribution, freq 1, lies between the variances for the peaked distribution, freq 2, and the "bimodal" distribution, freq 3. This cannot happen for ignorance since uniform ignorance is always the highest. Ignorance is less if there are any number of peaks, but the fewer the peaks the lower the ignorance.

The most important difference between ignorance and variance, or mean for that matter, is that ignorance *does not depend on the values*. It depends only on the distribution.

This is really handy if the values are not numbers. For example, nine apples have the histogram

values	green	red	yellow
# occurrences	2	4	3

What is the mean colour? The variance? The ignorance?

Being an average, ignorance is *additive* in the right circumstances. If our ignorance about what Joe is owed is lg 3 in Note 2 and our ignorance about what Sue is owed is also lg 3, then our combined ignorance is 2 lg 3. This is lg 9 and there are 9 combined possibilities.

We should be careful about this. From Note 2 our ignorance about what is owed to Joe *or* Sue is $\frac{5}{3}\lg 3 - \frac{4}{9}$ (it's just the ignorance for freq 2, above), which is not the same as (it's less than) 2 lg 3. But that distribution concerns the possibilities of *some person* owing $-2\$, -1\$, 0\$, 1\$$ or $2\$$. This

[4] ignorance for freq $2 = \lg 9 - \frac{2}{9}\lg 1 - \frac{4}{9}\lg 2 - \frac{3}{9}\lg 3$; ignorance for freq $3 = \lg 9 - \frac{8}{9}\lg 2 - \frac{1}{9}\lg 1$

is less than treating the Joe and Sue distributions independently because to get it we have added the information that it is the same set of people owing Joe and Sue. Information dispels ignorance, so the result is less ignorance than if we didn't have the information.

"Ignorance" is my name for this quantity. It has other names. One is obvious (what dispels ignorance?). One might have been "uncertainty" but that has been taken by the Heisenberg uncertainty principle of quantum mechanics (see Week 9). The third name for ignorance was used by Shannon because von Neumann told him that "nobody knows what [it] really is", but this name has beem claimed by thermodynamics for a special case of this quantity, so it is controversial and I will not use it here.[5]

When we define it with logarithms to base 2, ignorance is measured in units of *bits*.

8. Does ignorance ever decrease? A student of mine had T-shirt saying "Hell hath no fury like a Dad whose tools are all messed up". Let's call him Stu, for anonymity, and snoop into home life with Stu and Dad. Every day Dad comes home from work and goes to his workshop `W`. (On weekends he goes to the workshop in the afternoon after shopping with Mom or picnicking with Mom and Stu.) Sometimes the tool he needs is not there but in Stu's room `S`. So he goes to find it and brings it back. Let's focus on one tool, say the pliers.

Stu has increased Dad's ignorance of the whereabouts of the tool. If Stu has taken the pliers every other day, the ignorance is maximized:

$$\begin{aligned} I &= -\frac{1}{2}\lg(1/2) - \frac{1}{2}\lg(1/2) \\ &= \frac{1}{2}\lg(2) - \frac{1}{2}\lg(1) + \frac{1}{2}\lg(2) - \frac{1}{2}\lg(1) \\ &= \lg(2) \\ &= 1 \end{aligned}$$

If Stu has the tool 5 days out of 6, Dad's ignorance is less: $\lg 6 - (5/6)\lg 5 = 0.65$ bits. This is because Dad is pretty sure he'll find the pliers in Stu's room. (Dad's ignorance is the same as this if Stu has the pliers 1/6 of the time: why?)

If Stu has the tool 2 days out of 3, Dad's ignorance is $\lg 3 - 2/3 = 0.918$ bits.

But Stu's projects are more ambitious than pliers alone can accomplish. There are a dozen tools which can go missing from the workshop: pliers, hammer, saw, screwdriver, soldering iron, drill, chisel, socket wrench, fretsaw, file, shears and pick. Suppose Dad comes home one day and finds that the pliers, saw, screwdriver, chisel, file and shears had all wound up in Stu's room that day. *We* can represent this situation as `SWSSWWSWWSSW`. Because each tool has a 50% chance of being either in `W` or `S`. Dad's ignorance is again 1 bit per tool. (Do the calculation!) That is, 12 bits total for the 12 tools: without checking, all *Dad* knows is `????????????` with each `?` having a 50-50 chance of being `W` or `S`.

The way Dad decreases his ignorance is by tidying up: afterwards, he has `WWWWWWWWWWWW` and his ignorance is 0.

But Dad could equally well decrease ignorance by just going to Stu's room and *recording* the `SWSSWWSWWSSW`. Then he can locate any tool by just consulting the record. His ignorance is 0.

If Dad were to keep such a record he could also tidy up and then *reverse* his tidying-up by replacing in Stu's room the tools Stu had taken in the first place. Dad has no plausible reason for wanting to do this, but I want to make the point that zero ignorance implies reversibility.

If Dad were to lose his record, his ignorance would again be 1 bit/tool and he could not reverse any

[5] Although the example I've chosen suggests that ignorance is something we *could* know but just happen not to know, there is nothing in the word that implies this. Ignorance could also refer to what we don't know and *could never* know, such as, perhaps, the mechanisms underlying the amplitudes and probabilities of quantum physics.

tidying he might have done. If Dad kept such a record *every day* for every tool, he could reproduce any of Stu's depredations for any day in the past, even after a daily tidying. Dad's ignorance would be always 0. But he might soon lose the slips of paper, or run out of room in his record book and have to erase earlier records to make room for new records. Then, again, his ignorance increases by the extent to which he has lost or erased data.

Stu and Dad are a little fantasy of mine. But now think of a gas of molecules. The physics of billiard balls and of molecules colliding with each other and with the walls of the container is precise and time-reversible. If we are writing a computer simulation, we keep a record of all the positions (for a while). But if we are *observing* the gas we do not keep a record, so our ignorance does not decrease and what we observe is not reversible. We will be looking at this a lot more closely in the following Notes.

Stu and Dad can tell us about combined ignorance. Consider two tools—pliers and hammer. Suppose the pliers are in Stu's room 5/6 of the time. As above,

$$I_{\mathrm{p}} = \lg(6) - \frac{5}{6}\lg(5) = 0.65$$

Suppose the hammer is in Stu's room 1/3 of the time

$$I_{\mathrm{h}} = \lg(3) - \frac{2}{3} = 0.918$$

Dad's ignorance of *both* is just the sum of his ignorance of either

$$I_{\mathrm{ph}} = 2\lg(3) - \frac{5}{6}\lg(5) + 1/2 = 1.568$$

assuming that Stu makes off with the hammer independently of his making off with the pliers.

This is because, with independence, the combined frequencies are

$$\begin{array}{rcl}
\text{p in } \texttt{W}\text{, h in } \texttt{W}\text{: } \texttt{WW} & = & \frac{1}{6}\frac{2}{3} = \frac{2}{18} \\
\text{p in } \texttt{W}\text{, h in } \texttt{S}\text{: } \texttt{WS} & = & \frac{1}{6}\frac{1}{3} = \frac{1}{18} \\
\text{p in } \texttt{S}\text{, h in } \texttt{W}\text{: } \texttt{SW} & = & \frac{5}{6}\frac{2}{3} = \frac{10}{18} \\
\text{p in } \texttt{S}\text{, h in } \texttt{s}\text{: } \texttt{SS} & = & \frac{5}{6}\frac{1}{3} = \frac{5}{18}
\end{array}$$

By calling $\frac{1}{6}$ "w_1", $\frac{5}{6}$ "w_2", $\frac{2}{3}$ "s_1" and $\frac{1}{3}$ "s_2" we can see formally (by remembering $w_1 + w_2 = 1$ and $s_1 + s_2 = 1$) that

$$\begin{array}{rcl}
\sum_{jk} w_j s_k \lg(w_j s_k) & = & w_1 s_1 \lg(w_1 s_1) \\
& & +w_1 s_2 \lg(w_1 s_2) \\
& & +w_2 s_1 \lg(w_2 s_1) \\
& & +w_2 s_2 \lg(w_2 s_2) \\
& = & w_1 s_1 (\lg(w_1) + \lg(s_1)) \\
& & +w_1 s_2 (\lg(w_1) + \lg(s_2)) \\
& & +w_2 s_1 (\lg(w_2) + \lg(s_1)) \\
& & +w_2 s_2 (\lg(w_2) + \lg(s_2)) \\
& = & w_1 \lg(w_1) + s_1 \lg(s_1) \\
& & +w_2 \lg(w_2) + s_2 \lg(s_2) \\
& = & (\sum_j w_j \lg(w_j)) + (\sum_k s_k \lg(s_k))
\end{array}$$

Hence $I_{\text{ph}} = I_{\text{p}} + I_{\text{h}}$ as we said above.

We can also see that a combined average based on these frequencies is the sum of the individual averages. Let's suppose Dad hands out one demerit point for each tool he finds in Stu's room. Then the average demerit points for the pliers is

$$< \text{dem}_{\text{p}} >= 0 * \frac{1}{6} + 1 * \frac{5}{6} = \frac{5}{6}$$

and the average demerit points for the hammer is

$$< \text{dem}_{\text{h}} >= 0 * \frac{2}{3} + 1 * \frac{1}{3} = \frac{1}{3}$$

Then the combined average demerit points is

$$< \text{dem}_{\text{ph}} >= \frac{5}{6} + \frac{1}{3} = \frac{7}{6}$$

We can also see this formally:

$$\begin{aligned} 0 \times w_1 s_1 + 1 \times (w_1 s_2 + w_2 s_1) + 2 \times w_2 s_2 &= w_1 s_2 + w_2 s_2 + w_2 s_1 + w_2 s_2 \\ &= (0 \times s_1 + 1 \times s_2) + (0 \times w_1 + 1 \times w_2) \end{aligned}$$

Note that we have made an additional assumption in calculating demerits, namely that we don't care to discriminate between one pair of pliers in Stu's room and one hammer. We only care that there is one tool. So the four frequencies above become three: 2/18, 11/18 = 1/18 + 10/18 and 5/18. This in turn reduces the ignorance from 1.5683 for four frequencies to 1.2997 for three. (Work this out!)

For a final insight into ignorance, suppose that shortly after all this fury and demerits blow up, Mom travels to a conference. She wanted to forget the details of the family stress but couldn't get it totally out of her mind. So on the train coming home she decides to reconstruct the frequencies. All she remembers is that the average demerits for pliers and hammer were 7/6 as we just worked out above.

Mom reasons that since she knows only this, her ignorance about everything else must be maximum. So she decides to maximize ignorance, subject to a)

$$p_0 + p_1 + p_2 - 1 = 0$$

and b)

$$p_0 \times 0 + p_1 \times 1 + p_2 \times 2 - \frac{7}{6} = 0$$

In these equations, Mom uses p_j to represent the frequencies she does not know. There are three of them, p_0 for no tools in Stu's room, p_1 for one tool and p_2 for two tools. The symbol p stands for "probability", which we can now define for the first time.

A *collection of probabilities* is a collection (not a set: there may be duplicate values) of non-negative numbers which sum to 1 and which generally behave just like a collection of frequencies. "Probability" is a tricky and often vague concept which can be made much clearer by just substituting the word "frequency", and, if helpful, by working out a concrete example of actual or invented frequencies.

The two equations mean (a) the probabilities sum to 1 and (b) the mean is 7/6.

This is a problem in constrained maximization, with equations (a) and (b) giving the constraints. It can be solved by Lagrange multipliers (Week 8 Note 11) which Mom proceeds to do. Since this will involve finding slopes, Mom uses natural logarithms, ln, instead of logarithms to base 2, lg.

To maximize ignorance, Mom maximizes

$$p_0 \ln(p_0) + p_1 \ln(p_1) + p_2 \ln(p_2)$$

(or the negative of this: why does it make no difference?) subject to the above conditions. She uses as Lagrange multipliers two new variables, α' and β, which multiply the constraint expressions (a) and (b) respectively before adding them to the above expression for the ignorance, to be maximized.

$$
\begin{aligned}
0 &= \begin{pmatrix} \text{slope}_{p_0} \\ \text{slope}_{p_1} \\ \text{slope}_{p_2} \\ \text{slope}_{\alpha'} \\ \text{slope}_{\beta} \end{pmatrix} p_0 \ln(p_0) + p_1 \ln(p_1) + p_2 \ln(p_2) + \alpha'(p_0 + p_1 + p_2 - 1) + \beta(p_0 \times 0 + p_1 \times 1 + p_2 \times 2 - \frac{7}{6}) \\
&= \begin{pmatrix} \ln(p_0) + 1 + \alpha' + 0 \times \beta \\ \ln(p_1) + 1 + \alpha' + 1 \times \beta \\ \ln(p_2) + 1 + \alpha' + 2 \times \beta \\ p_0 + p_1 + p_2 - 1 \\ p_0 \times 0 + p_1 \times 1 + p_2 \times 2 - \frac{7}{6} \end{pmatrix} \\
&= \begin{pmatrix} \ln & & & 1 & \\ & \ln & & 1 & 1 \\ & & \ln & 1 & 2 \\ 1 & 1 & 1 & & \\ & 1 & 2 & & \end{pmatrix} \begin{pmatrix} p_0 \\ p_1 \\ p_2 \\ \alpha \\ \beta \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ \frac{7}{6} \end{pmatrix}
\end{aligned}
$$

where $\alpha = 1 + \alpha'$ is the variable she'll use from now on instead of α'.

Unlike Week 8 Note 11, we cannot just use linear algebra to solve this because of the logarithms, ln, that appear in the matrix. But it is fairly easy to solve if we take α and β as parameters:

$$
\begin{aligned}
\ln(p_0) = -\alpha \quad &\text{so} \quad p_0 = e^{-\alpha} \\
\ln(p_1) = -\alpha - \beta \quad &\text{so} \quad p_1 = e^{-\alpha-\beta} \\
\ln(p_2) = -\alpha - 2\beta \quad &\text{so} \quad p_2 = e^{-\alpha-2\beta}
\end{aligned}
$$

and Mom can get rid of the α since

$$1 = p_0 + p_1 + p_2 = e^{-\alpha}(1 + e^{-\beta} + e^{-2\beta})$$

so

$$e^{\alpha} = 1 + e^{-\beta} + e^{-2\beta}$$

and we call this the "partition function" because it shows how the new distribution is divided up among the three probabilities.

In this simple case Mom can even find the values for β, although it is not always necessary to calculate them right out. Define $x = e^{-\beta}$ and note that $e^{-2\beta}$ is x^2. Then use the constraint equation for the average demerit

$$
\begin{aligned}
\frac{7}{6} &= (x + 2x^2)/(1 + x + x^2) \\
\text{to get } 0 &= \frac{5}{6}x^2 - \frac{1}{6}x - \frac{7}{6} \\
\text{or } 0 &= 5x^2 - x - 7 \\
\text{so } x &= (1 \pm \sqrt{1 + 140})/10 \\
&= 1.2874
\end{aligned}
$$

choosing the + from $\pm$, and so

$$\beta = -\ln(x) = -0.2526$$

So the final answer for the probabilities (which needs only x, not β or α) is, using the partition function,

$$Z = 1 + x + x^2 = 1 + x(1 + x) = 3.9448$$

$$\begin{aligned} p_0 &= \frac{1}{Z} = 0.2535 \\ p_1 &= \frac{x}{Z} = 0.3264 \\ p_2 &= x \times p_1 = 0.4201 \end{aligned}$$

which sum to 1 and give $p_1 + 2 \times p_2 = 1.1666$, i.e., 7/6 to the precision calculated.

These probabilities are quite different from the original frequencies, which Mom looked up when she got home: 2/18, 11/18, 5/18 = 0.1111, 0.6111, 0.2778. We didn't say this would reproduce the original frequencies. Mom's ignorance has been maximized, increasing from Dad's ignorance of 1.2997 (over the three frequencies used to calculate demerits) to 1.554. (Check this!)

It should be apparent from the form of the maximization that the resulting probabilities will either increase or decrease exponentially, and certainly cannot go up then down the way the original frequencies do.

Maximizing ignorance to get the partition function, if not the probabilities directly, is a central operation in thermodynamics. I have tried to demystify it in this Note by using it in the Stu-Dad-Mom fable so that you can see it as a mathematical argument independent of the complications of physics.

Following up on the Excursion "Increasing ignorance" we see that in a complicated process, one might start off with a pretty good idea of the inner workings—at least of the frequencies and that certain values result—but as time goes on one loses more and more of this knowledge until one is left only something measurable, say an average, and ignorance is maximized. This can be thought of as a settling of the process into an "equilibrium" of maximum ignorance.

9. Inside knowledge: the clients of Joe and Sue revisited.[6] When in Note 2 we put together the two uniform distributions

# people	1	1	1
owes Joe	−1$	0$	1$

# people	1	1	1
owes Sue	−1$	0$	1$

to get the sum

# ways	1	2	3	2	1	total 9
owes Joe or Sue	−2$	−1$	0$	1$	2$	

we made an assumption.

Not only did we suppose that Joe and Sue are dealing with the *same* three people, which I stated, but we also supposed that these three people had no preferences for Joe or Sue, or for borrowing from both or lending to both, or all sorts of possible exclusions. This assumption we made is plausible if we are completely ignorant of the interactions among these three people and Joe and Sue.

Our ignorance-based assumption must be modified if we learn something about the three. For instance, if each of them is exclusively either a borrower or a lender or disdaining of any interaction with Joe or Sue, then the "sum" would be

# ways	3	0	3	0	3	total 9
owes Joe or Sue	−2$	−1$	0$	1$	2$	

[6]Notes 9, 10 and 11 are important refinements needed by statisticians, and Note 11 is often used in artificial intelligence. You may skip them if you want to go straight on to thermodynamics.

How do we get this? Let's look at the details. We'll have to name the three clients, so we'll call them a, b and c.

How many ways can a, b and c, in combination, owe Joe $-1\$$, $0\$$ and $1\$$, leading to one of them owing only, one being owed only and one free of financial interaction? There are six ways, which we can codify as

$$(a, b, c)\epsilon\{(-1, 0, 1), (-1, 1, 0), (0, -1, 1), (0, 1, -1), (1, -1, 0), (1, 0, -1)\}$$

where we leave off the \$ sign from now on to save encumbrance.

The meaning of this codification can be illustrated by the first case.

$$(a, b, c) = (-1, 0, 1)$$

means a owes Joe $-1\$$ (a lent the dollar to Joe), b is aloof (b neither owes Joe or is owed) and c owes Joe $1\$$.

The same codification gives the six ways a, b or c can owe Sue, so that Sue's distribution is also uniform over the three possibilities.

Combining Joe and Sue gives a table of $6 \times 6 = 36$ ways a, b and c can separately interact with them, leaving each of Joe and Sue with uniform distributions.

Sue $abc =$		–101	–110	0−11	01−1	1−10	10−1
Joe $abc =$	–101	–202	–211	–1−12	–110	0−11	000
	–110	–211	–220	–101	–12−1	000	01−1
	0−11	–1−12	–101	0−22	000	1−21	1−10
	01−1	–110	–12−1	000	02−2	10−1	11−2
	1−10	0−11	000	1−21	10−1	2−20	2−1−1
	10−1	000	01−1	1−10	11−2	2−1−1	20−2

Each entry is just the sum of the corresponding margin entries.

Out of this table we can construct *joint distributions* for the combined financial interactions of Joe and Sue with a, b and c. These joint distributions, or histograms, are, once again, abstractions, eliminating a, b and c from consideration.

We'll start with the assumption of complete ignorance, which we originally used in Note 2. That is, we'll forget for the moment that a and b and c might be exclusively a borrower, a lender, or not transacting at all. We must extract from the table a matrix giving the nine elements

$$\begin{aligned}(\text{owes Joe, owes Sue}) \quad \epsilon\{ \quad & (-1\$, -1\$), (-1\$, 0\$), (-1\$, 1\$), \\ & (0\$, -1\$), (0\$, 0\$), (0\$, 1\$), \\ & (1\$, -1\$), (1\$, 0\$), (1\$, 1\$)\}\end{aligned}$$

The first element, $(-1\$, -1\$)$, means there will be a -2 in the table. This happens 12 times. So the $(-1, -1)$ element of the matrix is 12.

The element $(-1\$, 0\$)$ corresponds to -1s in the table, but only those that are owed to Joe. These are 12 of the 24 -1s. The other 12 are for element $(0\$, -1\$)$.

Of the 36 0s in the table, 12 lead to $(-1\$, 1\$)$, 12 lead to $(1\$, -1\$)$ and 12 lead to $(0\$, 0\$)$.

Completing this argument with the remaining positive elements, $(1\$, 0\$), (0\$, 1\$)$ and $(1\$, 1\$)$, we get

owes Joe	owes Sue Joe\Sue	$-1\$$ 1	$0\$$ 1	$1\$$ 1
$-1\$$	1	12	12	12
$0\$$	1	12	12	12
$1\$$	1	12	12	12

I've shown the uniform histograms for Joe and Sue at the edges. You can sort of see how they can be generated from the joint histogram in the matrix.

Let's normalize the histograms into distributions (divide each matrix element by $108 = 3 \times 36$, their sum).

owes Joe	owes Sue	−1$	0$	1$
	Joe\Sue	1/3	1/3	1/3
−1$	1/3	1/9	1/9	1/9
0$	1/3	1/9	1/9	1/9
1$	1/3	1/9	1/9	1/9

Now we see that the *marginal* distributions, the original uniform distributions for Joe and Sue, are just the row-wise and column-wise sums over the joint distribution.

We can now see the sum of histograms from Note 2 as an amalgamation of the joint distribution and the dollar contributions

	Sue	1/3	1/3	1/3
Joe	1/3	1/9	1/9	1/9
	1/3	1/9	1/9	1/9
	1/3	1/9	1/9	1/9

	Sue	−1$	0$	1$
Joe	−1$	−2$	−1$	0$
	0$	−1$	0$	1$
	1$	0$	1$	2$

The resulting sum uses the joint distribution as weights in counting the occurrences. Since the weights are all the same in this case, we get what we had before.

owes Joe or Sue	−2$	−1$	0$	1$	2$
# ways	1	2	3	2	1

There are other joint distributions which also sum to uniform marginal distributions. Here is one for exclusive lenders/borrowers/neutrals. You can construct it by eliminating all but the diagonal in the Sue *abc*/Joe *abc* table above.

