Probability problems for Sigma - part 2

Andy Mitchell

July 19, 2021

Hello Sigma members!
Following from my previous "Probability problems for Sigma - part 1", I had a crack at a slightly different problems - described below. As there is quite a lot of manipulation of formulae involving p and q , I'm starting with some potentially useful identities before launching in.... I'm also using some recurrence relations, so feel free to look at part 1 if you wish.

1 Formulae involving p and q

This section includes some random identities involving the probabilities p and $q = 1 - p$:

$$p^2 + q^2 \equiv 1 - 2pq \tag{PQ1}$$

$$p + q^2 \equiv q + p^2 \tag{PQ2}$$

$$q + pq + p^2 \equiv 1 \tag{PQ3}$$

$$p^3 + q^3 \equiv 1 - 3pq \tag{PQ4}$$

$$p^4 + q^4 \equiv 1 - 4pq + 2p^2q^2 \tag{PQ5}$$

2 Runs of either heads or tails

2.1 Question

A biased coin shows heads with probability $p = 1 - q$ whenever it is tossed.
What is the probability that in n tosses, there is no run of length R of *either* heads or tails? $(n \geq 1)$

2.2 Comment 1

I suspect that a general solution, even establishing a general recurrence relation for any R is quite tricky. All I have done is to evaluate some particular cases, and presume that the same method can applied for any larger specific value of R.

2.3 Answer for $R = 2$

The question for $R = 2$ is actually very easy, since after the first toss, all subsequent values must alternate between heads and tails to avoid a run of 2.
For even values of n , $v_n = 2p^{n/2}q^{n/2}$, and for odd values of n,
$v_n = p^{(n+1)/2}q^{(n-1)/2} + p^{(n-1)/2}q^{(n+1)/2} = p^{(n-1)/2}q^{(n-1)/2}$, however this is a useful example to cut our teeth on!
Let A_j represent the event that, in j tosses, no pair of heads or tails occur successively and let $v_j := \mathbf{P}(A_j)$.
Also, define H_j as $\mathbf{P}(A_j \mid \text{first toss is a head})$ and T_j as $\mathbf{P}(A_j \mid \text{first toss is a tail})$.

Conditioning on the result of the first toss for v_{n+2},

$$v_{n+2} = \mathbf{P}(A_{n+2}) = H_{n+2}p + T_{n+2}q$$

But, H_{n+2} is the probability of no pair occurring in $n+2$ tosses given a head on the first toss, which means that the subsequent toss can only be a tail (with probability q) and the 'state' is then reset to that in which there are $n+1$ tosses remaining with the "first" toss being a tail. So

$$H_{n+2} = qT_{n+1} \qquad \text{and similarly } T_{n+2} = pH_{n+1} \tag{1}$$

which gives :

$$v_{n+2} = pq(T_{n+1} + H_{n+1})$$

Using (1) again, but with n reduced by 1, this leads to

$$\begin{aligned} v_{n+2} &= pq(pH_n + qT_{n+1}) \\ &= pqv_n \end{aligned}$$

Also, we can determine by looking at the possibilities, that $v_1 = 1$ and $v_2 = pq + qp = 2pq$. From this, we can directly obtain the solution mentioned at the beginning of this section. Alternatively, we can solve the recurrence relation :
The auxiliary equation is $\theta^2 - pq = 0 \implies \theta = \pm\sqrt{pq}$, yielding the general solution :

$$v_n = C(-\sqrt{pq})^n + D(\sqrt{pq})^n \text{ for some } C \text{ and } D.$$

From the initial conditions,

$$\begin{aligned} v_1 = 1 &\implies 1 = (-C + D)\sqrt{pq} \\ v_2 = 2pq &\implies 2pq = (C + D)pq \implies C + D = 2 \\ &\implies C = 1 - \frac{1}{2\sqrt{pq}} \text{ and } D = 1 + \frac{1}{2\sqrt{pq}} \end{aligned}$$

So :

$$v_n = \left(1 - \frac{1}{2\sqrt{pq}}\right)(-\sqrt{pq})^n + \left(1 + \frac{1}{2\sqrt{pq}}\right)(\sqrt{pq})^n$$

This looks quite unusual, but for even and odd values of n, different pairs of terms cancel each other out to give the result we had before.
Note that for $p = q = \frac{1}{2}$, the values of v_1, v_2, v_3, v_4 etc. are 1, $\frac{1}{2}$,$\frac{1}{4}$,$\frac{1}{8}$ ....

2.4 Answer for $R = 3$

This time the solution is not obvious (to me anyway!), but we can try the same approach.
Let A_j represent the event that, in j tosses, no run of three heads or tails occur successively and let $v_j := \mathbf{P}(A_j)$.
Also, define H_j as $\mathbf{P}(A_j \mid \text{first toss is a head})$ and T_j as $\mathbf{P}(A_j \mid \text{first toss is a tail})$.
Conditioning on the result of the first toss for v_{n+4},

$$v_{n+4} = \mathbf{P}(A_{n+4}) = H_{n+4}p + T_{n+4}q$$

But, H_{n+4} is the probability of no run of three occurring in $n+4$ tosses given a head on the first toss, which means that either :

(i) The next toss is a tail (with probability q) and the 'state' is then reset to that in which there are $n+3$ tosses remaining with the "first" toss being a tail

(ii) The next toss is a head, followed by a tail (with probability pq) and the 'state' is then reset to that in which there are $n+2$ tosses remaining with the "first" toss being a tail

So

$$H_{n+4} = qT_{n+3} + pqT_{n+2} \tag{2}$$

Similarly :

$$T_{n+4} = pH_{n+3} + qpH_{n+2} \tag{3}$$

which gives after substituting :

$$v_{n+4} = pqT_{n+3} + p^2qT_{n+2} + pqH_{n+3} + q^2pH_{n+2}$$

As for the $R = 2$ solution