	Sue *abc* =	–101	–110	0–11	01–1	1–10	10–1
Joe *abc* =	–101	–202					
	–110		–220				
	0–11			0–22			
	01–1				02–2		
	1–10					2–20	
	10–1						20–2

getting

	Sue	1	1	1
Joe	1	6		
	0		6	
	1			6

i.e.,

	Sue	1/3	1/3	1/3
Joe	1/3	1/3		
	1/3		1/3	
	1/3			1/3

Using this as weights for the dollar combination, we get the new sum of the Joe and Sue histograms

owes Joe or Sue	−2$	−1$	0$	1$	2$
# ways	3	0	3	0	3

We get a third joint distribution if we go on to imagine that no client is willing just to transfer money between Joe and Sue, i.e., (owes Joe, owes Sue) = $(-1\$, 1\$)$ and $(1\$, -1\$)$ are excluded.

Sue abc =		–101	–110	0–11	01–1	1–10	10–1
Joe abc =	–101	–202	–211	–1–12			
	–110	–211	–220		–12–1		
	0–11	–1–12		0–22		1–21	
	01–1		–12–1		02–2		11–2
	1–10			1–21		2–20	2–1–1
	10–1				11–2	2–1–1	20–2

giving

	Sue	1	1	1
Joe	1	12	6	
	0	6	6	6
	1		6	12

i.e.,

	Sue	1/3	1/3	1/3
Joe	1/3	2/9	1/9	
	1/3	1/9	1/9	1/9
	1/3		1/9	2/9

and

owes Joe or Sue	−2$	−1$	0$	1$	2$
# ways	2	2	1	2	2

One more: an asymmetrical variant in which only (owes Joe, owes Sue) = $(-1\$, 1\$)$ is excluded, i.e., the clients are only unwilling to transfer money from Joe to Sue but have no problem the other way.

Sue abc =		–101	–110	0–11	01–1	1–10	10–1
Joe abc =	–101	–202	–211	–1–12	–110		
	–110	–211	–220	–101	–12–1		
	0–11	–1–12		0–22		1–21	1–10
	01–1		–12–1		02–2	10–1	11–2
	1–10	0–11		1–21		2–20	2–1–1
	10–1		01–1		11–2	2–1–1	20–2

giving

	Sue	1	1	1
Joe	1	12	12	
	0	6	6	12
	1	6	6	12

i.e.,

	Sue	1/3	1/3	1/3
Joe	1/3	2/12	2/12	
	1/3	1/12	1/12	2/12
	1/3	1/12	1/12	2/12

and

owes Joe or Sue	−2$	−1$	0$	1$	2$
# ways	2	3	2	3	2

10. Correlation and co-ignorance. Matrices such as

Sue			
Joe	1/9	1/9	1/9
	1/9	1/9	1/9
	1/9	1/9	1/9

or

Sue			
Joe	1/3		
		1/3	
			1/3

are just distributions, but in two dimensions.

Distributions have moments so we should be able to find means, variances, etc. from these 2-D distributions.

The mean of

	Sue	−1$	0$	1$
Joe	−1$	−1, −1	−1, 0	−1, 1
	0$	0, −1	0,0	0,1
	1$	1, −1	1,0	1,1

under

	Sue	1/3	1/3	1/3
Joe	1/3	1/9	1/9	1/9
	1/3	1/9	1/9	1/9
	1/3	1/9	1/9	1/9

would be

$$\sum \begin{pmatrix} (-1,-1) & (-1,0) & (-1,1) \\ (0,-1) & (0,0) & (0,1) \\ (1,-1) & (1,0) & (1,1) \end{pmatrix} .* \begin{pmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{pmatrix} = (0,0)$$

(where .* is MATLAB notation and means multiply the matrices element by element, not as matrices).

This is just the two original means, 0$ for Joe and 0$ for Sue. You can readily show that this is true for any joint distribution giving back the original (uniform in this case) marginal distributions.

So the 2-D mean does not give us anything new.

Let's try the 2-D variances, with a twist. The 1-D variance, say for Joe, is $\sum(J-\mu_J)^2 d_j$, where $J\epsilon\{-1\$, 0\$, 1\$\}$ and $d_j\epsilon(1/3, 1/3, 1/3)$ and similarly for Sue. (And $\mu_J = 0 = \mu_S$ in this example of centralized distributions.)

We now have d_j, the 2-D joint distribution, and $(J-\mu_J)$ and $(S-\mu_S)$. It is tempting to multiply them together instead of squaring each: $\sum(J-\mu_J)(S-\mu_S)d_{js}$, i.e.,

$$\sum \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} .* \begin{pmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{pmatrix} = 0$$

Before we try to interpret this, let's try the diagonal joint distribution.

$$\sum \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} .* \begin{pmatrix} 1/3 & & \\ & 1/3 & \\ & & 1/3 \end{pmatrix} = \frac{2}{3}$$

Compare this with the two variances (where the squares on the vectors are applied element-by-element)

$$\sigma_J^2 = \sum(J-\mu_J)^2 d_j = \sum \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix}^2 .* \sum \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} = \frac{2}{3}$$

and

$$\sigma_S^2 = \sum(S-\mu_S)^2 d_s = \sum(-1,0,1)^2 .* (1/3, 1/3, 1/3) = 2/3$$

We would call $\sum(J-\mu_J)(S-\mu_S)d_{js}$ the *covariance*, σ_{JS}^2. Normalizing by dividing by the two standard deviations we get what is called the *correlation*:

$$\rho_{JS} = \frac{\sigma_{JS}^2}{\sigma_J \sigma_S} = \frac{\sum(J-\mu_J)(S-\mu_S)d_{js}}{\sqrt{\sum(J-\mu_J)^2 d_j}\sqrt{\sum(S-\mu_S)^2 d_s}}$$

For our two joint distributions so far, the correlations are

$$\rho_{JS} = \frac{0}{\sqrt{2/3}\sqrt{2/3}} = 0$$

for

$$\sum \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} .* \begin{pmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{pmatrix}$$

and

$$\rho_{JS} = \frac{2/3}{\sqrt{2/3}\sqrt{2/3}} = 1$$

for

$$\sum \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} .* \begin{pmatrix} 1/3 & & \\ & 1/3 & \\ & & 1/3 \end{pmatrix}$$

You can check that

$$\rho_{JS} = \frac{-2/3}{\sqrt{2/3}\sqrt{2/3}} = -1$$

for

$$\sum \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} .* \begin{pmatrix} & & 1/3 \\ & 1/3 & \\ 1/3 & & \end{pmatrix}$$

Here is the interpretation. The first joint distribution does not correlate Joe and Sue at all. They are *independent* of each other. Their clients treat them independently.

The second joint distribution is maximally *correlated*: if you know Joe's lending/borrowing status then you know Sue's and vice-versa. Since their clients are exclusively borrowers or exclusively lenders or exclusively neutral, we found out that the client borrowing from Joe also borrows from Sue and so on.

The third joint distribution is maximally *anticorrelated*: you can show that each client is simply transferring money from Joe to Sue or vice-versa. Again, if you know Joe's transactions then you know Sue's and vice-versa, but they are opposite.

We can also find two-dimensional *ignorance*.

The *co-ignorance* for the distribution

$$\begin{pmatrix} 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \\ 1/9 & 1/9 & 1/9 \end{pmatrix}$$

is $-\sum \frac{1}{9} \lg \frac{1}{9} = \lg 9 = 2 \lg 3 = 3.17$.

For

$$\begin{pmatrix} 1/3 & & \\ & 1/3 & \\ & & 1/3 \end{pmatrix}$$

the co-ignorance is $-\sum \frac{1}{3} \lg \frac{1}{3} = \lg 3 = 1.58$ which is just the same as each of the marginal ignorances for Joe and for Sue.

We can adjust co-ignorance to give a measure of the dependence/independence described by the joint distribution, just as we modified covariance to correlation.

In that case we normalized by *dividing* the covariance by the two standard deviations. But ignorance is *additive* as we saw in Note 7. So we should *subtract*.

So we define *mutual ignorance* as the co-ignorance minus the two marginal ignorances, $I_{JS} - I_J - I_S$. The *mutual information* is the negative of this

$$I_{J;S} = I_J + I_S - I_{JS}$$

and be careful! I've used the same letter I for both ignorance and information: use the semicolon to flag the inversion of meaning.

For the first joint distribution (independent)

$$I_{J;S} = \lg 3 + \lg 3 - 2 \lg 3 = 0$$

For the second (fully equivalent)

$$I_{J;S} = \lg 3 + \lg 3 - \lg 3 = \lg 3$$

You can show that for the anticorrelated joint distribution

$$I_{J;S} = \lg 3 + \lg 3 - \lg 3 = \lg 3$$

—the same!

Here is the interpretation. For uncorrelated data, the joint distribution allows us to infer nothing about either distribution given the other.

For fully correlated data, the joint distribution tells us everything about either distribution given the other.

For fully anticorrelated data, the joint distribution dispels exactly the same amount of ignorance. Since ignorance derives only from the distribution and not from the underlying values, anticorrelation cannot be distinguished from correlation in terms of ignorance.

11. Conditional distributions and ignorance. What is the joint histogram of Joe's debts with the sum of Joe's or Sue's debts?

	Joe or Sue		1 −2$	2 −1$	3 0$	2 1$	1 2$
Joe	3	−1$	1	1	1		
	3	0$		1	1	1	
	3	1$			1	1	1

If 1$ is owed Joe there is one way that 2$ is owed Joe or Sue (i.e., 1$ is also owed Sue); there is one way that 1$ is owed Joe or Sue (i.e., 0$ is owed Sue); there is one way that 0$ is owed Joe or Sue (i.e., −1$ is owed Sue: Sue owes 1$); and there are no ways that either −1$ or −2$ are owed Joe or Sue. This explains the bottom line of the joint histogram. The other two lines are explained similarly.

Note that the marginal histograms are the row and column sums of the joint histogram, respectively.

The same thing for **max** is even more useful for us. I am doing this to produce final histograms with nonuniform marginals to motivate the discussion of conditional distributions.

	Joe **max** Sue		1 −1$	3 0$	5 1$
Joe	3	−1$	1	1	1
	3	0$		2	1
	3	1$			3

Each of the rows of this joint histogram is itself a histogram, and has a corresponding distribution, mean, variance and ignorance.

	Joe **max** Sue		1 −1$	3 0$	5 1$	μ	σ^2	I
Joe	3	−1$	1/3	1/3	1/3	0	2/3	$\lg 3 = 1.58$
	3	0$		2/3	1/3	1/3	2/9	$\lg 3 - 2/3 = 0.92$
	3	1$			1	1	0	0

And for sum

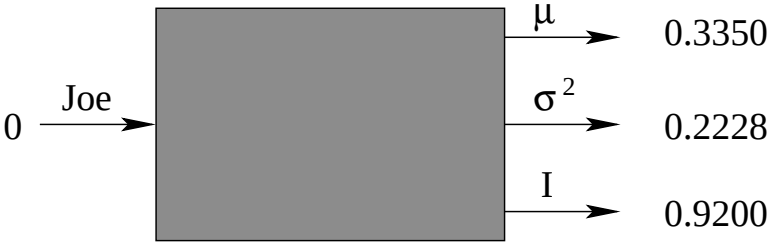
	Joe or Sue		1 −2$	2 −1$	3 0$	2 1$	1 2$	μ	σ^2	I
Joe	3	−1$	1/3	1/3	1/3			−1	2/3	$\lg 3$
	3	0$		1/3	1/3	1/3		0	2/3	$\lg 3$
	3	1$			1/3	1/3	1/3	1	2/3	$\lg 3$

Now suppose we have a "black box", which is an opaque device accepting inputs and responding with outputs but having completely hidden inner workings. Suppose we are allowed to input Joe's debt and the black box responds with mean, variance and ignorance.

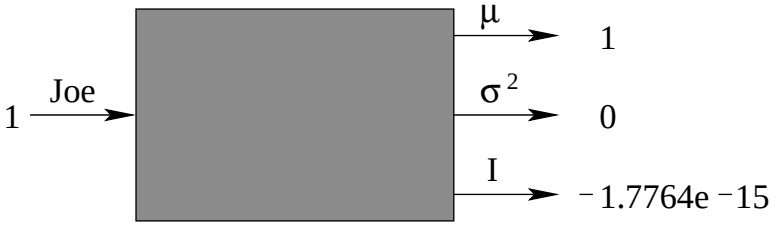
We would like to infer from the response what is going on inside the black box. Suppose

Of the two possible outputs from the **or** (+) and **max** conditional distributions corresponding to Joe owing −1$, the output numbers look closer to those for **max**: −0.0270 is sort of 0, 0.6563 is not far from 2/3, and 1.5839 is pretty close to lg 3. We might guess that the black box is generating a certain quantity of numbers uniformly distributed over the values $-1, 0$ and 1, then taking the max of the numbers it finds with the input and returning the three statistics.

A further test strengthens this hypothesis.



Finally



These distributions are called *conditional distributions* because they are conditional on the selected value of Joe's debt.

$$\begin{aligned} d_{jm}^{\mathbf{max}|J} &= d_{jm}^{J,\mathbf{max}} / d_j^J \\ &= \begin{pmatrix} 1/9 & 1/9 & 1/9 \\ & 2/9 & 1/9 \\ & & 3/9 \end{pmatrix} ./ \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} \\ &= \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ & 2/3 & 1/3 \\ & & 1 \end{pmatrix} \end{aligned}$$

The conditional distribution $d^{\mathbf{max}|J}$ is pronounced "distribution of the maximum *given* J".

There is a transpose conditional distribution

$$d_{jm}^{J|\mathbf{max}} = d_{jm}^{J,\mathbf{max}} / d_m^{\mathbf{max}}$$

$$= \begin{pmatrix} 1/9 & 1/9 & 1/9 \\ & 2/9 & 1/9 \\ & & 3/9 \end{pmatrix} ./(1/9, 3/9, 5/9)$$
$$= \begin{pmatrix} 1 & 1/3 & 1/5 \\ & 2/3 & 1/5 \\ & & 3/5 \end{pmatrix}$$

with corresponding means $(-1, -1/3, 2/5)$ (read down the columns for $d^{J|\mathbf{max}}$ as opposed to along the rows as for $d^{\mathbf{max}|J}$), variances (0,1/3,4/5) and ignorances $(0, \lg 3 - 2/3, \lg 5 - (3/5)\lg 3)$.

The ignorance associated with a conditional distribution may be called *conditional ignorance* and written more explicitly $I^{\mathbf{max}|J}$, say:

$$I_j^{\mathbf{max}|J} = \begin{pmatrix} \lg 3 \\ \lg 3 - 2/3 \\ 0 \end{pmatrix}$$

"Conditional mean" and "conditional variance" are terms not generally used.

When the black box returns mean, variance and ignorance we can uniquely determine which of $+$. $\times$,$\wedge$, **max** or **min** is being generated (see the MATLAB blackboxes program in excursion "Black boxes" for this Note). Only $+$ and $-$ are indistinguishable by these three measures.

$\star\star\star$

Now let's suppose the black box returns only means, and let's do some "Bayesian inference" on our experiment. If we start with some guess as to the likelihood of a certain cause (we'll call this guess a *hypothesis*, H) then *evidence*, E, will refine our guess through the conditional distribution $d^{H|E}$.

A black box which generates one of the sum, difference, product, max or min of the number input and an internally generated random value will give the following means for an input of -1.

	$+$	$-$	$*$	**max**	**min**
mean, μ_{-1}	-1	-1	0	0	-1

For an input of 0 the means will be

	$+$	$-$	$*$	**max**	**min**
mean, μ_0	0	0	0	1/3	1/3

And for an input of 1 the means will be

	$+$	$-$	$*$	**max**	**min**
mean, μ_1	1	1	0	1	0

Check these out!

If we know only that the black box uses these five operators we should start by granting equal likelihood to each of the five. Here is the marginal distribution (for the hypothesis as to which of the five is being used) that we can construct from this supposition.

H	$+$	$-$	$*$	**max**	**min**
d^H	1/5	1/5	1/5	1/5	1/5

Now suppose we input -1 to the black box. The possible means, μ_{-1}, are, from above, -1 and 0. We can use these, and how they are arrived at, to construct a joint distribution (for hypothesis H and evidence E):

d^{HE}	H	$+$	$-$	$*$	**max**	**min**
$E:\mu_{-1}$	-1	1/5	1/5			1/5
	0			1/5	1/5	

This joint distribution must be consistent with the marginal distribution, above, for H. It is easy to invent in this case, because the resulting means have disjoint probabilities: the mean is either -1 or 0 and cannot be both.

Finally, from the joint distribution, we have the other marginal distribution, for E.

E	d^E
-1	3/5
0	2/5

Hitherto we have been deriving both marginal distributions from the joint distribution. Here we proceed by steps from the marginal distribution for the hypothesis to the joint distribution—which is not necessarily unique—then to the marginal distribution for the evidence.

That is, d^H is prior, d^{HE} must be consistent, and d^E can be derived.

What happens next is we put a -1 into the black box and see what comes out. Then we must find the conditional distribution $d^{H|E}$ of the hypothesis given *this* evidence, so we can revise our estimate of the likelihood of the hypotheses given the evidence of the value for the mean returned by the black box.

$d^{H\|E}$	H	$+$	$-$	$*$	**max**	**min**
$E:\mu_{-1}$	-1	1/3	1/3			1/3
	0			1/2	1/2	

So, if we got -1 for μ_{-1} (i.e., for E) then we can revise d^H to

H	$+$	$-$	$*$	**max**	**min**
d^H	1/3	1/3			1/3

from

H	$+$	$-$	$*$	**max**	**min**
d^H	1/5	1/5	1/5	1/5	1/5

On the other hand, if we had got 0, we would have made a similar revision, to

H	$+$	$-$	$*$	**max**	**min**
d^H			1/2	1/2	

Let's say we got $\mu_{-1} = -1$ on our first probe of the black box, i.e., the first alternative. Then we can refine our ideas by a second probe. Say we input the number 0.

(This won't work with the program in the excursion. You'll have to write your own secret program.)

Here is the step from marginal distribution for the hypothesis to joint distribution for hypothesis and evidence. Since there are three possible results for the mean given 0 as input, $-1/3$, 0 and $1/3$, we have three rows in the joint distribution. The possibilities are again disjoint so we can go uniquely from the marginal to the joint distribution.

H	$+$	$-$	$*$	**max**	**min**
d^H	1/3	1/3			1/3

d^{HE}	H	$+$	$-$	$*$	**max**	**min**
$E:\mu_0$	$-1/3$					$1/5$
	0	$1/5$	$1/5$	$1/5$		
	$1/3$				$1/5$	

Well, I wrote it down completely, but since * and **max** could not have made the black box output -1 when -1 was input, I should have left these columns out and weighted the remaining probabilities in d^{HE} accordingly.

H	$+$	$-$	**min**
d^H	$1/3$	$1/3$	$1/3$

d^{HE}	H	$+$	$-$	**min**
$E:\mu_0$	$-1/3$			$1/3$
	0	$1/3$	$1/3$	

Then

$d^{H\|E}$	H	$+$	$-$	**min**
$E:\mu_0$	$-1/3$			1
	0	$1/2$	$1/2$	

If the evidence from this second probe was $\mu_0 = -1/3$ then we can revise

H	$+$	$-$	$*$	**max**	**min**
d^H	$1/3$	$1/3$			$1/3$

To

H	$+$	$-$	$*$	**max**	**min**
d^H					1

and we have discovered that the black box has been giving us **min**.

On the other hand, let's suppose that the alternative evidence resulted, $\mu_0 = 0$. Then the revised d^H is

H	$+$	$-$	$*$	**max**	**min**
d^H	$1/2$	$1/2$			

So we do a third probe of the black box, using 1 as input. The possible means are, we found above, 0 and 1.

d^{HE}	H	$+$	$-$	$*$	**max**	**min**
$E:\mu_1$	0			$1/5$		$1/5$
	1	$1/5$	$1/5$		$1/5$	

or, since the first two probes eliminated all but two, this should be

d^{HE}	H	$+$	$-$
$E:\mu_0$	0		
	1	$1/2$	$1/2$

and we calculate

$d^{H\|E}$	H	$+$	$-$
$E:\mu_0$	0		
	1	$1/2$	$1/2$

We've run out of probes and so we're left with being unable to distinguish + from − using means. (We saw earlier that variances or ignorances cannot distinguish them either.) So we stop, But we have refined the hypothesis likelihoods significantly.

$\star\star\star$

We do not always have the luxury of being able to figure out d^{HE}. Sometimes we know only $d^{E|H}$ and must derive $d^{H|E}$ from this.

"Bayes' theorem" says

$$d^{H|E} = d^{E|H}. \times d^{H}./d^{E}$$

where I am using MATLAB notation to indicate multiplication and division element-by-element, rather than of whole matrices.

Furthermore we also might not know d^{E}, say. But we can show

$$d^{E} = d^{E|H} d^{H}$$

where this time the multiplication *is* matrix multiplication.

I'll illustrate both of these assertions using the first example of this Note, where H is **max** and E is J. (Because $d^{\mathbf{max}J}$ is not as clearly the joint distribution as is d^{HE} where we have only single letters, I'll write the joint distribution with a comma: $d^{\mathbf{max},J}$.)

$$\begin{aligned} d^{\mathbf{max},J} &= d^{J|\mathbf{max}}. * d^{\mathbf{max}} \\ &= \begin{pmatrix} 1 & 1/3 & 1/5 \\ & 2/3 & 1/5 \\ & & 3/5 \end{pmatrix} . * (\frac{1}{9}, \frac{3}{9}, \frac{5}{9}) \\ &= \begin{pmatrix} 1/9 & 1/9 & 1/9 \\ & 2/9 & 1/9 \\ & & 3/9 \end{pmatrix} \end{aligned}$$

Then

$$\begin{aligned} d^{\mathbf{max}|J} &= d^{\mathbf{max},J}./d^{J} \\ &= \begin{pmatrix} 1/9 & 1/9 & 1/9 \\ & 2/9 & 1/9 \\ & & 3/9 \end{pmatrix} ./ \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} \\ &= \begin{pmatrix} 1/3 & 1/3 & 1/3 \\ & 2/3 & 1/3 \\ & & 1 \end{pmatrix} \end{aligned}$$

Which illustrates Bayes' theorem.

To illustrate the second assertion from this same example

$$\begin{aligned} d^{J} &= d^{J|\mathbf{max}} d^{\mathbf{max}} \\ &= \begin{pmatrix} 1 & 1/3 & 1/5 \\ & 2/3 & 1/5 \\ & & 3/5 \end{pmatrix} . \begin{pmatrix} 1/9 \\ 3/9 \\ 5/9 \end{pmatrix} \\ &= \begin{pmatrix} 1/3 \\ 1/3 \\ 1/3 \end{pmatrix} \end{aligned}$$

These two assertions form the basis of "Bayesian inference". Let's apply them to a classical problem of disease diagnosis (adapted on 09/9/30 from en.wikipedia.org/wiki/Bayesian-inference).

Suppose we are testing (E) for a disease (H) which every tenth person has.

H	d	d'
d^H	1/10	9/10

Suppose statistics are known to say that if you have disease (d) your test will be positive p nine times out of 10, and if you don't have it (d') your test will be positive two times out of ten.

$d^{E\|H}$	H	d	d'
E	p	9/10	2/10
	n	1/10	8/10

(Of course, the test coming out negative—which is good news for you but apparently bad news for the medical profession—is complementary to it coming out positive. We see here the typical $d^{E|H}$, that the columns, but not the rows, sum to 1.)

Now

$$\begin{aligned} d^E &= d^{E|H} d^H \\ &= \begin{pmatrix} 9/10 & 2/10 \\ 1/10 & 8/10 \end{pmatrix} \begin{pmatrix} 1/10 \\ 9/10 \end{pmatrix} \\ &= \begin{pmatrix} 27/100 \\ 73/100 \end{pmatrix} \end{aligned}$$

and

$$\begin{aligned} d^{H|E} &= d^{E|H} .* d^H ./ d^E \\ &= \begin{pmatrix} 9/10 & 2/10 \\ 1/10 & 8/10 \end{pmatrix} .* (\frac{1}{10}, \frac{9}{10}) ./ \begin{pmatrix} 27/100 \\ 73/100 \end{pmatrix} \\ &= \begin{pmatrix} 9/100 & 18/100 \\ 1/100 & 72/100 \end{pmatrix} ./ \begin{pmatrix} 27/100 \\ 73/100 \end{pmatrix} \\ &= \begin{pmatrix} 1/2 & 2/3 \\ 1/73 & 72/73 \end{pmatrix} \end{aligned}$$

(Now the rows, but not the columns, sum to 1.)

Let's think about "false positives". That is, the element (p, d'), which gives the probability that you do not have the disease but get a positive result anyway.

Looking at $d^{E|H}$ you might think this is 20%. But $d^{H|E}$ tells us it is 67%.

"False negatives" on the other hand are much less likely than you might have supposed: 1/73 instead of 10%.

These two counterintuitive results stem from the relative rarity of the disease. The rarer it is the more extreme the discrepancy.

Bayesian inference is controversial. It can be used to refine a vague guess as to a distribution into

something more precise, but beware GIGO (garbage in, garbage out).

12. A gas simulation 1: the collisions
13. A gas simulation 2: statistics
14. The Boltzmann and Maxwell distributions.
15. Fluctuations, variations and samples.
16. Entropy.
17. Temperature.
18. Pressure.
19. State function for monatomic gases.
20. Thermostatic equations of state.
21. More on multivariate slopes.
22. Work and heat.
23. Correlation.
24. Collision theory.
25. Mobility, diffusivity and Brownian motion.
26. Three potentials and dissipation.

27. Active transport and biochemistry.
28. Combined transport.
29. Phase transitions.
30. Phase transitions in random graphs.
31. Point-to-point resistance in a network.
32. Van der Waals.
33. Sublimation.
34. Ferromagnets.
35. Particle individuality and Bose-Einstein condensation.

II. **The Excursions**
You've seen lots of ideas. Now *do* something with them!

1. My sources for the two stories in Note 1 are [vB98, pp.157–8] and [Hoy60, p.211]. The first is a wonderful book on the history and ideas behind thermodynamics. The second is science fiction by an outstanding physicist and cosmologist. Look them up. What does Alexandrov, the character in *The Black Cloud* who makes the argument about the golf ball, have to say earlier on about correlation versus prediction in science?

2. When somebody puts a hair on the diary cover to detect if a sibling has been snooping, is the hair-and-diary an ordered system?

3. a) What is the sum of the temperature histogram with itself? The difference? The product?
   b) What is the sum of

# ways	1	1	1
value	1	2	3

with itself?

c) What is the sum of

# ways	1	1	1
value	−2$	0$	2$

and

# ways	1	1	1
value	1$	2$	3$

?

4. Here is a different take on the product of histograms. How does this relate to the sum of histograms in Note 2?

	*	1	1	1
	*	$\frac{1}{2}$	1	2
1	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$	1
1	1	$\frac{1}{2}$	1	2
1	2	1	2	4

and

	*	1	2	3	2	1
	*	$\frac{1}{4}$	$\frac{1}{2}$	1	2	4
1	$\frac{1}{4}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	1
2	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	1	2
3	1	$\frac{1}{4}$	$\frac{1}{2}$	1	2	4
2	2	$\frac{1}{2}$	1	2	4	8
1	2	1	2	4	8	16

If there are enough of these, what is the limiting distribution?

5. Not all distributions are symmetric. The *binomial* distribution concerns positive values only. Here, for two arbitrary probabilities, p and q, which must sum to 1, is the beginning of its construction (by histogram addition). (I've picked values 1 and 2 arbitrarily.)

	+	p	q
	+	1	2
p	1	2	3
q	2	3	4

*		p	q
		1	2
p	1	p^2	pq
q	2	pq	q^2

a) What is the resulting histogram?

b) Explore further sums of this histogram with itself.

c) What about differences?

6. What is the sequence of # *ways* arising from a repeated sum with itself of *two* equally likely outcomes? Show that it gives Pascal's triangle (Week ii, Note 6).
What is the equivalent of the summing rule used to build Pascal's triangle for the 3-way (uniform) histogram of Note 2? For a 4-way histogram? What do the rows sum to for the 3-way histogram? The 4-way?

7. **histogArith(hist1,op,hist2).** Write a MATLAB program to calculate the sum of two arbitrary histograms, and to plot the result. We will need this in Note 6, but you can start working on it after working through Note 2.
Hint. Use an extended histogram, which includes zero entries in the # ways row, to accumulate results and for the plot. Then compress it to eliminate the zero entries. E.g.,

# ways	2	3	1
value	−4	0	1

+

# ways	1	4	2
value	1	2	3

gives an extended histogram from −3 (= −4 + 1) to 4 (= 1 + 3). This is initialized to zeros, then products of # ways are added for each of the second set of values, for each of the first set of values at a time.

value		−3	−2	−1	0	1	2	3	4
# ways		0	0	0	0	0	0	0	0
2	−4	2	8	4					
3	0				3	12	6		
1	1					1	4	2	
# ways		2	8	4	0	3	13	10	2

⇓

# ways	2	8	4	3	13	10	2
value	−3	−2	−1	1	2	3	4

(Adaptations for the ranges of the extended histogram will allow this to work for subtraction, multiplication, max and min of histograms, too. What other histogram operations can you invent?)

8. What is Δv in Note 3 for converting a distribution of 7 values, over -1 to 1, to the continuous density?

9. What is the difference between the mean of the numbers in the two sets {1,2,3} and {1,4}, and the mean of the means in each set? What if the second set were {1,3,4}? {1,3}?

10. For the distribution

1/4	1/4	1/2
-1	0	1

show that the variance $\sigma^2 = 3/4$, the coefficient of skewness is $2/(3\sqrt{3})$ and the kurtosis is $-5/3$.

11. Confirm that $\text{slope}_v(1 + v + v^2/2! + v^3/3! + ..) = 1 + v + v^2/2! + v^3/3! + ..$ and so that $\text{antislope}_v(1 + v + v^2/2! + v^3/3! + ..) = 1 + v + v^2/2! + v^3/3! + ..$

12. a) Using the moment generating function e^{tv} show that the $k-1$st moment of the uniform density $1/(b-a)$ from a to b is

$$\frac{b^k - a^k}{k(b-a)}$$

and hence $\mu = (b-a)/2$ and $\sigma^2 + \mu^2 = (b^2 + ba + 2a)/3$ (so $\sigma^2 = (b-a)^2/12$).
b) From this show that the density is uniquely determined by the first two moments and in particular by μ and σ.
c) Work out the $k-1$st moment in terms of μ and σ.

13. **Mean, median, mode.** The mean, μ, is a measure of "central tendency"—an indication of where the centre of the distribution is.
Two other important measures of central tendency are the *mode* and the *median*. The mode is the location of the peak of the distribution. The median is the location of the middle element of the distribution. Note that mode and median both occupy actual values of a discrete distribution, while the mean may be any number.
Draw the histogram

occurrences	2	1	1	1
value	2	3	10	18

and show for it that

mode	median	mean
2	3	7

If these were salaries, in hundreds of dollars per week, would it be correct to tell a prospective employee that the average salary is 700 $ per week? Given that the CEO makes 1800 $, the chair of the board makes 1000 $, the secretary makes 300 $ and the workers make 200 $ each, would it be *fair* to tell the prospective worker that the average is 700 $? What would be the honest answer to the inquiry?
Look up the story by Martin Gardner [Gar82, p.114] highlighting the "paradox" of "average". ("Median" generalizes to "quantile", in that the median is the 1/2 quantile: the sum of the histogram occurences up to value 3, but not including it, is 2; the sum above it is also 2, so

value 3 is the middle. The sum up to 10 but not including it is 3 so the 3/4 quantile ("3rd quartile") occurs at value 10. (This notion is imprecise, but more precise for continuous distributions, i.e., densities [MGB74, p.73].) An equivalent to "quantile" is "percentile", with appropriate scaling: the median is the 50th percentile.)

14. **Geometric and harmonic means.** The mean we've discussed so far is the *arithmetic mean* of n numbers

$$\mu_{\rm A} = \frac{1}{n}\sum_{j=1}^{n} v_j$$

The *geometric mean*

$$\mu_{\rm G} = (\prod_{j=1}^{n} v_j)^{1/n}$$

tells, for instance in the case $n = 2$, the side of a square having the same area as a rectangle of sides v_1 and v_2, or in the case $n = 3$, the side of the cube having the same volume as a rectangular hexahedron of sides v_1, v_2 and v_3.
The *harmonic mean*

$$\mu_{\rm H} = \frac{n}{\sum_{j=1}^{n} \frac{1}{v_j}}$$

is used for rates and ratios.
a) What is the average rate of interest over three years if $v_1 = 5\%, v_2 = 8\%$ and $v_3 = 3\%$ for each of three successive years, respectively? By how much does this differ from the arithmetic mean?
b) If you drive 10 km at 50 km/hr then 10 more km at 100 km/hr, what is your average speed?

15. Show that the matrix $M = \sum_k w_k \vec{v_k} \vec{v_k^{\rm T}}$, for any set of orthonormal vectors, $\vec{v_k}$, has $\vec{v_k}$ as eigenvectors (Week 9 Note 1) with w_k as corresponding eigenvalues.

16. For any matrix $\mathcal{A}$ and any normalized vector $\vec{v}$ show that

$$\vec{v^{*\rm T}} = {\rm Tr}(\vec{v}\vec{v^{*\rm T}}\mathcal{A}) = {\rm Tr}(\rho\mathcal{A})$$

for the density matrix $\rho = \vec{v}\vec{v^{*\rm T}}$.

17. What are the expected values for rotation by angle β about the z-axis for state $(1/\sqrt{3},\ 1/\sqrt{3},\ 1/\sqrt{3})$? $(1/\sqrt{4},\ 1/\sqrt{2},\ 1/\sqrt{4})$? $(i/\sqrt{2},\ 0,\ -i/\sqrt{2})$?

18. In a (1-number) state (p, q), what are the expected values of the three Pauli matrices (Week 6 Note 5)?

19. Are the sums and products of density matrices themselves density matrices? Hint: multiply $(I + \sigma_x)/2$ and $(I + \sigma_y)/2$ where σ_x and σ_y are Pauli matrices (Week 6 Note 5).

20. Density matrices were proposed by John von Neumann in his book [vN55]. We're not ready for that book yet but you can check out some of the many stories about him. Are the following two true or apocryphal?
a) During the Manhattan project, Fermi, von Neumann and Feynman were doing a calculation together. Feynman took to his calculator. Fermi whipped out the slide rule he always carried. Von Neumann looked at the ceiling. Von Neumann got the answer first.
b) Two trains approach each other (each on its own track) each doing 40 km/hr. When they were 1 km apart a bee left the front of one, flew to the other at 120 km/hr, turned immediately on encountering the second, flew back to the first at 120 km/hr, and kept back

and forth at this speed until the moment the fronts of the trains passed each other. How many km did the bee fly?
This problem can be solved in two ways. A physicist would supposedly calculate the time it takes for the trains to meet then calculate how far the bee can fly in that time. This can be done mentally. A mathematician, again supposedly, would find the series and sum it, an exercise usually needing pencil and paper.
When tested with this problem, von Neumann looked at the ceiling then promptly gave the correct answer. "Aha! You're a physicist at heart." But when the reason for this conclusion was explained, von Neumann said "Oh, no, I summed the series."

21. I really cheated in arguing in Note 6 that the moment generating function of the sum of m distributions tends to $e^{\sigma^2 t^2/2}$ because I omitted higher powers of t in $(1+\sigma_1^2 t^2/2!+..)^m$. Show for m *uniform* distributions that these higher powers of t are all divided by higher powers of m and so can be neglected.

22. **Notation.** a) The argument showing $I^2 = \pi a$ in Note 6 was quite long, involving careful thinking at each step. The length was because of this careful thinking. I'd like to show you how good notation actually captures operations of thought and can make the argument much shorter.
The slope/antislope notation I've been using is my own, and is not as good as the standard notation in calculus. Its advantage is that it requires explicit thought at each step, and this is very important for beginners. But it is a crutch which you can eventually throw away.
Standard notation writes $\frac{\mathrm{d}f(x)}{\mathrm{d}x}$ for $\mathrm{slope}_x f(x)$ and $\int f(x)\mathrm{d}x$ for $\mathrm{antislope}_x f(x)$. Here is the $I^2 = \pi a$ argument from Note 6 in standard notation. (It is a slight adaptation of the footnote in [FLS64, p.40-6].)

$$\begin{aligned} I &= \int_{-\infty}^{\infty} e^{-x^2/a}\mathrm{d}x \\ I^2 &= (\int_{-\infty}^{\infty} e^{-x^2/a}\mathrm{d}x)(\int_{-\infty}^{\infty} e^{-y^2/a}\mathrm{d}y) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/a}\mathrm{d}x\mathrm{d}y \\ &= \int_0^{\infty} e^{-r^2/a}2\pi r\mathrm{d}r \\ &= -\pi a\int_0^{\infty} e^{-s}\mathrm{d}s \\ &= \pi a \end{aligned}$$

Only the swap of sum and product, the change of variables from x and y to r and the subsequent change from r^2/a to s are given explicitly in this version of the argument. The Fundamental Theorem of Calculus is embedded in the notation, especially in the appearance of $\mathrm{d}x$, $\mathrm{d}y$, $\mathrm{d}r$ and $\mathrm{d}s$ in the "integrals" (antislopes). We don't even see the use of $\frac{\mathrm{d}f(x)}{\mathrm{d}x}$ as the slope, but it is implicit in $\mathrm{d}s = 2r\mathrm{d}r$.
The importance of good notation is captured by Alfred North Whitehead

> Civilization advances by extending the number of important operations which we can perform without thinking about them. Operations of thought are like calvary charges in battle—they are strictly limited in number, they require fresh horses, and must only be made at decisive moments.

quoted by Strachey [Str66]. Strachey never explicitly cites this introductory quotation nor says why he quotes it, but the implication of the article is that a good notation, and in particular a good programming language, captures past operations of thought thus saving us

repeating them.
Without looking back at the text, use the above to reproduce the argument of Note 6.
b) Language is also, among other aspects, notation, and grammar can incorporate important knowledge. For example "I teach you" is grammatical; "I learn you" is not. "Teach" is a transitive verb, but the intransitivity of "learn" (strictly speaking, *ambitransitivity*: "learn" can also take an object as in "learn French") indicates the gap between teaching and learning: you must do the learning yourself. "Learn" is also an active verb, and the passive form, "you were learned by me", is not grammatical.
Is the grammar for "to learn" correct? Try to find counterexamples to the idea that you can only learn by doing. When did you ever really learn by not doing anything?
(Syntax cannot be counted on to correct every error. It allows oxymorons such as "win the war" and "unshared language", and could never eliminate one-word oxymorons such as "object" and "jihad". Could some of this be fixed?)

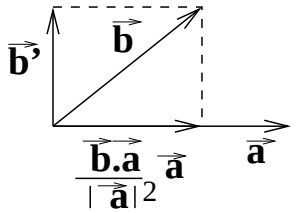
23. For a non-central normal distribution the density is

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(v-\mu)^2/(2\sigma^2)}$$

Work out the first four moments, showing in particular that μ is the mean and σ^2 the variance.

24. **Legendre and Hermite polynomials.** When there is a moment generating function, the moments completely determine the density or distribution. This is reminiscent of the way the Fourier transform (Week 9) completely determines the original function.
The Fourier transform maps a function written in terms of position, x, say, to a representation in terms of frequency, ω.
The moments might be thought of as describing the density in terms of the basis $1, x, x^2, x^3, ..$ This is not exactly so, because the moments in fact are the antislopes over a fixed range of the product of density and $1, x, x^2, x^3, ..$
But the idea is suggestive and I'd like to explore in this excursion the "basis" $1, x, x^2, x^3, ..$ of an infinite-dimensional space of functions.
The most important operation on a set of vectors, $\vec{a}, \vec{b}, ..$ which may form a basis of a "vector space" is the inner or dot product, $\vec{a} \bullet \vec{b}, ...$ This determines if the vectors are orthogonal to each other, in which case, $\vec{a} \bullet \vec{b} = 0$. In general, $\vec{a} \bullet \vec{b}$ is the length of the *projection* of $\vec{a}$ onto $\vec{b}$ and of $\vec{b}$ onto $\vec{a}$.
It also allows us to construct orthogonal vectors if they are not already orthogonal.
Starting with $\vec{a}$ as a not necessarily normalized basis vector we would like to map $\vec{b}$ into $\vec{b'}$ orthogonal to $\vec{a}$.
We do this by subtracting from $\vec{b}$ its projection on $\vec{a}$

$$\vec{b'} = \vec{b} - \frac{\vec{a} \bullet \vec{b}}{|\ \vec{a}\ |^2}\vec{a}$$



If the space has higher dimensions, so that there is a vector $\vec{c}$ not coplanar with $\vec{a}$ and $\vec{b'}$, we can find $\vec{c'}$ orthogonal to both $\vec{a}$ and $\vec{b'}$ by a similar process.

$$\vec{c'} = \vec{c} - \frac{\vec{c} \bullet \vec{b'}}{|\ \vec{b'}\ |^2}\vec{b'} - \frac{\vec{c} \bullet \vec{a}}{|\ \vec{a}\ |^2}\vec{a}$$

And so on.

$$\vec{d'} = \vec{d} - \frac{\vec{d} \bullet \vec{c'}}{|\ \vec{c'}\ |^2}\vec{c'} - \frac{\vec{d} \bullet \vec{b'}}{|\ \vec{b'}\ |^2}\vec{b'} - \frac{\vec{d} \bullet \vec{a}}{|\ \vec{a}\ |^2}\vec{a}$$

This process is called Gram-Schmidt orthogonalization.
We are going to find that $1, x, x^2, x^3, ..$ are not "orthogonal" to each other. We will need to define an "inner product" $< f(x), g(x) >$ of two *functions* $f(x)$ and $g(x)$, akin to $< \vec{a}, \vec{b} >= \vec{a} \bullet \vec{b}$ for vectors.
The vector inner product just sums the product of coefficients

$$\vec{a} \bullet \vec{b} = \sum_j a_j b_j$$

a) So it is plausible that a functional inner product will be given by the area under the product of the functions

$$\begin{aligned} < f(x), g(x) > &= \sum_x f(x) g(x) \Delta x \\ &= \text{antislope} f(x) g(x) \end{aligned}$$

Since the inner product is a number, not a function itself, we should take this antislope over a definite range of values of x.
Let's start with the range -1 to 1, which we used for uniform distributions.

$$< 1, x >= \text{antislope}_x 1 \times x \mid_{-1}^{1} = \left. \frac{x^2}{2} \right|_{-1}^{1} = 0$$

So 1 and x are orthogonal according to the inner product.
Similarly, x and x^2 are orthogonal.

$$< x, x^2 >= \text{antislope}_x x \times x^2 \mid_{-1}^{1} = \left. \frac{x^4}{4} \right|_{-1}^{1} = 0$$

Show that $< j \mid k >=< x^j, x^k >= 0$ for any $j + k$ odd.
But 1 and x^2 are not orthogonal.

$$< 1, x^2 >= \text{antislope}_x 1 \times x^2 \mid_{-1}^{1} = \left. \frac{x^3}{3} \right|_{-1}^{1} = \frac{2}{3}$$

So we need to make a function out of x^2 which is orthogonal to both 1 and x. We'll call it $L_2(x)$.
(We'll let $L_0(x) = 1$ and $L_1(x) = x$.)

$$\begin{aligned} L_2(x) &= x^2 - \frac{< x^2, L_1(x) >}{< L_1(x), L_1(x) >} L_1(x) - \frac{< x^2, L_0(x) >}{< L_0(x), L_0(x) >} L_0(x) \\ &= x^2 - \frac{< x^2, x >}{< x, x >} x - \frac{< x^2, 1 >}{< 1, 1 >} 1 \\ &= x^2 - 0x - \frac{2/3}{2} 1 \\ &= x^2 - \frac{1}{3} \end{aligned}$$

where you should show that $< 1, 1 >= 2$.
Two changes before we go on: first, we'll drop the (x) after the functions L_j, L_k in $< L_j, L_k >$,

just to keep it shorter; second we'll stop writing every second term, which you should show always has $j+k$ odd and so is zero.

$$\begin{aligned}
L_3(x) &= x^3 - \frac{< x^3, L_1 >}{< L_1, L_1 >} L_1(x) \\
&= x^3 - \frac{\text{antislope } x^3 \times x \mid_{-1}^{1}}{\text{antislope } x \times x \mid_{-1}^{1}} x \\
&= x^3 - \frac{x^5/5 \mid_{-1}^{1}}{x^3/3 \mid_{-1}^{1}} x \\
&= x^3 - \frac{3}{5} x \\
L_4(x) &= x^4 - \frac{< x^4, L_2 >}{< L_2, L_2 >} L_2(x) - \frac{< x^4, L_0 >}{< L_0, L_0 >} L_0(x) \\
&= x^4 - (\frac{x^7/7 - x^5/15) \mid_{-1}^{1}}{(x^4/5 - 2x^3/9 + x/9) \mid_{-1}^{1}} (x^2 - \frac{1}{3}) - \frac{x^5/5 \mid_{-1}^{1}}{2} 1 \\
&= x^4 - \frac{6}{7} x^2 + \frac{6}{21} - \frac{1}{5} \\
&= x^4 - \frac{6}{7} x^2 + \frac{3}{35}
\end{aligned}$$

These calculations will get larger and larger because L_k will have k projections on $L_{k-1}, .., L_0$ to be subtracted. (Well, half of them are zero, but it's still long.) Let's see if there is a shortcut.
Try

$$\begin{aligned}
L_k &= x L_{k-1} - c_k L_{k-2} \\
L_2 &= x L_1 - c_2 L_0 \\
&= x^2 - c_2 \\
L_3 &= x L_2 - c_3 L_1 \\
&= x(x^2 - \frac{1}{3}) - c_3 x \\
&= x^3 - (c_3 + \frac{1}{3}) x \\
L_4 &= x L_3 - c_4 L_2 \\
&= x(x^3 - \frac{3}{5} x) - c_4 (x^2 - \frac{1}{3}) \\
&= x^4 - (c_4 + \frac{3}{5}) x^2 + \frac{c_4}{3}
\end{aligned}$$

By comparing these with what we found previously we see

$c_2 = \frac{1}{3}$	which we used in L_3
$c_3 = \frac{3}{5} - \frac{1}{3} = \frac{2^2}{5 \times 3}$	which we used in L_4
$c_4 = \frac{3^2}{7 \times 5}$	which we could use in L_5

and here we spot a pattern.
If

$$c_k = \frac{(k-1)^2}{(2k-1)(2k-3)}$$

then we have a much shorter way of finding all the $L_k(x)$.
These are a variant of the *Legendre polynomials* that are normalized so that the leading

coefficient is 1.
[Wei09] tells us that, starting with

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \end{aligned}$$

and given a polynomial inner product $< f, g >=< f(x), g(x) >$, then

$$P_{k+1}(x) = (x - \frac{< xP_k, P_k >}{< P_k, P_k >})P_k(x) - \frac{< P_k, P_k >}{< P_{k-1}, P_{k-1} >}P_{k-1}(x)$$

This is a recurrence relation involving only the previous two polynomials, instead of all $k+1$ previous. Check that it works for the Legendre polynomials.
b) If we extend the range to $-\infty$ to ∞, which is used in the normal distribution, simple integration will give infinite answers: try, say,

$$\text{antislope } x^2 \mid_{-\infty}^{\infty}$$

So we must modify our definition of the inner product to unclude a *weight function* which will keep things small out at the extremes. To beat down a polynomial we'll need an exponential, and to beat it down in both positive and negative directions, it must be exponential in x^2. Try $e^{-x^2/2}$:

$$< f, g >=< f(x), g(x) >= \text{antislope } f(x)g(x)e^{-x^2/2} \mid_{-\infty}^{\infty}$$

Now we need the calculus we've developed in Note 6: for present calculations we can just set $\sigma = 1$.

$$< 1, x >= \text{antislope } 1 \times xe^{-x^2/2} \mid_{-\infty}^{\infty} = 0$$

and $< j \mid k >=< x^j, x^k >= 0$ for $j + k$ odd, as we saw in Note 6.

$$< x, x >=< 1, x^2 >= \text{antislope } x^2e^{-x^2/2} \mid_{-\infty}^{\infty} = \sqrt{2\pi}$$

So 1 and x^2 are not orthogonal, again. We must create a polynomial orthogonal, under the new definition, to both 1 and x. We'll call it $H_2(x)$ (with $H_0(x) = 1, H_1(x) = x$).

$$\begin{aligned} H_2(x) &= x^2 - \frac{< x^2, H_1 >}{< H_1, H_1 >}H_1(x) - \frac{< x^2, H_0 >}{< H_0, H_0 >}H_0(x) \\ &= x^2 - \frac{< x^2, x >}{< x, x >}x - \frac{< x^2, 1 >}{< 1, 1 >}1 \\ &= x^2 - 0x - \frac{\text{antislope } x^2e^{-x^2/2} \mid_{-\infty}^{\infty}}{\text{antislope } e^{-x^2/2} \mid_{-\infty}^{\infty}}1 \\ &= x^2 - 1 \\ H_3(x) &= x^3 - \frac{< x^3, H_1 >}{< H_1, H_1 >}H_1(x) \\ &= x^3 - \frac{\text{antislope } x^4e^{-x^2/2} \mid_{-\infty}^{\infty}}{\text{antislope } x^2e^{-x^2/2} \mid_{-\infty}^{\infty}}x \\ &= x^3 - \frac{3\sqrt{2\pi}}{\sqrt{2\pi}}x \\ &= x^3 - 3x \\ H_4(x) &= x^4 - \frac{< x^4, H_2 >}{< H_2, H_2 >}H_2(x) - \frac{< x^4, H_0 >}{< H_0, H_0 >}H_0(x) \\ &= x^4 - 6x^2 + 3 \end{aligned}$$

Check that $H_k(x) = xH_{k-1}(x) - (k-1)H_{k-2}(x)$
Check that this is also given by the [Wei09] recurrence relationship.
These are the variant of the *Hermite polynomials* that are normalized so the leading coefficient is 1. We saw other variants in Book 8c Note 39.
Both the Legendre and the Hermite polynomials form orthogonal bases for infinite-dimensional function spaces. They have different ranges $(-1:1, -\infty:\infty)$ and weight functions $(1, e^{-x^2/2})$, respectively.
c) Other ranges and weight functions lead to other sets of orthogonal polynomials. Look some up.

25. What functions other than negative logarithms have values at 0 and 1 that would make them suitable for surprisal?

26. What is the surprisal of each of 27 values uniformly distributed?

27. Where is the maximum of $-x \ln x$? $-x \lg x$? Plot them on your calculator and also find the max.

28. Just as the curtain rises on the dancers, two and a half minutes into Igor Stravinski's "Le Sacre du Printemps", he sounds an arpeggio of four E-minor notes with the conventional 4/4 rhythm **1** 2 **3** 4. Twenty seconds later, these notes are played as a chord along with the corresponding E-major chord, with the surprising rhythm **1** 2 3 4 5 6 7 8. This is immediately followed by more repetitions of the chord with the emphasis on the second beat and further surprises 1 **2** 3 **4** 5 6 7 8 / 1 **2** 3 4 **5** 6 7 8. We would now expect 1 **2** 3 4 5 **6** 7 8 but instead we next hear **1** 2 3 4 5 **6** 7 8. Discuss this development in terms of surprisal.
(This astute musical analysis must make me look pretty good, so I'll say I got it all from Robert Harris on Michael Enwright's Canadian Broadcasting Corporation "Sunday Edition" on Sept. 27 2009. I had to run my recording into a visual display before I could count the rhythms. Harris went on to point out that the 1913 premier of this ballet, which triggered a riot in the audience, anticipated the outbreak of World War I fifteen months later, which the artist in Stravinski foresaw and which probably accounts for the violence of the reception.)

29. "It's a crime for a man to go philandering/and fill his wife's poor heart with grief and doubt/.." I originally wanted to use "doubt" for "ignorance", so invent an example for this and analyse it in terms of "doubt", i.e., ignorance in the technical sense. The rest of this verse is also apposite. Dig up George Bernard Shaw's Cinderella story "Pygmalion" in its musical incarnation "My Fair Lady", particularly the performance of the song by Stanley Holloway.

30. a) An abstraction throws away information considered to be irrelevant. What does this do to ignorance?
b) A theory could be thought of as a compressing of all the empirical data that could be deduced from it. Does compression change ignorance? Consider both lossy and lossless compression.
c) Speculate, in terms of ignorance increasing, about the immolation of Black Clouds in Fred Hoyle's novel of the same name when they get too close to solving the "deep problems" of the universe [Hoy60, pp.203–4].

31. If Stu never removes a tool (or always replaces it before Dad gets home), what is Dad's ignorance (Note 8)?

32. Occasionally Stu is on his way to his room with the pliers when Mom calls "Drop everything! Supper's ready". So he drops them in the hall. What happens to Dad's ignorance? Suppose, say, that this happens once a week, with the other likelihoods, S and W, remaining equal to each other.

33. Mom had only an arithmetic calculator with her on the train, so she had to calculate $\beta = \ln(x)$ for $x = 1.2874$ (the last step in Mom's exercise of calculating the probablities in Note 8) by using

$$\ln(1+q) = q - \frac{q^2}{2} + \frac{q^3}{3} - \frac{q^4}{4} + ..$$

She even had to remind herself of this series by showing that it is the series that inverts

$$e^r = 1 + r + \frac{r^2}{2!} + \frac{r^3}{3!} + \frac{r^4}{4!} + ..$$

(That one is easy to remember because $\text{slope}_r e^r = e^r$.)
Having read these three steps, put this text aside and reproduce Mom's calculation on your own.

34. **Increasing ignorance.** This excursion shows ignorance increasing as Dad forgets details from Note 8. Consider three independent tools, starting with frequencies

	W	S
1	1/6	5/6
2	2/3	1/3
3	1/4	3/4

a) Show that the combined frequencies for demerit points are

0	1	2	3
(3W)	(2W1S)	(1W2S)	(3S)
2/72	17/72	38/72	15/72

for an average demerit of 23/12 and a ignorance of 1.1044
b) Show that the constrained maximization performed by Mom in Note 8 gives $x = e^{-b}$, $e^a = Z = 1 + x + x^2 + x^3$ and $p_0 = 1/Z$, $p_1 = x/Z$, $p_2 = x^2/Z$, $p_3 = x^3/Z$. Show that the average demerit is thus $(x + 2x^2 + 3x^3)/(1 + x + x^2 + x^3) = 23/12$ so $0 = 13x^3 + x^2 - 11x - 23$. Solve this using MATLAB to give $x = 1.4112$ and $Z = 7.2616$ so the probabilities are $p_0 = .1386, p_1 = .1993, p_2 = .2761, p3 = .3896$ giving ignorance = 1.3157, $\beta = -0.3444$ and $\alpha = 1.9758$.
c) Now suppose Dad also remembered $p_3 = 15/72$. Confirm the following. $1 = p_0 + p_1 + p_2 + p_3 = (1 + x + x^2)/Z + 15/72$ so $1/Z = (57/72)/(1 + x + x^2)$, $138/72 = 23/12 = p_1 + 2p_2 + 3p_3 = (x + 2x^2)/Z + 45/72$ so $93/72 = (x + 2x^2)/Z$, i.e., $0 = 21x^2 - 36x - 93$ so $x_1 = 3.1294, Z_1 = 17.5866$ and $\text{probs}_1 = (.0569, .1779, .0554, 15/72)$. This gives finally, ignorance = 1.1230, $\beta = -1.1408$ and $\alpha = 2.8671$.
d) Put these three results together in a progression of increasing ignorance as Dad remembers: all frequencies, the average and p_3, and the average only. (The more Dad forgets the greater his ignorance.)

35. My discussion of maximizing ignorance to give the partition function is based on 1957 work by E T Jaynes, cited by [Rob93, p.5]. This book assumes familiarity with statistical physics and will be frustrating reading for a beginner. But it is worth dipping into in many places and should provide some further insights even at this level.

36. a) Does ignorance really never decrease? "The rich get richer and the poor get poorer": invent an example for this and see if ignorance decreases in your analysis. b) It is said that the more we know the more questions we have. Discuss this in terms of increasing ignorance.

37. We can use Mom's maximization of ignorance in Note 8 to invert the process of going from probabilities (frequencies) to moments and go instead from moments to probabilities. Let's work an example in which we do not have a complete set of moments.
Consider the values 3, 1, 2, 4, 2, for which the first three moments are 12/5, 34/5 and 108/5. There are four different values, the set {1, 2, 3, 4}.
a) To maximize the ignorance

$$p_1 \ln p_1 + p_2 \ln p_2 + p_3 \ln p_3 + p_4 \ln p_4$$

subject to

$$\begin{aligned} p_1 + p_2 + p_3 + p_4 &= 1 \\ 1p_1 + 2p_2 + 3p_3 + 4p_4 &= \frac{12}{5} \\ 1p_1 + 4p_2 + 9p_3 + 16p_4 &= \frac{34}{5} \\ 1p_1 + 8p_2 + 27p_3 + 64p_4 &= \frac{108}{5} \end{aligned}$$

show that

$$\begin{aligned} Z &= e^{-\beta-\gamma-\delta} + e^{-2\beta-4\gamma-8\delta} + e^{-3\beta-9\gamma-27\delta} + e^{-4\beta-16\gamma-64\delta} \\ \frac{12}{5}Z &= e^{-\beta-\gamma-\delta} + 2e^{-2\beta-4\gamma-8\delta} + 3e^{-3\beta-9\gamma-27\delta} + 4e^{-4\beta-16\gamma-64\delta} \\ \frac{34}{5}Z &= e^{-\beta-\gamma-\delta} + 4e^{-2\beta-4\gamma-8\delta} + 9e^{-3\beta-9\gamma-27\delta} + 16e^{-4\beta-16\gamma-64\delta} \\ \frac{108}{5}Z &= e^{-\beta-\gamma-\delta} + 8e^{-2\beta-4\gamma-8\delta} + 27e^{-3\beta-9\gamma-27\delta} + 64e^{-4\beta-16\gamma-64\delta} \end{aligned}$$

where $\alpha - 1, \beta, \gamma$ and δ are the four Lagrange multipliers, and α is represented by the partition function $Z = e^{\alpha}$ as Mom found out in Note 8.
b) Divide every term by $e^{-\beta-\gamma-\delta}$ and show that the following equations must be solved

$$\begin{pmatrix} 2-12/5 & 3-12/5 & 4-12/5 \\ 4-34/5 & 9-34/5 & 16-34/5 \\ 8-108/5 & 27-108/5 & 64-108/5 \end{pmatrix} \begin{pmatrix} e^{-\beta-3\gamma-7\delta} \\ e^{-2\beta-8\gamma-26\delta} \\ e^{-3\beta-15\gamma-63\delta} \end{pmatrix} = \begin{pmatrix} 1-12/5 \\ 1-34/5 \\ 1-108/5 \end{pmatrix}$$

giving

$$\begin{pmatrix} e^{-\beta-3\gamma-7\delta} \\ e^{-2\beta-8\gamma-26\delta} \\ e^{-3\beta-15\gamma-63\delta} \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}$$

c) That was step 1. Now take the logarithm and show that the following equations must now be solved.

$$\begin{pmatrix} 1 & 3 & 7 \\ 2 & 8 & 26 \\ 3 & 15 & 63 \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \\ \delta \end{pmatrix} = \begin{pmatrix} \ln 2 \\ 0 \\ 0 \end{pmatrix}$$

giving

$$\begin{pmatrix} \beta \\ \gamma \\ \delta \end{pmatrix} = \begin{pmatrix} -6.5849 \\ 2.7726 \\ -0.3466 \end{pmatrix}$$

d) From this, show that the numerators of the probabilities are exponentials of

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & 4 & 8 \\ 3 & 9 & 27 \\ 4 & 16 & 64 \end{pmatrix} \begin{pmatrix} \beta \\ \gamma \\ \delta \end{pmatrix} = \begin{pmatrix} 4.1589 \\ 4.8520 \\ 4.1589 \\ 4.1589 \end{pmatrix}$$

giving

$$\begin{pmatrix} 64 \\ 128 \\ 64 \\ 64 \end{pmatrix}$$

and hence probabilities

$$\begin{pmatrix} .2 \\ .4 \\ .2 \\ .2 \end{pmatrix}$$

Show that the ignorance is thus 1.3322 (using ln).
e) That was step 2. To check the final answer, confirm that the three moments are indeed

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \end{pmatrix} \begin{pmatrix} .2 \\ .4 \\ .2 \\ .2 \end{pmatrix} = \begin{pmatrix} 12/5 \\ 34/5 \\ 108/5 \end{pmatrix}$$

f) Note that this approach does not always work:
If we had only two moments, we would have two equations in two unknowns, say β and γ, but breaking the solution into two steps causes the first matrix problem to be underdetermined with two equations in three unknown exponentials, $e^{-\beta-3\gamma}, e^{-2\beta-8\gamma}$ and $e^{-3\beta-15\gamma}$. Check this claim.
If we had all four moments, we would have four equations in four unknowns, β, γ, δ and, say, ε, but the first matrix problem would be overdetermined: four equations in three exponential unknowns.
So we need a way to do the solution in one step, not the two we used for this excursion.
If we have n different values and know m moments, how must m and n relate in order for us to use the above two-step solution?

38. **Newton's method 1.** This is motivated by the previous excursion, but is self-contained.
Given the values, 3, 1, 2, 4, 2, for which the first three moments are 12/5, 34/5 and 108/5, let's try to reconstruct the probabilities (0.2, 0.4, 0.2, 0.2) for the four different values 1, 2, 3, 4, respectively, *assuming we know only the first moment*, 12/5.
a) Show that

$$\begin{aligned} Z &= e^{-\beta} + e^{-2\beta} + e^{-3\beta} + e^{-4\beta} \\ \frac{12}{5}Z &= e^{-\beta} + 2e^{-2\beta} + 3e^{-3\beta} + 4e^{-4\beta} \end{aligned}$$

so that we must solve

$$0 = (1 - \frac{12}{5})e^{-\beta} + (2 - \frac{12}{5})e^{-2\beta} + (3 - \frac{12}{5})e^{-3\beta} + (4 - \frac{12}{5})e^{-4\beta}$$

or, simpler,

$$0 = -7 - 2e^{-\beta} + 3e^{-2\beta} + 8e^{-3\beta}$$

b) This is a bit more elaborate than, say, a quadratic equation for which we know a simple formula. So we introduce *Newton's method*.
Isaac Newton (1642–1727) invented the calculus and found a way to find, in principle, the zeros of *any* expression.
Newton's idea is easy to visualize (see Week v Notes 1–5). We have a known curve, $f(x)$, whose zero (where it crosses the x-axis) we want to find. So we guess at a value of x which might be the zero, or at least close to it, and we call this guess x_1.
The slope of the curve, slope f (or slope$_x f(x)$), approximately obeys the relation

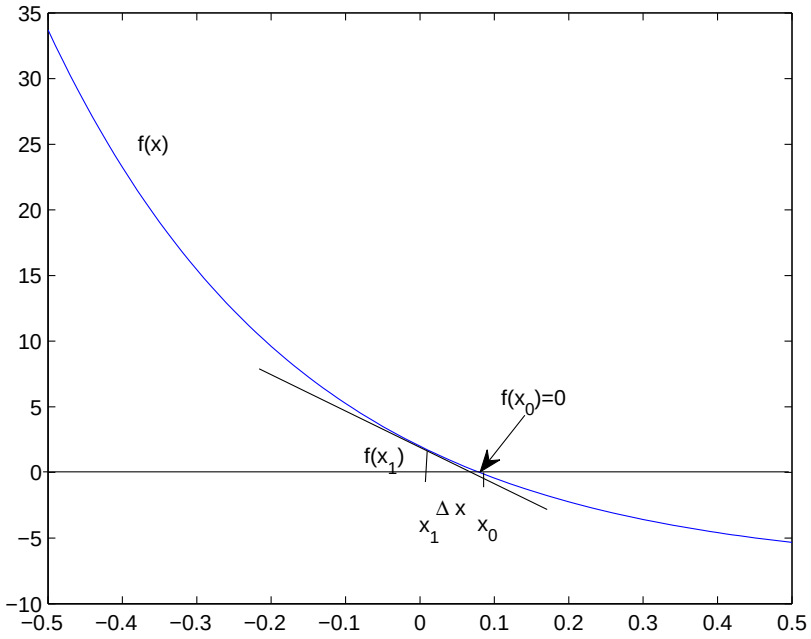
$$(\text{slope} f)_{x_1} \Delta x = f(x_1)$$

if Δx is the x-distance between the guess x_1 and x_0 (which is what we are looking for), where the function goes to zero, $f(x_0) = 0$.
So if we solve this equation for Δx (in MATLAB notation)

$$\Delta x = (\text{slope} f)_{x_1} \backslash f(x_1)$$

we can change our guess by Δx to

$$x_2 = x_1 - \Delta x$$



The figure shows why we must have the minus: the slope of the curve in the figure (it is just the above expression we are trying to solve for β) is negative, so $\Delta x \approx x_1 - x_0$ is negative if our first guess is x_1 as shown. So to advance to where the black line crosses the x-axis in the figure (about 0.08) we must subtract Δx from x_1

$$x_0 \approx x_1 - \Delta x \approx x_1 - (x_1 - x_0)$$

If we keep repeating this process, $x_1 \to x_2 \to x_3 \to ..$, eventually we should get close enough to x_0. Let's try it for the probability problem, starting with $\beta = 0$ as our first guess.

First we need slope_β for the function

$$\begin{aligned} f(\beta) &= 8e^{-3\beta} + 3e^{-2\beta} - 2e^{-\beta} - 7 \\ &= (-2, 3, 8) \begin{pmatrix} e^{-\beta} \\ e^{-2\beta} \\ e^{-3\beta} \end{pmatrix} - (7) \\ \text{slope}_\beta f = ff(\beta) &= (-2, 3, 8) \begin{pmatrix} -e^{-\beta} \\ -2e^{-2\beta} \\ -3e^{-3\beta} \end{pmatrix} \end{aligned}$$

Now calculate:

β	f	ff	$\Delta\beta = ff \backslash f$
0	2	-28	-0.0714
0.0714	0.1955	-22.7100	-0.0086
0.0800	0.0024	-22.1434	-1.1055×10^{-4}
0.0801	3.9725×10^{-7}	-22.1362	-1.7946×10^{-8}

That converged to four decimal places for β in four iterations. You can see that each successive $\Delta\beta$ is approximately the square of its predecessor: Newton's method converges "quadratically".

c) Now work out the probabilities, starting with the numerators.

$$pnum = (e^{-\beta}, e^{-2\beta}, e^{-3\beta}, e^{-4\beta})$$

and divide by Z

$$\begin{aligned} prob &= \frac{pnum}{\text{sum}(pnum)} \\ &= (0.2808,\ 0.2592,\ 0.2392,\ 0.2208) \end{aligned}$$

Compare our ignorance

$$\begin{aligned} ign &= -\text{sum}(prob. \times \log(prob)) \\ &= 1.3823 \end{aligned}$$

with our ignorance in the previous excursion, where we assumed we knew three moments, not just one.

Finally, check the moments

$$moments = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 1 & 4 & 9 & 16 \\ 1 & 8 & 27 & 64 \end{pmatrix} \times prob' = \begin{pmatrix} 2.4000 \\ 7.0032 \\ 22.9440 \end{pmatrix}$$

The moment we started with ($12/5 = 2.4$) comes back exactly. The other two are not bad.

39. **Newton's method 2.** Here is a more advanced form of Newton's method, which we can motivate by supposing we are given two moments, 12/5 and 34/5, for the set of values 1, 2, 3, 4 from the previous two excursions.

a) Show that

$$\begin{aligned} Z &= e^{-\beta-\gamma} + e^{-2\beta-4\gamma} + e^{-3\beta-9\gamma} + e^{-4\beta-16\gamma} \\ \frac{12}{5}Z &= e^{-\beta-\gamma} + 2e^{-2\beta-4\gamma} + 3e^{-3\beta-9\gamma} + 4e^{-4\beta-16\gamma} \\ \frac{34}{5}Z &= e^{-\beta-\gamma} + 4e^{-2\beta-4\gamma} + 9e^{-3\beta-9\gamma} + 16e^{-4\beta-16\gamma} \end{aligned}$$

so we must solve (simplified as in the previous two excursions)

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -2 & 3 & 8 \\ -14 & 11 & 46 \end{pmatrix} \begin{pmatrix} e^{-\beta-3\gamma} \\ e^{-2\beta-8\gamma} \\ e^{-3\beta-15\gamma} \end{pmatrix} - \begin{pmatrix} 7 \\ 29 \end{pmatrix} = f(\beta,\gamma)$$

The Newton method equations

$$(\text{slope} f)_{x_1} \Delta x = f(x)$$

also holds up if $x, \Delta x$ and $f(x)$ are vectors, as above. (slope $f)_{x_1}$ must be a matrix. It is called the *Jacobian* of f and consists of partial slopes.
Here it is for $f_1(x,y)$ and $f_2(x,y)$

$$\begin{pmatrix} \text{slope}_x f_1 & \text{slope}_y f_1 \\ \text{slope}_x f_2 & \text{slope}_y f_2 \end{pmatrix}$$

To be specific, write it out for our probability problem

$$\text{Jacobian} = ff(\beta,\gamma) = \begin{pmatrix} -2 & 3 & 8 \\ -14 & 11 & 46 \end{pmatrix} \begin{pmatrix} -e^{-\beta-3\gamma} & -3e^{-\beta-3\gamma} \\ -2e^{-2\beta-8\gamma} & -8e^{-2\beta-8\gamma} \\ -3e^{-3\beta-15\gamma} & -15e^{-3\beta-15\gamma} \end{pmatrix}$$

In the second matrix, see how the first column is the slope with respect to β of the vector in $f(\beta,\gamma)$, and the second column is its slope with respect to γ. This Jacobian is a 2×2 matrix: I've separated out the coefficients (first matrix) from the dependence on β and γ.
The iteration then is

$$\begin{pmatrix} \Delta\beta \\ \Delta\gamma \end{pmatrix} = ff \backslash f$$
$$\text{and} \quad \begin{pmatrix} \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} \beta \\ \gamma \end{pmatrix} - \begin{pmatrix} \Delta\beta \\ \Delta\gamma \end{pmatrix}$$

until the error is small enough.

$\begin{pmatrix} \beta \\ \gamma \end{pmatrix}$	f	ff	$ff \backslash f$
$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 2 \\ 14 \end{pmatrix}$	$\begin{pmatrix} -28 & -138 \\ -146 & -736 \end{pmatrix}$	$\begin{pmatrix} 1 \\ -0.2174 \end{pmatrix}$
$\begin{pmatrix} -1 \\ 0.2174 \end{pmatrix}$	$\begin{pmatrix} 0.2253 \\ 0.8924 \end{pmatrix}$	$\begin{pmatrix} -23.4455 & -115.1033 \\ -115.0461 & -586.3209 \end{pmatrix}$	$\begin{pmatrix} -0.0582 \\ -0.0099 \end{pmatrix}$
$\begin{pmatrix} -0.9418 \\ 0.2075 \end{pmatrix}$	$\begin{pmatrix} 0.0036 \\ 0.0138 \end{pmatrix}$	$\begin{pmatrix} -22.7637 & -111.8194 \\ -111.8178 & -569.1081 \end{pmatrix}$	$\begin{pmatrix} -0.0011 \\ 0.0002 \end{pmatrix}$
$\begin{pmatrix} -0.9407 \\ -0.2073 \end{pmatrix}$	$\begin{pmatrix} 0.0995 \\ 0.3538 \end{pmatrix}$	$\begin{pmatrix} -22.7525 & -111.7655 \\ -111.7655 & -569.8488 \end{pmatrix}$	$\begin{pmatrix} -0.3622 \\ 0.0648 \end{pmatrix}_{10^{-6}}$

Another fast convergence.
c) Probabilities, ignorance and check:

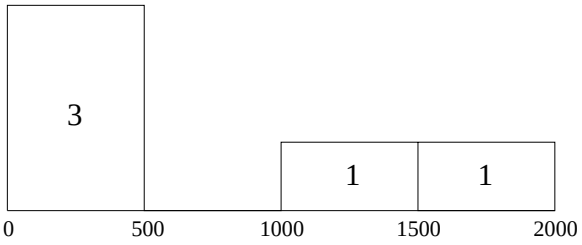
$$\begin{aligned} pnum &= (e^{-\beta-\gamma}, e^{-2\beta-4\gamma}, e^{-3\beta-9\gamma}, e^{-4\beta-16\gamma}) \\ prob &= (0.2285,\ 0.3144,\ 0.2856,\ 0.1715) \\ ign &= 1.3614 \\ moments &= (2.4000,\ 6.8000,\ 21.4287) \end{aligned}$$

What should they be, exactly?
d) Work out the case of three given moments using Newton's method instead of the two-step exact solution of two excursions ago.
Take the last step for the four different values in 3, 1, 2, 4, 2 and find the probabilities and ignorance if you are given four moments.

40. **Continuous distributions.** The discussion of continuous distributions in Note 3 omits some tricky considerations. What can we make of a distribution of salaries such as

180$ 210$ 300$ 1000$ 1800$

A continuous histogram for this could be



It does not give satisfactory aggregates. For example the mean might calculate to

$$\frac{1}{5}(3 \times 250 + 1250 + 1750) = 750$$

whereas the true mean is 698.
So there are problems designing histograms to describe continuous distributions: what range do we use? how many buckets? what width(s) for the buckets?
A better way would be to round off each value, say for this example, to one significant figure:

200$ 200$ 300$ 1000$ 2000$

giving a *discrete* histogram

occurrences	2	1	1	1
values	200	300	1000	2000

The mean of this is 740, still not close to 698 but better.
But we can maximize our ignorance and find probabilities (to replace what I just now put down as occurrences) which gives back the aggregate exactly.
Show that the function of β we need to get to zero is

$$f = -498 - 398e^{-100\beta} + 302e^{-800\beta} + 1302e^{-1800\beta}$$

find the slope ff (the "Jacobian") and calculate the sequence

β	f	ff	Δ
0	708	-2.55×10^6	-2.78×10^{-4}
2.78×10^{-4}	145.8	-1.58×10^6	-9.26×10^{-5}
3.71×10^{-4}	11.0	-1.34×10^6	-8.19×10^{-6}
3.79×10^{-4}	0.08	-1.33×10^6	-5.81×10^{-8}
3.79×10^{-4}	3.8×10^{-6}		

Find the resulting probabilities and check that they give a mean of 698.
Do it again to match the mean 698 *and* the second moment 881300.

41. **More on continuous distributions.** On the other hand, using an equal-width distribution allows us to find β analytically, without writing a program, and gives $\beta = 1/a$ in some limiting cases, where a is the mean value.
Let's try this for the distribution of the previous excursion. The partition function

$$\begin{aligned} Z &= e^{-250\beta} + e^{-750\beta} + e^{-1250\beta} + e^{-1750\beta} \\ &= e^{-250\beta} Y \end{aligned}$$

where

$$\begin{aligned} Y &= 1 + e^{-500\beta} + e^{-1000\beta} + e^{-1500\beta} \\ &\approx \frac{1}{1 - e^{-500\beta}} \end{aligned}$$

because

$$1 = (1-x)(1 + x + x^2 + x^3 + ..)$$

The mean, a, satsifies

$$\begin{aligned} aZ &= 250e^{-250\beta} + 750e^{-750\beta} + 1250e^{-1250\beta} + 1750e^{-1750\beta} \\ &= e^{-250\beta}(250 + 750e^{-500\beta} + 1250e^{-1000\beta} + 1750e^{-1500\beta}) \\ &= e^{-250\beta}(250(1 + e^{-500\beta} + e^{-1000\beta} + e^{-1500\beta}) + 500e^{-500\beta} + 1000e^{-1000\beta} + 1500e^{-1500\beta}) \\ &= e^{-250\beta}(250Y - \text{slope}_\beta Y) \end{aligned}$$

So

$$\begin{aligned} aY &= 250Y - \text{slope}_\beta Y \\ 0 &= (a - 250)Y + \text{slope}_\beta Y \end{aligned}$$

a) Show that this means (yes, really *multiply* by $1 - e^{-500\beta}$)

$$\begin{aligned} 0 &\approx (a - 250)(1 - e^{-500\beta}) - 500e^{-500\beta} \\ &= (a - 250) - (a + 250)e^{-500\beta} \end{aligned}$$

b) Show then that

$$\beta \approx \frac{1}{500} \ln\left(\frac{a + 250}{a - 250}\right)$$

and for $a = 698$ this gives $\frac{1}{\beta} = 667$ to three significant figures, closer to a than 740 or 750 from the previous excursion.
c) The width of the histogram buckets is $w = 500$ in this example, so

$$\begin{aligned} e^{w\beta} &\approx \frac{a + w/2}{a - w/2} \\ &= \frac{1 + w/(2a)}{1 - w/(2a)} \\ &= (1 + \frac{w}{2a})(1 + \frac{w}{2a} + \frac{w}{2a}^2 + ..) \\ &\approx 1 + \frac{w}{a} + \frac{1}{2}(\frac{w}{a})^2 \end{aligned}$$

to the second power of w/a supposing $w \ll a$. Show that this gives $\beta = 1/a$ to this second power.

d) Repeat the calculation to show that if the start value for the first bucket in the histogram is s then

$$\beta = \frac{1}{a - s}$$

under the assumptions above. These assumptions were (i) that the histogram has a large number of buckets and (ii) that the bucket size is very small compared to the mean value. A third assumption (iii) that $s = 0$ gives

$$\beta = \frac{1}{a}$$

e) Since the width of the buckets for the example of this and the previous excursion is too big to make the analytic approximations, and since some buckets are empty in that example, write a program to calculate β along the lines of this excursion, and hence the mean value of exactly 698.

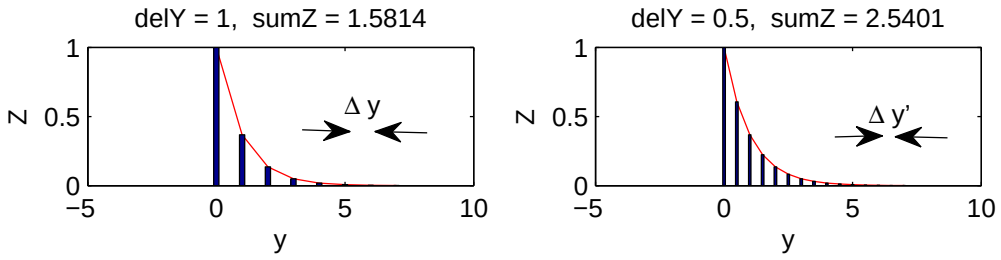
42. **Ignorance and continuous distributions.** Show that our ignorance of a continuous distribution is unbounded ("infinite").
a) For a finite, equal-width approximation, $y_j = j\Delta y$ (Δy is the bucket width), to a continuous distribution, the ignorance

$$\begin{aligned} ign &= -\sum_j (Pr) \ln(Pr) \\ &= -\sum_j (Pr) \ln\left(\frac{e^{-\beta y_j}}{\sum_k e^{-\beta y_k}}\right) \\ &= \sum_j (Pr)(\beta y_j + \ln(Z)) \\ &= \left(\sum_j (Pr)\beta y_j\right) + \ln(Z) \\ &= \beta\left(\sum_j (Pr) y_j\right) + \ln(Z) \\ &= \beta\overline{y} + \ln(Z) \end{aligned}$$

where the probability $Pr = e^{-\beta y_j}/Z$, $Z = \sum_k e^{-\beta y_k}$ and $\overline{y}$ is the average of y_j. Check this argument, then use the previous excursion to show

$$ign \approx 1 + \ln \sum_k e^{-\beta y_k}$$

b) As β gets smaller, $\overline{y}$ is a better approximation to the true average of the continuous distribution and β gets closer to $\overline{y}$ so there is no problem about the 1 in this expression for the ignorance.
The problem is with $Z = \sum_k e^{-\beta y_k} = \sum_k e^{-\beta k \Delta y}$. This gets arbitrarily large as Δy gets small and the number of terms in the sum increases to compensate. Even if the upper limit is ∞ in both cases this remains a problem.



It is easy to see that the sum on the right is about twice the sum on the left because $\Delta y' = \Delta y/2$.
So as $\Delta y \to 0, Z \to \infty$ and $\ln(Z) \to \infty$.
Write a program to draw the two plots and calculate the two sums.
c) The *area* under the curve does not depend (so heavily) on Δy. We can make a general comparison between areas and sums.
For example, show that

$$\begin{aligned} sum1 &= \sum_{j=0}^{\infty} e^{-\beta \Delta y j} = \frac{1}{1-e^{-\beta}} \\ sum2 &= \sum_{j=0}^{\infty} e^{-(\beta/2)\Delta y j} = \frac{1}{1-e^{-\beta/2}} \\ area &= \text{antislope}_y e^{-\beta y} \,|_0^{\infty} = \frac{1}{\beta} e^{-\beta y} \,|_{\infty}^{0} = \frac{1}{\beta} \end{aligned}$$

(For $\beta = 1$ in the above equations, $sum1 = 1.5820$, $sum2 = 2.5415$ and $area = 1$.) And show that the relationship between sum and area, as Δy gets small, is

$$\sum_{j=0}^{\infty} e^{-\beta \Delta y j} = \frac{1}{\Delta y} \text{antislope}_y e^{-\beta y} \,|_0^{\infty}$$

Hint. What is the meaning of $\Delta y \sum_{j=0}^{\infty} e^{-\beta \Delta y j}$?
d) You might want to motivate part (c) by comparing sums and areas for a simpler function.

$$\begin{aligned} \sum_{0}^{4} j &= \text{antislope}_x x \,|_{-0.5}^{4.5} \\ \frac{1}{2} \sum_{0}^{8} \frac{j}{2} &= \text{antislope}_x x \,|_{-0.25}^{4.25} \end{aligned}$$

Draw a picture to convince yourself that these equalities should hold.
Why do corresponding equalities not hold (exactly) for j^2 and x^2? How close do they come?
e) Why don't we change the definition so that we use areas instead of sums for Z and for the ignorance?

$$\begin{aligned} Z &= \text{antislope}_y e^{-\beta y} \,|_0^{\infty} = \frac{1}{\beta} \\ ign &= \text{antislope}_y \beta y \frac{e^{-\beta y}}{Z} \,|_0^{\infty} + \ln(Z) = 1 - \ln(\beta) \end{aligned}$$

This is OK for the ignorance, since it is the ratio of two areas, which will be the same as the ratio of the two corresponding sums.

And it will leave Z finite, and thus similarly, $\ln(Z)$.
But Z will have the wrong physical dimensions. Suppose y is an energy. We say it has physical dimension E. Then β will have physical dimension $1/E$, so βy is "dimensionless", E/E. This is necessary in an exponent, as in $s^{-\beta y}$.
Z originally was just a sum of dimensionless numbers, $e^{-\beta y}$, and so Z was also originally dimensionless. But converting from sum to area multiplies the sum by Δy, which has physical dimension E, like y itself. So the new definition of Z has dimension E.
Since $\ln(Z)$ can be expanded in a power series in Z ($\ln(Z) = (Z-1)-(Z-1)^2/2+(Z-1)^3/3-..$ is one possibility), $\ln(Z)$ has undefined physical dimensions $(E, E^2, E^3, ..)$ if Z has physical dimension E. This is not acceptable and the area definition of Z is wrong.
Argue that exponents must be dimensionless, as in $s^{-\beta y}$.

43. **Continuous distributions by antislope: average distance.** The previous excursion looked at equal-width distributions and at the limit of very small width. In these conditions we can use antislopes.
Here is an (eventually) purely analytical average for which antislopes are appropriate: show that the average distance from the centre of a disc approaches the radius of the disc, R, as the dimension of the disc increases.
(A 1-D disc is a line from $-R$ to R; a 2-D disc is bounded by a circle of radius R; a 3-D disc is bounded by a sphere of radius R; and so on. We'll look at these three cases.)
a) Write a program which constructs a grid in any number of dimensions and counts the grid points that lie within distance R of the centre, and sums the radii of each of these grid points. The average distance is this sum, s, divided by the count, c.
Here are two examples in 2-D, $R = 1$.

$\frac{s}{c} = \frac{4}{5}$ $\qquad$ $\frac{s}{c} = \frac{4}{13}(1 + \frac{1}{\sqrt{2}} + \frac{1}{2}) = 0.68$

What grid size gives $s/c = 2/3$ to four significant figures? What grid spacing, w, gives $cw^2 = \pi$ to four significant figures? What is s/c in 1-D? 3-D?
Hint. You don't need to write a separate program, with d nested loops, for each dimension d. A grid from $(0 : n-1)^d$ can be built from a single index $0{:}n^d - 1$ along the following (3-D) lines—I've used three different values for n, i.e., $n_1 = 3, n_2 = 4, n_3 = 3$, to make the example easier to figure out. The single index, a, runs from 0 to $n_1 \times n_2 \times n_3 - 1 = 23$, and is shown for each grid point.

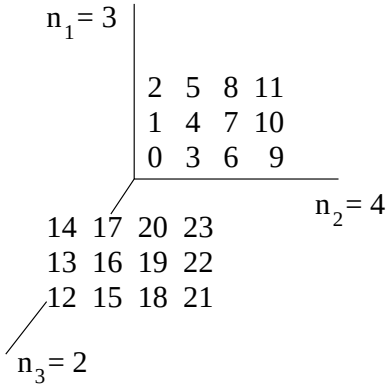


Figure out how to construct the three coordinates for each value of a shown, e.g., $a = 0 \rightarrow (0, 0, 0); a = 7 \rightarrow (1, 2, 0); a = 19 \rightarrow (1, 2, 1)$
b) Areas. The area (length) of a 1-D disc is easy to find.

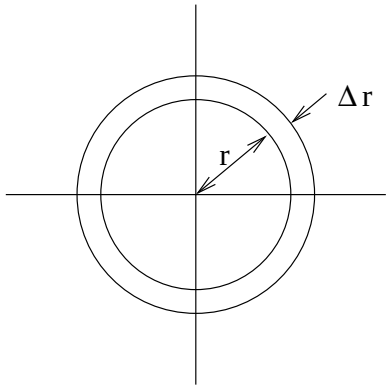
$$\text{antislope}_x 1 \mid_{-R}^{R} = 2 \text{ antislope}_r 1 \mid_0^R = 2R$$

where I've switched from x to r as variable.
The area of a 2-D disc

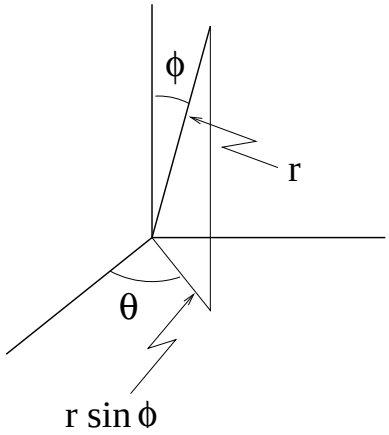
$$\begin{aligned} \text{antislope}_x \text{antislope}_y 1 \mid_{\circledR} &= \text{antislope}_r (r \text{ antislope}_\theta \mid_0^{2\pi}) \mid_0^R \\ &= 2\pi \text{ antislope}_r r \mid_0^R \\ &= 2\pi R^2 \end{aligned}$$

where $S \mid_{\circledR}$ means limit the antislopes by the circle of radius R, where I've switched from variables x, y to variables r, θ, and where the extra r arises as follows.
Antislope$_x$ functions as a sum, like the count, c, in (a), over a lot of little steps, Δx. Similarly, antislope$_y$ sums many Δy. When we switch variables from x, y to r, θ, we must find the small increments of r and θ that correspond to Δx and Δy and in particular to their product, the small area $\Delta x \Delta y$. This must be $\Delta r(r\Delta\theta)$, because the area swept out by Δr at a distance r from the centre of a circle when the angle changes by $\Delta\theta$ is proportional to r.

Another way of seeing this, in this case, is to consider the angle fully swept out, giving a ring of area approximately $2\pi r \Delta r$.



The area (volume) of a 3-D disc uses similar reasoning, this time for two angles: the small volume element is $\Delta r(r\Delta\phi)(r \sin\phi \Delta\theta)$



So

$$\begin{aligned}
\text{antislope}_x\text{antislope}_y\text{antislope}_x 1 \mid_{\textcircled{B}} &= \text{antislope}_r(\text{antislope}_\phi r\sin\phi(\text{antislope}_\theta r \mid_0^{2\pi}) \mid_0^{\pi}) \mid_0^R \\
&= \text{antislope}_r(\text{antislope}_\phi r^2\sin\phi(2\pi) \mid_0^{\pi}) \mid_0^R \\
&= 2\pi\ \text{antislope}_r r^2(2) \mid_0^R \\
&= 4\pi\frac{R^3}{3}
\end{aligned}$$

c) Total and average distances. The total distances are also sums expressed as antislopes, by analogy with (b).

$$\begin{array}{lll}
\text{1-D} & 2\ \text{antislope}_r r \mid_0^R & = \frac{2R^2}{2} \\
\text{2-D} & 2\pi\ \text{antislope}_r r^2 \mid_0^R & = \frac{2\pi R^3}{3} \\
\text{3-D} & 4\pi\ \text{antislope}_r r^3 \mid_0^R & = \frac{4\pi R^4}{4}
\end{array}$$

So the averages are these totals divided by the areas from (b).

$$\begin{array}{lll}
\text{1-D} & \frac{R^2}{2R} & = \frac{R}{2} \\
\text{2-D} & \frac{2\pi R^3}{3\pi R^2} & = \frac{2R}{3} \\
\text{3-D} & \frac{\pi R^4}{4\pi R^3/3} & = \frac{3R}{4}
\end{array}$$

Compare these results with your program from (a).
d) Beta. Maximizing ignorance gives probabilities

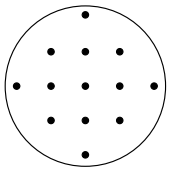
$$\frac{e^{-\beta r}}{\text{antislope}_x ... e^{-\beta r} \mid_{\textcircled{B}}}$$

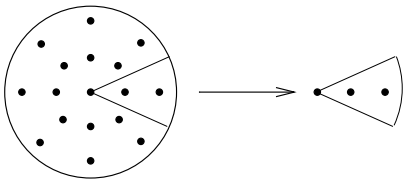
for each radius r, where there is one antislope for each dimension.
So the averages are

$$\frac{\text{antislope}_x ... r e^{-\beta r}}{\text{antislope}_x ... e^{-\beta r} \mid_{\textcircled{B}}}$$

Show, by comparing these expressions for the averages with those in (c), that $e^{-\beta r} = 1$ hence $\beta = 0$ in all three cases. Do the integration by parts if you need to convince yourself.
What, then, is the ignorance? Is this plausible: how does the probability vary, across all locations, of there being a distance from the centre to that location?

e) Instead of counting all the grid points within a disc in 2-D we can use the change of variables in (b) without going to antislopes.
For the disc of radius $R = 1$ and points separated by $\Delta x = \Delta y = 1/2$ or $\Delta r = 1/2$, here are the rectangular-grid and polar-grid pictures.





count	c = 13	c = 17	c = 3
area	$a = c\Delta x\Delta y$		$a = 2\pi\Delta r\,(0 + \frac{1}{2} + 1)$
	$= \frac{13}{4}$		$= 2\pi\,(\Delta r)^2\,(0 + 1 + 2)$
	$= 3.25$		$= 2\pi\frac{3}{4}$
			$= \frac{3\pi}{2}$

Note that $0 + \frac{1}{2} + 1$ is the sum of the radii over the 3 points, $\sum_{r=0}^{2} r\Delta r$.
Both of these results for the area can be taken as approximations to π. The rectangular one is better in this example of spacing 1/2. But the polar one also gets arbitrarily close to π as the number of points increases and Δr commensurately decreases.
Show that

$$a = 2\pi\frac{1}{(c-1)^2}\sum_{r=0}^{c-1} r = \frac{c}{c-1}\pi$$

which approaches π as c increases for a disc of radius 1 and hence area π.

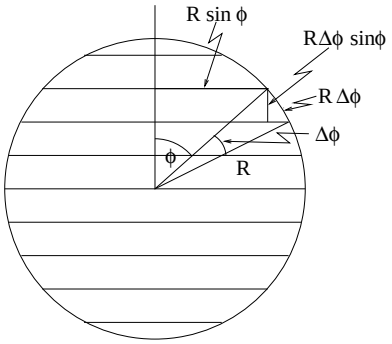
44. **Hyperspheres.** Part (c) of the previous excursion found average distances in "spheres" of 1, 2 and 3 dimensions. We can go further. Here we look at the volumes and surface areas of "hyperspheres" in these and higher dimensions.
The excursion on hyperspheres in Week 1 worked out the Cartesian coordinates for the surface of a d-dimensional hypersphere in terms of its radius and $d-1$ angles. Here we use the way we visualized hyperspheres in that excursion, but we will not need the coordinates.
We need to visualize hyperspheres the way we did in Week 1 rather than the way we did in the previous excursion. We start at the same place, with a 1-D hypersphere of radius R whose "volume" is

$$V_1 = 2R$$

Note that its "surface area" is just the two endpoints

$$A_1 = 2$$

and note that this is consistent with saying $A_1 = \text{slope}_R V_1$.
We build on this.
a) A 2-D hypersphere is made up of many 1-D hyperspheres, separated by distances $R\Delta\phi\sin\phi$. Each one has a 1-D "volume" of $2R\sin\phi$.



To add them all up we need

$$\sum_{\phi_j=0}^{\pi} 2R^2 \sin^2 \phi_j \Delta\phi$$

which we'll write as an antislope.

$$\text{antislope}_\phi 2R^2 \sin^2 \phi \mid_0^\pi = 2R^2 \text{antislope}_\phi \sin^2 \phi \mid_0^\pi$$

An important side benefit of this excursion will be finding antislopes of powers of sin. We need to use "integration by parts" (cf Note 6 and Book 8c Note 37).
To the product rule (Book 8c Note 30)

$$\text{slope } uv = u \text{ slope } v + v \text{ slope } u$$

we can apply antislopes

$$\text{antislope}(u \text{ slope } v) = uv - \text{antislope}(\text{v slope u})$$

and this may help us in evaluating the first if we can devise a u and a v such that u slope v gives the function we want the antislope of *and* v slope u is easier to find the antislope of.
For powers of sin we'll find ourselves taking two integration-by-parts steps each time. Here is $\sin^2$.
Pick $u = \sin\phi$, $\text{slope}_\phi v = \sin\phi$; so $v = -\cos\phi$ and $\text{slope}_\phi u = \cos\phi$. Then

$$\text{antislope}_\phi \sin\phi \sin\phi = -\sin\phi\cos\phi + \text{antislope}_\phi \cos\phi \cos\phi$$

Thus

$$\text{antislope}_\phi \cos\phi \cos\phi = \cos\phi \sin\phi + \text{antislope}_\phi \sin\phi \sin\phi$$

This first time, we need not do it again for $\cos^2$ because $\cos^2 = 1 - \sin^2$ so

$$\text{antislope}_\phi \sin^2\phi = -\sin\phi\cos\phi + \text{antislope}_\phi(1 - \sin^2\phi)$$

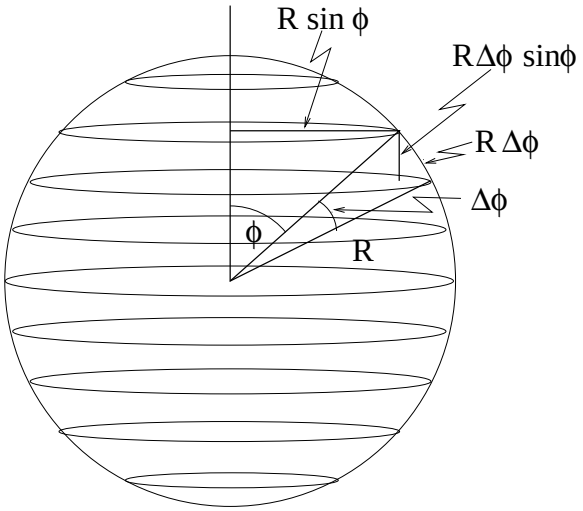
i.e.,

$$2 \text{ antislope}_\phi \sin^2\phi = \phi - \sin\phi\cos\phi$$

since antislope_ϕ 1 is ϕ (plus a constant which I've omitted since we are about to put definite limits on this antislope).
Back to the 2-D "volume".

$$\begin{aligned} 2R^2 \text{antislope}_\phi \sin^2\phi \mid_0^\pi &= 2R^2 \frac{1}{2}(\phi - \sin\phi\cos\phi) \mid_0^\pi \\ &= R^2 \end{aligned}$$

The volume is $V_2 = \pi R^2$ and the surface area is its slope $A_2 = 2\pi R$. (*Why* does $\text{slope}_R V_2 = A_2$?)
b) A 3-D hypersphere (and this is the special hypersphere which is just called a sphere) is made up of many 2-D hyperspheres in exactly the same way, each separated by a distance $R\Delta\phi \sin\phi$ and each of radius $R\sin\phi$.



The sum

$$\sum_{\phi_j=0}^{\pi} \pi(R\sin\phi)^2 R\sin\phi\Delta\phi$$

goes over to

$$\pi R^3 \text{antislope}_\phi \sin^3\phi \,|_0^\pi$$

and we must digress to find the antislope of $\sin^3$.
Pick $u = \sin^2\phi$, $\text{slope}_\phi v = \sin\phi$; so $v = -\cos\phi$ and $\text{slope}_\phi u = 2\sin\phi\cos\phi$. Let's abbreviate $\sin\phi$ to s and $\cos\phi$ to c and understand implicitly that slopes and antislopes are with respect to ϕ. Then

$$\text{antislope } s^3 = -s^2 c + \text{antislope } sc^2$$

Now we must integrate by parts a second time, to deal with antislope sc^2.
Pick $u = sc$, slope $v = c$; so $v = s$ and slope $u = c^2 - s^2 = 1 - 2s^2$ (because $c^2 + s^2 = 1$). So

$$\begin{aligned} \text{antislope } sc^2 &= s^2 c - \text{antislope}(s - 2s^3) \\ &= s^2 c - \text{antislope } s + 2\text{antislope } s^3 \end{aligned}$$

We're back to antislope s^3, but we can put our two results together.

$$\text{antislope } s^3 = 4 \text{ antislope } s^3 + s^2 c - 2 \text{ antislope } s$$

so

$$3 \text{ antislope } s^3 = -s^2 c + 2 \text{ antislope } s$$

i.e.,

$$\begin{aligned} \text{antislope } s^3 &= -\frac{1}{3}s^2 c + \frac{2}{3}\text{antislope } s \\ &= -\frac{1}{3}(s^2 + 2)c \end{aligned}$$

Back to the 3-D sphere:

$$V_3 = \pi R^3 \text{ antislope } s^3 \mid_0^\pi = \frac{4}{3}\pi R^3$$

and

$$A_3 = \text{slope}_R V_3 = 4\pi R^2$$

c) Show, by generalizing the procedure of part (b), that

$$\text{antislope } s^d = \frac{1}{d} s^{d-1} c + \frac{d-1}{d} \text{antislope } s^{d-2}$$

d) Hence use the 1-D and 2-D results to build up the following list (a blank entry is taken to be 1 when multiplied)

d	V_{d-1}	R	$\frac{d-1}{d}$	$\int_0^\pi s^d$	V_d
1		R		2	$2R$
2	$2R$	R	1/2	π	πR^2
3	πR^2	R	2/3	2	$(4\pi R^3)/3$
4	$(4\pi R^3)/3$	R	3/4	$\pi/2$	$(\pi^2 R^4)/2$
5	$(\pi^2 R^4)/2$	R	4/5	4/3	$(8\pi^2 R^5)/15$
6	$(8\pi^2 R^5)/15$	R	5/6	$(3\pi)/8$	$(\pi^3 R^6)/6$
7	$(\pi^3 R^6)/16$	R	6/7	16/15	$(16\pi^3 R^7)/105$
8	$(16\pi^3 R^7)/105$	R	7/8	$(15\pi)/48$	$(\pi^4 R^8)/24$
:					

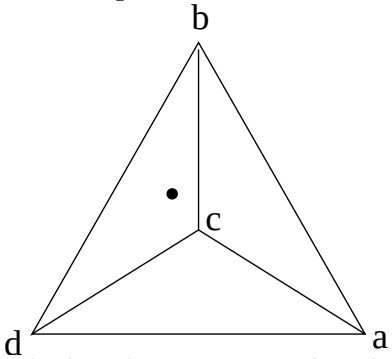
I've written $\text{antislope}_\phi \sin^d \phi \mid_0^\pi$ almost in its conventional form, $\int_0^\pi \sin^d \phi \mathrm{d}\phi$, because it is shorter. Notice that this is constructed from the product of $(d-1)/d$ two lines earlier and itself two lines earlier.
Notice that the even dimensions, $d = 2n$, give the simple expressions

$$V_{2n} = \frac{\pi^n R^{2n}}{n!}$$

What is the relationship between V_{n+1} and V_{2n-1}? Write the expression for v_{2n-1} directly as a power of R. Show from the recurrence in part (c) that these are always true.
e) You might want to work out the 4-D case explicitly, using the visualization in the hypersphere excursion of Week 1.
f) Show that almost all the volume of a hypersphere is found within a shell of thickness $\Delta R = R/d$ at the surface. (Hint. Use $A_d = \text{slope}_R V_d$.)
How does this confirm and explain the tendency we noticed in part (c) of the *previous* excursion?
g) Show that, as the number of dimensions increases, the ratio of the volume of the hypersphere to the volume of the hypercube whose sides it touches goes to zero.

45. Here are some other possible exclusions to continue Note 9. Find the joint distribution in each case and the "owes Joe or Sue" distribution. Also find correlations and co-ignorance.
a) No client will lend to Joe while remaining neutral to Sue: exclude (owes Joe, owes Sue) = $(-1\$, 0\$)$;
b) No client may lend 2\$: exclude (owes Joe, owes Sue) = $(-1\$, -1\$)$ or exclude 2\$ from result;
c) No client may borrow 1\$: exclude $-1\$$ from result;
d) No client will lend to Sue at all;
e) Either every client treats Joe and Sue identically, or no client is neutral.

46. **Geometrical interpretations of constant sums.** A hypersphere represents a constant sum of squares: e.g., $p^2+q^2+r^2+s^2$ gives the radius of a 4D hypersphere. Show that a constant sum (such as probabilities summing to 1) is represented by a regular simplex: e.g., $p+q+r+s$ equals the height of a regular tetrahedron, and is independent of x, y, z, where p, q, r and s are respectively the distances of an arbitrary point (x, y, z) from the *faces* of the tetrahedron.
Here it is for the tetrahedron. You can work it out for $p+q+r$ and a point (x, y) in an equilateral triangle.



$$\begin{aligned} a &= (1,0,0) \\ b &= (1/2, \sqrt{3}/2, 0) \\ c &= (1/2, 1/2/\sqrt{2}, \sqrt{2/3}) \\ d &= (0,0,0) \end{aligned}$$

The tetrahedron has vertices a, b, c, d as shown and the point shown has coordinates (x, y, z). To calculate the distance of that point from any face we can equate two formulas for the volume of any tetrahedron.

$$\frac{1}{6}\left|\begin{array}{c} a \vec{-} d \\ b \vec{-} d \\ c \vec{-} d \end{array}\right| = \frac{1}{3}Ah$$

where A is the area of a face the tetrahedron, h is the height from that face to the opposite vertex, and the determinant shown consists of three rows, each being the vector of three coordinates shown. For any face of the regular tetrahedron, $A = \sqrt{3}/4$. (And, for the full tetrahedon, $h = \sqrt{2/3}$ so the volume is $1/6\sqrt{2}$, but we will be replacing each of the vertices in turn by the point (x, y, z), in order to find the distance from that point to each face.)
Here are the four determinants.

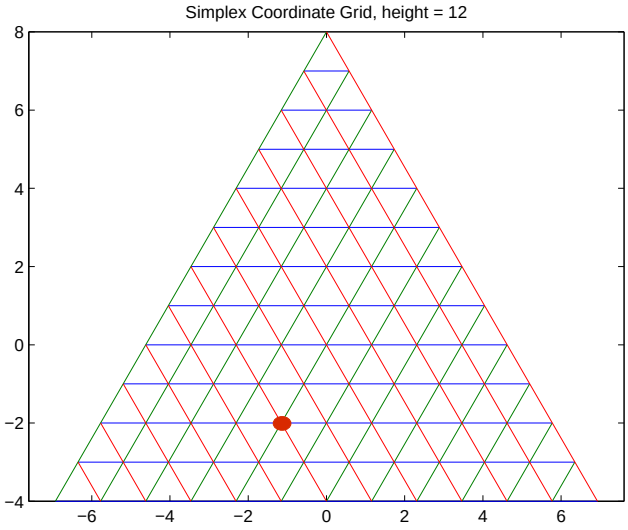
$$\begin{aligned} \det([b \vec{-} a; [x,y,z] - \vec{a}; c \vec{-} a]) &= \left|\begin{array}{ccc} -1/2 & \sqrt{3}/2 & 0 \\ x-1 & y & z \\ -1/2 & 1/2/\sqrt{2} & \sqrt{2/3} \end{array}\right| \\ &= 1/\sqrt{2} - x/\sqrt{2} - (y/2)\sqrt{2/3} - z/2/\sqrt{3} \\ \det([a \vec{-} d; [x,y,z]; c \vec{-} d]) &= \left|\begin{array}{ccc} 1 & 0 & 0 \\ x & y & z \\ 1/2 & 1/2/\sqrt{2} & \sqrt{2/3} \end{array}\right| \\ &= y\sqrt{2/3} - z/2/\sqrt{3} \\ \det([a \vec{-} d; b \vec{-} d; [x,y,z]]) &= \left|\begin{array}{ccc} 1 & 0 & 0 \\ 1/2 & \sqrt{3}/2 & 0 \\ x & y & z \end{array}\right| \\ &= \sqrt{3}z/2 \\ \det([b \vec{-} d; c \vec{-} d; [x,y,z]]) &= \left|\begin{array}{ccc} 1/2 & \sqrt{3}/2 & 0 \\ x & y & z \\ 1/2 & 1/2/\sqrt{2} & \sqrt{2/3} \end{array}\right| \\ &= x/\sqrt{2} - (y/2)\sqrt{2/3} - z/2/\sqrt{3} \end{aligned}$$

You can check that these sum to $1/\sqrt{2}$—and so the volumes (1/6 of each determinant) sum to the volume $(1/6/\sqrt{2})$ of the original tetrahedron—independently of the position (x, y, z) of the point. The distance from each face to (x, y, z) are $3/A$ times these volumes and so sum to $\sqrt{2/3}$, independently of (x, y, z). If we call the four distances p, q, r, s then $p+q+r+s = \sqrt{2/3}$, a constant. Note that the determinants must be taken with the vertices in the right order: what is it?

What are the determinants that do the same job for the equilateral triangle? What about a 1-dimensional simplex?

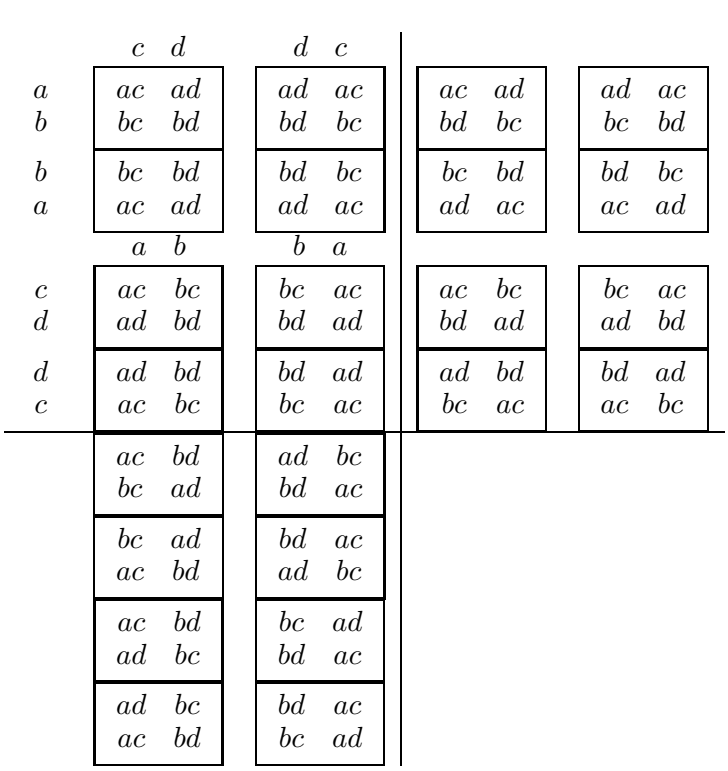
The four numbers p, q, r, s can be called *simplex coordinates*. They are constrained to sum to a constant and so there is one more of them than the number of dimensions of the simplex (line, triangle, tetrahedron, ..).

47. **Simplex coordinates.** There are $d + 1$ "simplex coordinates", discussed in the previous Excursion, in a d-dimensional simplex. The constraint that the coordinates sum to the height of the simplex reduces the dimensionality to d. Each coordinate of a given point is simply the distance of the point from one of the sides of the simplex. Do the calculations and write the program to generate the grid of all integer simplex coordinates for an equilateral triangle $(d = 2)$ of height 12 as shown in the figure.



The red spot marks the location of coordinate $(b, g, r) = (2, 4, 6)$ where b, g and r stand for "blue", "green" and "red" respectively. Show that this has Cartesian coordinates $(-2/\sqrt{3}, -2)$ for this particular triangle.

48. **When is a distribution the product of its marginals?** (See Note 9.) a) Let's explore two different binary distributions: AB has probabilities a and b with $a + b = 1$; CD has probabilities c and d with $c + d = 1$. We can generate $2(2!)^2 = 8$ product distributions, included in a total of $(2^2)! = 24$ joint distributions. Here are the 24, with the 8 singled out by showing the marginals.



What if both distributions were the same (e.g., both a, b)? What if one or both distributions were balanced (e.g., a, a or c, c)?

b) Note that when the distribution is the product of its marginals, it does not correlate the two variables (Note 10). Suppose the value associated with the probabilities a and c is 0, and with the probabilities b and d is 1. Then the average value of distribution AB is $\mu_{\rm AB} = a \times 0 + b \times 1 = b$ and the average value of distribution CD is $\mu_{\rm CD} = c \times 0 + d \times 1 = d$. The vector AB$-\mu_{\rm AB}$ takes on the two values $0 - b = -b$ and $1 - b = a$. Similarly CD$-\mu_{\rm CD}$ is $0 - d = -d$ and $1 - d = c$. The outer product of these vectors element-wise times the joint distribution (the first one above) is

$$\begin{pmatrix} -b \\ a \end{pmatrix} (-d \;\; c) \cdot * \begin{pmatrix} ac & ad \\ bc & bd \end{pmatrix} = \begin{pmatrix} bd & -bc \\ -ad & ac \end{pmatrix} \cdot * \begin{pmatrix} ac & ad \\ bc & bd \end{pmatrix} = abcd \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

which sums to zero. Work out the mutual information for this distribution—hint: the logarithm of a product is the sum of the logarithms. What happens to this cancellation in the cases that the joint distribution is not the product of its marginals?

c) In the case of certainty (a or b is 1 and c or d is 1) show that the joint "distribution" is always a product of its marginals.

d) Check that the numbers (8 and 24) in (a) above also hold if the three distributions are represented as density matrices (Note 5), e.g.,

$$\begin{pmatrix} a & \\ & b \end{pmatrix} \text{ and } \begin{pmatrix} c & \\ & d \end{pmatrix}$$

combine via *tensor product* (see Note 20 in Part IV of Book 11d) to give

$$\begin{pmatrix} a & \\ & b \end{pmatrix} \overleftarrow{\times} \begin{pmatrix} c & \\ & d \end{pmatrix} = \begin{pmatrix} ac & & & \\ & ad & & \\ & & bc & \\ & & & bd \end{pmatrix}$$

49. **Mutual information from Chesapeake clams.** Here is an alternative derivation of the idea of mutual information, following [Ula86, Ex.5.4]. Three categories of clam harvest in Chesapeake Bay can be distinguished by their tonnages: $b_1 > 3000, 1500 \leq b_2 < 3000, b_3 < 1500$. The minimum January air temperatures (in Celcius) the year before the harvest can be put into four classes: $a_1 < -12, -12 \leq a_2 < -10, -8 \leq a_3 < -10, a_4 > -8$. Hypothetical data for the numbers of harvests in each category and class over 50 seasons could be

	a_1	a_2	a_3	a_4
b_1	9	3	3	1
b_2	2	6	5	2
b_3	2	3	4	10

which turn into joint and marginal frequency distributions

$p(a_j, b_i)$	a_1	a_2	a_3	a_4	$p(b_i)$
b_1	.18	.06	.06	.02	.32
b_2	.04	.12	.10	.04	.30
b_3	.04	.06	.08	.20	.38
$p(a_j)$	.26	.24	.24	.26	1.00

So we might use this to forecast the probability of a good harvest (category b_1) as $p(b_1) = .32$. But if we additionally knew that last January's temperature had averaged $-12.7\,\mathrm{degC}$ then we could revise this probability upwards to $0.18/0.26 = 0.69$, i.e., $p(a_1, b_1)/p(a_1)$.
So although our surprisal about b_i is $-\lg p(b_i)$ a priori, after a_j is known, our surprisal becomes $-\lg p(b_i \mid a_j) = -\lg p(a_j, b_i)/p(a_j)$.
Averaging these two surprisals, our ignorance is reduced from

$$-\sum_{ij} p(a_j, b_i) \lg p(b_i)$$

to

$$-\sum_{ij} p(a_j, b_i) \lg p(a_j, b_i)/p(a_j)$$

i.e., by

$$\begin{aligned}
\sum_{ij} p(a_j, b_i)(\lg p(a_j, b_i)/p(a_j) - \lg p(b_i)) &= \sum_{ij} p(a_j, b_i) \lg p(a_j, b_i)/p(a_j)p(b_i) \\
&= \sum_{ij} p(a_j, b_i)(\lg p(a_j, b_i) - \lg p(a_j) - \lg p(b_i)) \\
&= \sum_{ij} p(a_j, b_i) \lg p(a_j, b_i) - \sum_{ij} p(a_j, b_i) \lg p(a_j) - \sum_{ij} p(a_j, b_i) \lg p(b_i) \\
&= \sum_{ij} p(a_j, b_i) \lg p(a_j, b_i) - \sum_{j} p(a_j) \lg p(a_j) - \sum_{i} p(b_i) \lg p(b_i) \\
&= -(I_{ab} - I_a - I_b) \\
&= I_a + I_b - I_{ab}
\end{aligned}$$

It takes information to reduce ignorance, so this quantity is reasonably called "mutual information": how much our knowledge of one distribution reduces our ignorance of a related distribution.
a) Check the above derivations.
b) Write a MATLAB `function [mI,tot] = mutInfo(M)` which calculates the mutual information, `mI`, of a given unnormalized matrix, `M`, and also calculates `tot = sum(sum(M))`, the

total of all entries in `M`. You may find it useful to display `M` extended by its row and column sums and by `tot`. There are several ways to find `mI` as we saw above: which is the clearest and involves the least amount of calculation?
c) Calculate the three harvest surprisals (b) and the twelve revised surprisals given specific knowledge of a for this clam data; confirm your result by calculating the average value of the differences and comparing with `mI` from `mutInfo()`.

50. a) Show that if the joint distribution of two variables a and b, $p(a,b) = p(a) \times p(b)$ (the variables are independently distributed), then the mutual information is zero.
b) It is plausible that if one variable, b, is a function of the other, a, the mutual information will be a maximum. This will be the case if the m-by-m matrix giving the joint distribution has exactly one entry in each row and each column. In particular, show that the mutual information is maximized in such a case if all entries have the same value. (What if the matrix is m-by-n, not square?) Relate this to our discovery in Note 10 that the mutual information is the same for purely correlated and purely anticorrelated distributions.
c) Modify the 4-by-4 matrix as shown, where $p+q = 1$, and show that the mutual information decreases in the modified matrix.

$$\begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix} \begin{pmatrix} p & & q & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}$$

Do this again with a modification which preserves both row and column sums:

$$\begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix} \begin{pmatrix} p & & q & \\ & 1 & & \\ q & & p & \\ & & & 1 \end{pmatrix}$$

Does this support your intuition that mutual information is maximum if one variable is a function of the other?

51. **Ascendency.** The matrix over which we calculate mutual information does not need to describe a joint distribution. Ulanowicz [Ula86] makes interesting use of it for matrices describing *flows* and in particular "trophic" flows in an ecology (Greek $\tau\rho o\phi\iota\kappa o\varsigma$, nourishment)—the flows of energy or carbon or nitrogen or phosphorus transferred between species which prey on or otherwise nourish each other.
Here is the matrix of the flows of energy (kcal/m^2/year) for the Cone Spring ecosystem on [Ula86, p.32]

in/out	plants	detrit	bacter	detriv	carniv	dissip	totals
0	11184	635	0	0	0	0	11819
300	0	8881	0	0	0	2003	11184
860	0	0	5205	2309	0	3109	11483
255	0	1600	0	75	0	3275	5205
0	0	200	0	0	370	1814	2384
0	0	167	0	0	0	203	370
1415	11184	11483	5205	2384	370	10404	42445

Here the flows go from row to column, so, for instance, the flow of 11184 is from outside the ecosystem ("in/out") to the "plants" category (representing an aggregation of species). The abbreviations are "detrit*us*", "bacter*ia*", "detriv*ores*", "carniv*ores*", and "dissip*ation*". This latter covers all output from the ecosystem which cannot be used by any other species or

system, and so is lost.
I have added the last column and the last row as checks: they sum the earlier rows and columns, respectively. Note that the column sums match the row sums for each (aggregated) species: energy (and flows generally) are *conserved* and do not accumulate in any node. (Accumulations may also be modelled, however, by self-loops, which would be nonzero entries on the diagonal of the matrix. The row sums and column sums would still be equal.) Note that the sum of useful and dissipated outputs, 1415 + 10404 = 11819, the sum of the inputs.
Running the program from the earlier Excursion *Mutual information from Chesapeake clams* on this,

```
[mI,tot] = mutInfo(coneSpring2)
```

gives

```
mI = 1.3364 tot = 42445
```

where `tot` is just the grand total of all the matrix entries, and is considered in this context to be the *total flow* in the ecosystem.
Ulanowicz defines *ascendency* to be the product, in this case `mI*tot = 56725`.
Ulanowicz claims that the ascendency tends to increase. This would give ecologists a systems-level quantity to indicate the state of development of an ecology: a *quantity* permits measurement and precision; *systems-level* makes the measure specifically ecological, rather than something microscopic and complicated such as the biochemistry of nutrition. He is circumspect in his claims, so some of the explorations in this and the next Excursions might attempt to go too far. But it is worthwhile to explore the concept itself.
The first thing to do is to ask whether ascendency really does increase with ecosystem development. I am not an ecologist and so will depend on Ulanowicz' book (and its sequel [Ula97] which is less technical in an attempt to be accessible to Ulanowicz' ecological colleagues). The obvious example given consists of two related "marsh gut ecosystems" in Crystal River, Florida. One is stressed by its location near a power plant outflow, increasing its temperature by 6 C°. This is not historical development of a single system, but one would suppose that the ascendency of the stressed system would be less than that of the unstressed system.
a) Code the matrices for the carbon flows (mg/m^2/day) of the marsh gut ecosystems given in [Ula86, pp.69,75]. I copied the networks of 17 taxa into 18×19 matrices (and probably made many mistakes) and got `mI = 1.2640, tot = 22414, mI*tot = 28331` for the unstressed system and `mI = 1.2417, tot = 18048, mI*tot = 22410` for the stressed system, bearing out the supposition. Check my work!
A possible invariant among all these numbers is the *efficiency* of a species. That is the complement to the proportion of flow that is dissipated. If we sum all the inputs in the Cone Spring ecosystem above we get

in/out	plants	detrit	bacter	detriv	carniv
1415	11184	11483	5205	2384	370

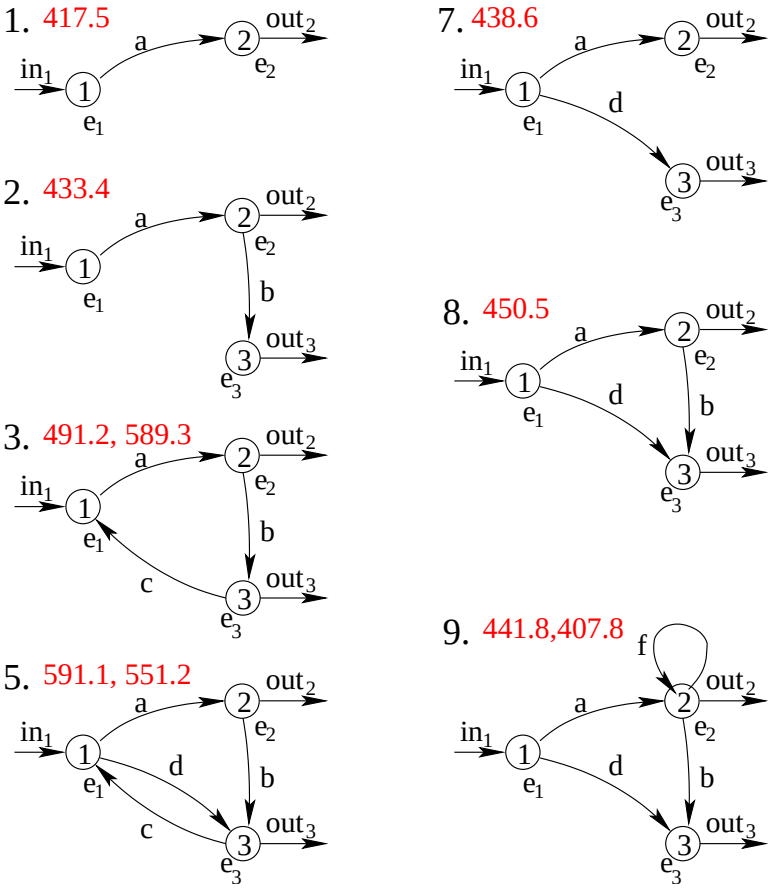
Dividing the dissipation column of the Cone Spring system above by these numbers respectively gives 1− the efficiencies of each species. So the efficiencies are

in/out	plants	detrit	bacter	detriv	carniv
1.0000	0.8209	0.7293	0.3708	0.2391	0.4514

(with the 1.0 for the "efficiency" of the in/out element: only the last five numbers count for efficiencies of actual species).
Ulanowicz [Ula86, p.58] shows an ecology of four species in which one is replaced by a more efficient species. Does this actually happen?
b) Code the two cases of this example and show that the ascendency in the first part, with the less efficient species, is 562.8816, while the ascendency in the second part, with the slightly more efficient species, is less, 554.4457. But then in this second case, Ulanowicz has killed

the output, so the comparison is unequal. So we'll need to do some more programming to be able to vary the ecosystems we are experimenting with.
c) Ulanowicz also gives an example (p.33) of cash flows in an economy. You can calculate the ascendency for this system, but economics is a less plausible candidate for a quantity such as ascendency to be useful. For one thing, economics is open to micromanaging, say by governments. More significantly, economics is susceptible to large-scale psychological effects, as when a large portion of the population loses confidence in the banks or, through a prevalent ideology of greed, the trust basic to agreements and contracts is eroded.

52. **A playpenful of baby ecologies.** Here are some baby ecologies you can play with in this Excursion.



Assume that the efficiencies of each species is given: $(e_1, e_2, e_3) = (0.8, 0.9, 0.7)$ until we get to ecology 4 (the second part of what is shown as 3 in the figure) when e_1 changes to 0.9, supposing replacement of species 1. The dissipations are not shown explicitly, but are implied by the efficiencies.
Assume that the input, in_1, is always 100. The outputs are usually given but sometimes may be derived from the internal flows, a, b, c, d, f.
The numbers in red are the ascendencies, which we will now proceed to calculate along with the internal flows.
The only other fixtures available to us, apart from the efficiencies, are the equations that say that flow is conserved at each node. There is one such equation for each node. Sometimes the flows are fewer and overdetermined: for the network to be connected there can only be one less flows than nodes, so in such a case we'll simply allow an output to be determined as well as the internal flows. Sometimes there are more flows than nodes, and so the flows are

underdetermined: in such cases, superfluous flows can be treated as parameters and supplied with values.
It's probably most straightforward to start with the ecology shown as network 3 in the figure. Here the number of flows equals the number of nodes. So $\texttt{in}_1$, $\texttt{out}_2$ and $\texttt{out}_3$ can be given and the three conservation equations solved to find the three flows a, b and c.
Here is the network matrix.

in/out	1	2	3	dissip
0	$\texttt{in}_1$	0	0	0
0	0	a	0	$(1-e_1)(\texttt{in}_1+c)$
$\texttt{out}_2$	0	0	b	$(1-e_2)a$
$\texttt{out}_3$	c	0	0	$(1-e_3)b$

To constrain node 1 to conserve flows, we must set its input (column sum) equal to its output (row sum) including dissipation:

$$\texttt{in}_1 + c = a + (1-e_1)(\texttt{in}_1 + c)$$

From this

$$a = e_1(\texttt{in}_1 + c)$$

so we see that we can just equate the bottom row with the rightmost column in

in/out	1	2	3	
0	$\texttt{in}_1$	0	0	
0	0	a	0	a
$\texttt{out}_2$	0	0	b	$\texttt{out}_2 + b$
$\texttt{out}_3$	c	0	0	$\texttt{out}_3 + c$
	$e_1(\texttt{in}_1 + c)$	$e_2 a$	$e_3 b$	

to give the three conservation equations. These follow as

$$\begin{pmatrix} 1 & 0 & -e_1 \\ e_2 & -1 & 0 \\ 0 & e_3 & -1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} e_1 \texttt{in}_1 \\ \texttt{out}_2 \\ \texttt{out}_3 \end{pmatrix}$$

If we put numbers into this, $(e_1, e_2, e_3) = (0.8, 0.9, 0.7)$, $\texttt{in}_1 = 100$, $\texttt{out}_2 = 52$ and $\texttt{out}_3 = 5$, we get $(a, b, c) = (94.5, 33.1, 18.1)$. Finally, running the network matrix, with these values plugged in, through `mutInfo()` (previous two Excursions) we get $\texttt{mI} = 1.42$, $\texttt{tot} = 345.7$ and ascendency 491.2 which is the first red value for network 3 in the figure.
a) Write a program which accepts symbolic inputs such as (corresponding to network 3)

```
LHS3 = [1,0,-sym('eff1');sym('eff2'),-1,0;0,sym('eff3'),-1]
```

and

```
RHS3 = [sym('eff1')*sym('in1');sym('out2');sym('out3')]
```

and uses

```
eval(RHS)\eval(LHS)
```

to solve for the flows a, b, c.
Your program should also use these flows and the earlier values given symbolically to construct the network matrix, in this case supplied symbolically as

```
net3 = [0,sym('in1'),0,0,0;0,0,sym('flow1'),0,...
(1-sym('eff1'))*(sym('in1')+sym('flow3'));sym('out2'),0,0,...
sym('flow2'),(1-sym('eff2'))*sym('flow1');sym('out3'),sym('flow3'),...
0,0,(1-sym('eff3'))*sym('flow2')]
```

This program will have to be adapted for overdetermined and underdetermined networks

(coming up).
b) Networks 1, 2 and 7 in the figure are overdetermined: two equations for one flow for (1) and three equations for two flows for the other two. In each case an output will also be determined. Show that the network matrix and resulting equations for network 1 are

in/out	1	2	3	
0	$\mathtt{in}_1$	0	0	
0	0	a	0	a
$\mathtt{out}_2$	0	0	0	$\mathtt{out}_2$
$\mathtt{out}_3$	0	0	0	$\mathtt{out}_3$
	$e_1\mathtt{in}_1$	e_2a	0	

and

$$\begin{pmatrix} 1 \\ e_2 \\ 0 \end{pmatrix} \begin{pmatrix} a \end{pmatrix} = \begin{pmatrix} e_1\mathtt{in}_1 \\ \mathtt{out}_2 \\ \mathtt{out}_3 \end{pmatrix}$$

Adapt your program from (a) to use only the appropriate part of the left-hand matrix to solve only for a, and then to go on to solve also for $\mathtt{out}_2$. You should find $a = 80$ and $\mathtt{out}_2 = 72$. From this derive the ascendency shown for network 1 of the figure.
Do the same for network 2, having specified $\mathtt{out}_2 = 62$. Where does the missing 10 go?
For network 7 I used $\mathtt{out}_2 = 52$.
c) Networks 5 and 9 are underdetermined: three equations to find four flows. So I used flow d in network 5 and flow f in network 9 as additional givens: 1 in both cases (but I tried again with 10 to give networks 6 and 10, which are otherwise identical respectively with 5 and 9). Show that the matrix and conservation equations for network 5 are

in/out	1	2	3	
0	$\mathtt{in}_1$	0	0	
0	0	a	d	$a+d$
$\mathtt{out}_2$	0	0	b	$\mathtt{out}_2+b$
$\mathtt{out}_3$	c	0	0	$\mathtt{out}_3+c$
	$e_1(\mathtt{in}_1+c)$	e_2a	$e_3(b+d)$	

and

$$\begin{pmatrix} 1 & 0 & -e_1 \\ e_2 & -1 & 0 \\ 0 & e_3 & -1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} e_1\mathtt{in}_1 - d \\ \mathtt{out}_2 \\ \mathtt{out}_3 - e_3d \end{pmatrix}$$

Notice that flow d appears on the right-hand side.
Adapt your program to handle underdetermined equations. (Hint: the inputs can convey this, but there must be a way to supply values for the flows that are given.)
d) Now explore all ten possibilities presented in the figure and discuss the resulting ascendencies (discuss the contributions of `mI` and `tot`):

(1.) $e1 = 0.8, e_2 = 0.9, \mathtt{in}_1 = 100$

(2.) $e1 = 0.8, e_2 = 0.9, e_3 = 0.7, \mathtt{in}_1 = 100, \mathtt{out}_2 = 62$

(3.) $e1 = 0.8, e_2 = 0.9, e_3 = 0.7, \mathtt{in}_1 = 100, \mathtt{out}_2 = 52, \mathtt{out}_3 = 5$

(4.) $e1 = 0.9, e_2 = 0.9, e_3 = 0.7, \mathtt{in}_1 = 100, \mathtt{out}_2 = 52, \mathtt{out}_3 = 5$

(5.) $e1 = 0.9, e_2 = 0.9, e_3 = 0.7, \mathtt{in}_1 = 100, \mathtt{out}_2 = 52, \mathtt{out}_3 = 5, d = 1$

(6.) $e1 = 0.9, e_2 = 0.9, e_3 = 0.7, \mathtt{in}_1 = 100, \mathtt{out}_2 = 52, \mathtt{out}_3 = 5, d = 10$

(7.) $e1 = 0.9, e_2 = 0.9, e_3 = 0.7, \mathtt{in}_1 = 100, \mathtt{out}_2 = 52$

(8.) $e1 = 0.9, e_2 = 0.9, e_3 = 0.7, \mathtt{in}_1 = 100, \mathtt{out}_2 = 52, \mathtt{out}_3 = 21$

(9.) $e1 = 0.9, e_2 = 0.9, e_3 = 0.7, \mathtt{in}_1 = 100, \mathtt{out}_2 = 52, \mathtt{out}_3 = 21, f = 1$

(10.) $e1 = 0.9, e_2 = 0.9, e_3 = 0.7, \mathtt{in}_1 = 100, \mathtt{out}_2 = 52, \mathtt{out}_3 = 21, f = 10$

What do you speculate about small cycles in ecological networks? What about large cycles?
e) For the species substitution example of [Ula86, p.58] evaluate the flows both before and after the species replacement and confirm (almost) Ulanowicz' numbers.

53. **Maximizing ascendency.** Ulanowicz does not claim that ascendency is maximized, but it would be nice if it were and we could predict the future of an ecology by calculating its configuration under maximum ascendency.
We can use numerical techniques loosely based on what Sketchpad does for nonlinear constraints (Week 8, Note 8): from guessed values of the parameters (e.g., outputs and underdetermined flows as discussed in the previous Excursion), try shifting to either side and see what the resulting ascendency calculates out to be. Then pick the best of these as a new guess and start again. (I wrote code to fit a parabola to the three resulting numbers, in case a maximum could be extrapolated directly, but the curves I got were always convex if not linear—at best they had minima not maxima.)
What happens is that flows soon go negative and resulting ascendencies become 2-numbers (complex numbers), so these must be watched for. Essentially, then, in the cases I tried, the program attempts to increase ascendency, using some predetermined step size, until any further steps produce impossible flows.
So it's not very good, but can still be played with and the results may be interesting.
a) Build such a program and see what it does to the system

in/out	1	2	
0	$\mathtt{in}_1$	0	
0	0	a	a
$\mathtt{out}_2$	b	0	$\mathtt{out}_2 + b$
	$e_1(\mathtt{in}_1 + b)$	$e_2 a$	

Starting with $\mathtt{out}_2 = 0$ and stepping it by ± 5 units each time (except where this would make it go negative, when the steps are forced to be 5 and 10) I got the following "evolution" until flow b started to go negative. (Efficiencies of the two nodes are set at 50%.)

$\mathtt{in}_1$	a	b	$\mathtt{out}_2$	ascendency
100	66.7	33.3	0	259
100	60	20	10	268
100	56.7	13.3	15	276
100	53.3	6.7	20	287
100	50	0	25	311

b) Run your program for the merged example [Ula86, p.58] with five species, one of them competing with a more efficient one. You should build in an additional constraint, that the outputs from the two competing species should sum to 10.

54. Shannon and Weaver's 1949 *channel coding theorem* says that when transmitting information across even a "noisy" channel the probability of the output differing from the input can be made arbitrarily small, provided that the number of messages to be sent is small (and certainly much smaller than the number of possible messages). How does mutual information figure in the proof?

55. What are the covariance, correlation, co-ignorance and mutual information for the two joint distributions discussed at the end of Note 9 but not in Note 10?

56. What are the correlation and mutual information for the following joint histograms? Are there conditions on the clients of Joe and Sue that would give rise to these joint histograms?

2	1		3	1		11	1		35	1		1	35	
	2	1		3	1		11	1		35	1		1	35
1		2	1		3	1		11	1		35	35		1

3			4			12			36			36			
	2	1		3	1		11	1		35	1		1	35	
	1	2		1	3		1	11		1	35		35	1	

57. Correlation is not causation. For example, among schoolchildren height is strongly correlated with weight but neither causes the other. Mathematical ability is correlated, less strongly, with both but does not cause them. Collect data to test these statements. What is the cause of all three? Can you measure it?

58. If Joe is owed 1\$ then the maximum owed to Joe or Sue cannot be -1\$ or 0\$. This explains the first two entries of the last line of the "Joe **max** Sue" vs. "Joe" table in Note 11. Explain the 3 that is the last entry of this last line.

59. **Black boxes.** Run the following MATLAB program and see if you can figure out each time what it is doing. Note that the conditional distributions for all three allowed inputs, $-1, 0$ and 1, are calculated each invocation.

```
% function blackboxes THM 090921
% Give x the values -1, 0, 1. z is some combination of x and a hidden variable y
% also -1, 0, 1. The combination is determined by an initial random choice of op
% Blackbox finds n different ys (n is a parameter set within blackbox, say 1000)
% and returns the resulting mean m, variance v and ignorance u of z.
% The user is intended to figure out what the combination is after trying
% several values for x.
function blackboxes
%function [m,v,u] = blackboxes(op) % test only
op = round(rand(1)*6 - 0.5); % range 0,1,2,3,4,5
n = 1000;
out = cell(4,4); % results will be displayed in  out
out(1,1) = {'x'}; out(1,2) = {'mean'}; % column titles: next 3 rows will
out(1,3) = {'var'}; out(1,4) = {'ignor'}; % correspond to x = -1, 0, 1
for x = -1:1
  t = 0; % total
  s = 0; % sum squares
  hist = zeros(1,5); % to calculate ignorance
  for k = 1:n
    y = round(rand(1)*3 - 1.5); % range -1,0,1
    switch op
    case 0
      z = x+y;
      hist(z+3) = hist(z+3)+1; % hist(1:5) for z = -2:2
%       1  2  3  2  1  m  v   u
%  x\y -2 -1  0  1  2
```

```
% 1 -1  1  1  1 -1 2/3 lg3
% 1  0     1  1  1  0 2/3 lg3
% 1  1        1  1  1  1 2/3 lg3
    case 1
      z = x-y;
      hist(z+3) = hist(z+3)+1;
% ditto
    case 2
      z = x*y;
      hist(z+2) = hist(z+2)+1;
%       2  5  2  m  v   u
%  x\y -1  0  1
% 1 -1  1  1  1  0 2/3 lg3
% 1  0     3  0  0   0
% 1  1  1  1  1  0 2/3 lg3
    case 3
      y = round(rand(1)*2 - 0.5)*2 - 1; % range -1,1 NB need nonzero y this case
      z = y^x;
      hist(z+2) = hist(z+2)+1;
%       2  4  m  v   u
%  x\y -1  1
% 1 -1  1  1  0  1   1
% 1  0     2  1  0   0
% 1  1  1  1  0  1   1
    case 4
      z = max(x,y);
      hist(z+2) = hist(z+2)+1;
%       1  3  5  m   v   u
%  x\y -1  0  1
% 1 -1  1  1  1  0  2/3 lg3
% 1  0     2  1 1/3 2/9 lg3-2/3
% 1  1        3  1   0   0
    case 5
      z = min(x,y);
      hist(z+2) = hist(z+2)+1;
%       5  3  1   m   v   u
%  x\y -1  0  1
% 1 -1  3  -1   0   0
% 1  0  1  2 -1/3 2/9 lg3-2/3
% 1  1  1  1  1   0  2/3 lg3
    end %switch
    t = t + z;
    s = s + z^2;
  end %for y
% hist % test only
  u0 = log2(n);
  sh = size(hist);
  for k = 1:sh(2)
    h = hist(k);
    if h>0
      u0 = u0 - h*log2(h)/n;
    end %if k
  end %for k
```

```
  out(x+3,1) = {x}; out(x+3,2) = {t/n};
  out(x+3,3) = {s/n - (t/n)^2}; out(x+3,4) = {u0};
end %for x
out
```

60. Using $d^J_j = \sum_m d^{J,\mathbf{max}}_{jm}$ show that $d^J_j = \sum_m d^{J|\mathbf{max}}_{jm} d^{\mathbf{max}}_m$, the second basis for Bayesian inference in Note 11. (Note that $d^{J,\mathbf{max}}$ and $d^{\mathbf{max},J}$ are two ways of writing the same thing, the joint distribution.)
If you know $d^{\mathbf{max}|J}$ and d^J how would you get $d^{\mathbf{max}}$? Using $d^J_j = \sum_m d^{J,\mathbf{max}}_{jm}$ show that

61. **Thongs.** How long must we type randomly before we produce a complete Shakespeare sonnet? I won't address this problem here, but look instead at ways of biassing the outcome in favour of success. We could start by adjusting the probabilities of striking the keys so they coincide with typical English usage. But then why not take into account what's already been typed? We could look one letter back and produce the next letter according to the probability that the resulting *dipthong* occurs in normal English usage. Or we could go further and look two letters back and use "tripthong" probabilities. Or three letters and "quadripthongs", and so on.
So we talk about "n-thongs" or *thongs* and their probabilities. We actually need conditional distributions (Note 11) as we can see from the text `abbaabaa`, drawn from a binary alphabet.

1-thongs	`a`	`b`
freqs.	5/8	3/8

2-thongs	`aa`	`ab`	`ba`	`bb`
freqs.	2/7	2/7	2/7	1/7

Here are the numbers of occurrences ..

	`a`	`b`	
`a`	2	2	4
`b`	2	1	3
	4	3	

.. and hence the frequencies.

	`a`	`b`	
`a`	1/2	1/2	4/7
`b`	2/3	1/3	3/7

Note that the denominators are 1 less than for the 1-thongs, because there are $n-1$ dipthongs in a text of length n.

3-thongs	`aab`	`aba`	`abb`	`baa`	`bba`
freqs.	1/6	1/6	1/6	2/6	1/6

With occurrences and frequencies

	`a`	`b`	
`aa`	0	1	1
`ab`	1	1	2
`ba`	2	0	2
`bb`	1	0	1
	4	2	

	`a`	`b`	
`aa`	0	1	1/6
`ab`	1/2	1/2	2/6
`ba`	2	0	2/6
`bb`	1	0	1/6

a) Write a program to analyze a text and calculate conditional probabilities for all thongs up to a given parameter n. Write another program which uses these probabilities for m-thongs, given parameter m, to generate a random text starting with the same m letters as a given text. Here are samples of what I got using, for both programs, the text
`the quick red fox jumps over the lazy brown dog.  but the dog did not bark.`

Using 1-thongs

```
t juid fothere bazy the but d ove qumps bazy bazy bumps fog downot bazy d
```

Using 2-thongs

```
the lazy brown did fox jumps over the dog dog did fox jumps over the quick
```

and

```
the lazy but bark.
```

(Note that there can be no statistics for the last dipthong in this text, so because the text is short it is quite possible to get to a point from which there is nowhere to go further.)

b) Run your programs on a file containing Shakespeare's sonnets. For what m do your m-thongs give results that are somewhat Shakespearean? What major aspects are missing?

c) My generalizations of "dipthong" are inelegant because I don't even use that word correctly. Find better terminology.

62. **Order of a symbol system.** How far is it valid to increase m in the previous Excursion? In the red fox example of that Excursion m=2 provided results containing only English words. But we'd have to go to bigger m if the statistics were drawn from a large corpus.

We can define the *order* of a system of symbols as the number of adjacent symbols from which the conditional distribution is extracted. Thus a 2nd-order system is constructed by considering pairs of symbols and using the frequencies in some given corpus by which a single symbol is followed by one of the symbols of the alphabet. Third order arises from triplets of symbols and frequencies with which the first pair is followed by the third symbol. Backtracking, 1st order arises from using the frequencies of occurrence of single symbols in the corpus, and 0th-order from simply assigning all symbols the same frequency.

As well as generating such systems, as we did in the previous Excursion, we can analyze existing systems, including whole languages if we have sufficiently large corpi. Apparently, human natural languages are, if we take the symbols to be words rather than letters, between 8th and 9th order. That is, at 1st through 8th or 9th order we can make better-than-random predictions about what will come next, but beyond that we do no better than random guessing. (Try this at third order: "How are —".) Apparently, too, dolphin symbols systems are about 4th-order. (But my source is shakey, so do your own research.)

Such predictions are best made using mutual information (Note 10). For example, using the frequencies from text `abbaabaa` of the previous Excursion, the 2nd-order mutual information is $I_{1;1} = I_1 + I_1 - I_2$ where

$$\begin{aligned} I_1 &= -\frac{4}{7}\lg\frac{4}{7} - \frac{3}{7}\lg\frac{3}{7} \\ &= \lg 7 - (8 + \lg 3)/7 \\ I_2 &= -(\frac{2}{7}\lg\frac{2}{7}) \times 3 - \frac{1}{7}\lg\frac{1}{7} \\ &= \lg 7 - \frac{6}{7} \\ I_{1;1} &= \lg 7 - \frac{10}{7} - \frac{2}{7}\lg 3 = 0.926 \end{aligned}$$

The 3rd-order mutual information is $I_{2;1} = I_1 + I_2 - I_3$ where

$$\begin{aligned} I_1 &= -\frac{4}{6}\lg\frac{4}{6} - \frac{2}{6}\lg\frac{2}{6} \\ &= \lg 6 - \frac{5}{3} \\ I_2 &= -(\frac{1}{6}\lg\frac{1}{6}) \times 2 - (\frac{2}{6}\lg\frac{2}{6}) \times 2 \\ &= \lg 6 - \frac{2}{3} \end{aligned}$$

$$\begin{aligned} I_3 &= -(\frac{1}{6}\lg\frac{1}{6}) \times 4 - \frac{2}{6}\lg\frac{2}{6} \\ &= \lg 6 - \frac{1}{3} \\ I_{2;1} &= 2\lg 3 - 2 = 1.17 \end{aligned}$$

a) Show that $I_{3;1} = \lg 5 - 8/5 = 0.7219$, $I_{4;1} = 2 - 3(\lg 3)/4 = 0.8113$, $I_{5;1} = \lg 3 - 2/3 = 0.9183$ and $I_{n;1} = 0$ for $n = 6$ and 7.
b) Write a program to calculate mutual information from an input file containing a text or a corpus of texts. Note how the denominators above decrease as the order increases. But with a large corpus they will not change much and you can get away with using directly the frequencies you get from the singletons, pairs, etc. For example in the above you can use 5/8 and 3/8 for `a` and `b`, respectively, instead of 4/7 and 3/7 for 2nd order, 4/6 and 2/6 for 3rd order, etc.
So you can build up lists of "thongs", rowsort to cluster duplicates, and just count occurrences instead of building matrices. But you may need matrices if you want to find conditional distributions.
c) Explore mutual information for various orders in a corpus such as Shakespeare's sonnets, and discuss. (Remember that we are still working with letters, not words, as the basic symbols.)

63. **Marilyn vos Savant on the Monty Hall problem.** In his game show *Let's Make a Deal* Monty Hall hid an expensive car and two goats behind three doors respectively and invited the contestant to choose one door. Monty then opened another door, revealing a goat. Monty next allowed the contestant to switch to the remaining door. What should the contestant do? In her column *Ask Marilyn* in *Parade* magazine, Marilyn vos Savant said: switch! She was vilified by thousands including many mathematicians, who *knew* that you can't change probabilities this way.
a) Write a computer simulation, `probWin = montyHall(switch)` in which `switch` is false if the contestant does not switch after Monty Hall reveals the goat behind one of the non-chosen doors, and true if hey does. Here are my results for 1000 trials

```
>> probWin = montyHall(false)
probWin = 0.3460
>> probWin = montyHall(true)
probWin = 0.6570
```

b) Why? Hint: what is the probability of the contestant *losing* as a result of switching?
c) Here is a more elaborate answer, applying Bayes' theorem. If you don't switch, the probabilities are unchanged as we all *know*: 1/3.
If you do switch, you're taking advantage of the additional knowledge that Monty won't reveal the car: $\text{Prob}(m_j \mid c_j) = 0$, where c_j is the event that the car is behind door j and m_k is the event that Monty opened door k.
The above does not apply to the case $j = d$ where d is the door you chose, because Monty does not open that door. Let's fix our ideas by supposing $d = 3$ since the analysis will be the same for any other d.
We need to find $\text{Prob}(c_j \mid m_k)$ for the two possibilities $j, k \in \{1, 2\}$ with $j \neq k$, given $\text{Prob}(m_k \mid c_j)$ and $\text{Prob}(c_j)$ for all j, k. We can use Bayes,

$$\text{Prob}(c_j \mid m_k) = \frac{\text{Prob}(m_k \mid c_j)\text{Prob}(c_j)}{\sum_e ll\text{Prob}(m_k \mid c_e ll)\text{Prob}(c_e ll)}$$

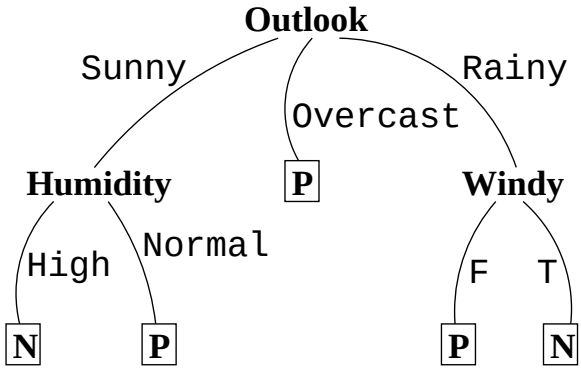
First, $\text{Prob}(c_j) = 1/3$ for all j, the original probabilities.
Second, with $d = 3$, $\text{Prob}(m_1 \mid c_1) = 0 = \text{Prob}(m_2 \mid c_2)$, because Monty never picks a winning

door.
Third, Prob($m_1 \mid c_2$) = 1 = Prob($m_2 \mid c_1$), because if the car is not behind door 3, Monty will open a door it is not behind.
Finally, Prob($m_1 \mid c_3$) = 1/2 = Prob($m_2 \mid c_3$) if we assume that Monty chooses randomly between door 1 and door 2 if the car is behind door 3.
From all this, show that Prob($c_1 \mid m_2$) = 2/3 = Prob($c_2 \mid m_1$), twice the original chance of winning.
(The clue that Bayes should be used here is that the contestant *learned* something by Monty opening a non-car door.)
d) Apart from the math, what is the psychological aspect? Why might it be best not to switch? (What have we assumed in the above solutions?)

64. **Decision trees.** The task of *classification algorithms* is to use a certain amount of *training data* to find rules for classifying any new data of the same form that may happen along in the future. This is also the task of science in general, here caricatured by application to very specific data of fixed format.
A classic example (and I will not say—but leave you to guess—what a `P` result tells us that the weather permits us to do, because that would interfere with your formal understanding of the process) classifies various types of weather as negative `N` or positive `P` [Qui86].

Training	(*Outlook*	*Temperature*	*Humidity*	*Windy*	*Class*)
1	`sunny`	`hot`	`high`	`f`	`N`
2	`sunny`	`hot`	`high`	`t`	`N`
3	`overcast`	`hot`	`high`	`f`	`P`
4	`rain`	`mild`	`high`	`f`	`P`
5	`rain`	`cool`	`normal`	`f`	`P`
6	`rain`	`cool`	`normal`	`t`	`N`
7	`overcast`	`cool`	`normal`	`t`	`P`
8	`sunny`	`mild`	`high`	`f`	`N`
9	`sunny`	`cool`	`normal`	`f`	`P`
10	`rain`	`mild`	`normal`	`f`	`P`
11	`sunny`	`mild`	`normal`	`t`	`P`
12	`overcast`	`mild`	`high`	`t`	`P`
13	`overcast`	`hot`	`normal`	`f`	`P`
14	`rain`	`mild`	`high`	`t`	`N`

From this we would like to build a *decision tree* which can classify any future configuration of weather—i.e., find out if it is `N` or `P`—in as few steps as possible. The decision tree we will find out how to build from the above training data is



Note that *Temperature* plays no role in the decision tree: look carefully at the original data and confirm that `cool` versus `hot` *Temperature* makes no difference to the classification if

every other attribute (*Outlook*, *Humidity* and *Windy*) stays the same.
a) Confirm that using this decision tree on the following new data will give the classifications *Class* as shown.

Outlook	*Temperature*	*Humidity*	*Windy*	*Class*
sunny	cool	high	t	N
rain	hot	high	f	P
rain	hot	high	t	N

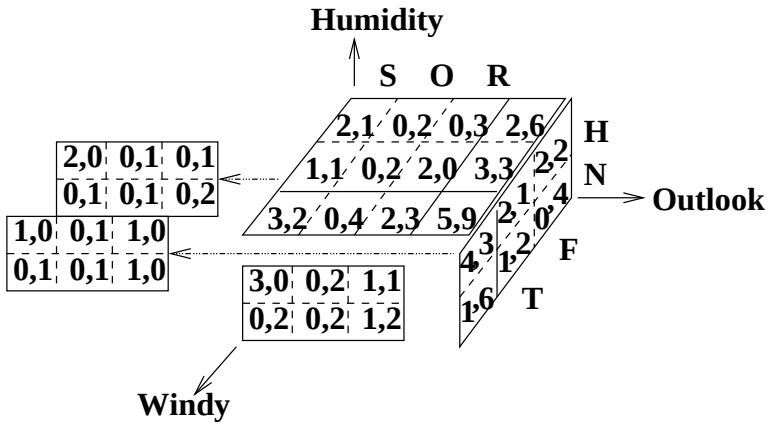
To construct this decision tree from the training data, we would like to minimize ignorance. The ignorance left us by the whole set of training data, which has 5 `N` and 9 `P` among its 14 entries, is

$$I(5,9) = -\frac{5}{14}\lg\frac{5}{14} - \frac{9}{14}\lg\frac{9}{14} = \lg 14 - (5\lg\frac{5}{14} + 9\lg\frac{9}{14})/14 = 0.940$$

bits.
Since we are using *counts* of the numbers of occurrences of `N` and `P`, let's recast the training data using pairs (#`N`,#`P`) and at the same time getting rid of the *Temperature* attribute, keeping only the two double counts (rows 1, 8 and 5, 10) that result from the presence of *Temperature* data.

Training	(*Outlook*	*Humidity*	*Windy*	#*Class*)
	sunny	high	f	(2,0)
	sunny	high	t	(1,0)
	overcast	high	f	(0,1)
	rain	high	f	(0,1)
	rain	normal	f	(0,2)
	rain	normal	t	(1,0)
	overcast	normal	t	(0,1)
	sunny	normal	f	(0,1)
	sunny	normal	t	(0,1)
	overcast	high	t	(0,1)
	overcast	normal	f	(0,1)
	rain	high	t	(1,0)

To decide which attribute to place at the root (or at any other node) of the decision tree, we must aggregate these counts in all possible ways. This is best shown as a *datacube* (and fortunately we got rid of *Temperature* so that the result is three-dimesional). Here it is, with the original data pulled out from inside so we can see it, and the aggregations shown as surfaces and edges and one corner.



b) Confirm the original data and all the aggregations shown in the datacube.
What we will call the "*Outlook* edge" of this datacube are the three pairs of counts, (3,2), (0,4) and (2,3). Similarly the "*Humidity* edge" is the two pairs, (1,6) and (4,3); and the "*Windy* edge" is (2,6) and (3,3).
Each pair in each of these edges has an associated ignorance, e.g., I(3,2), I(0,4) and I(2,3) for the *Outlook* edge. We can find the expected ignorance for each edge, using the function $I()$ introduced above.

$$\begin{array}{rcl} \textit{Outlook} & \frac{5}{14}I(3,2) + \frac{4}{14}I(0,4) + \frac{5}{14}I(2,3) & = 0.694 \\ \textit{Humidity} & \frac{7}{14}I(4,3) + \frac{7}{14}I(1,6) & = 0.788 \\ \textit{Windy} & \frac{8}{14}I(2,6) + \frac{6}{14}I(3,3) & = 0.892 \end{array}$$

Of these the least ignorance comes from the *Outlook* attribute, so this becomes the root node of the decision tree.
To find the subtrees, we repeat this process for the two aggregate planes containing this edge for *Outlook*. For *Outlook*=`sunny`, this requires us to compare

$$\frac{3}{5}\mathcal{I}(2,1) + \frac{2}{5}\mathcal{I}(1,1) = 0.951$$

(*Windy*) with

$$\frac{3}{5}\mathcal{I}(3,0) + \frac{2}{5}\mathcal{I}(0,2) = 0$$

(*Humidity*), which is smaller, so *Humidity* forms the subtree below *Outlook*=`sunny`.
For *Outlook*=`overcast`, the total ignorance, $\mathcal{I}(0,4)$, is already zero, so no subtree is needed: every *Class* for *Outlook*=`overcast` is `P`.
Finally, for *Outlook*=`rain`, the comparisons

$$\frac{3}{5}\mathcal{I}(0,3) + \frac{2}{5}\mathcal{I}(2,0) = 0$$

(*Windy*) with

$$\frac{2}{5}\mathcal{I}(1,1) + \frac{3}{5}\mathcal{I}(1,2) = 0.951$$

(*Humidity*) give the subtree to *Windy*.
c) Write a program which automates this process of building an optimum decision tree.
d) Write a program which uses the final decision tree to make further classifications.
e) Look up automatic classification. Where do Bayes' classifiers fit into the framework of aggregation we've examined above? "One-rule" classifiers? "Instance-based learning"?

65. Any part of the Preliminary Notes that needs working through.

References

[FLS64] R. P. Feynman, R. B. Leighton, and M. Sands. *The Feynman Lectures on Physics, Volume I.* Addison-Wesley, 1964.

[Gar82] Martin Gardner. *aha! Gotcha: Paradoxes to puzzle and delight.* W. H. Freeman and Company, New York, 1982. Derived from The Paradox Box, Scientific American 1975.

[Hoy60] Fred Hoyle. *The Black Cloud.* Penguin Books Ltd., Harmondsworth, UK, 1960. Heinemann 1957.

[MGB74] Alexander M Mood, Franklin A Graybill, and Duane C Boes. *Introduction to the Theory of Statistics, 3rd ed.* McGraw-Hill Inc., New York, 1974.

[Qui86] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

[Rob93] Harry S Robertson. *Statistical ThermoPhysics*. P T R Prentice-Hall, Inc., Engelwood Cliffs, NJ, 1993.

[Str66] Christopher Strachey. System analysis and programming. *Scientific American*, 215(3):112–24, Sept. 1966.

[Ula86] Robert E Ulanowicz. *Growth and Development: Ecosystems Phenomenology*. Springer-Verlag, New York, 1986.

[Ula97] Robert E Ulanowicz. *Ecology, the Ascendent Perspective*. Columbia University Press, New York, 1997. Complexity in Ecology Series, T. F. H. Allen and David W. Roberts, Eds.

[vB98] Hans Christian von Baeyer. *Maxwell's Demon: Why Warmth Disperses and Time Passes*. Random House, New York, 1998.

[vN55] John von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, Princeton, N.J., 1955. translated from the German ed. by Robert T. Beyer.

[Wei09] Eric Weisstein. Gram-Schmidt orthonormalization. mathworld.wolfram.com/Gram-SchmidtOrthonormalization.html accessed 09/9/6, 2009.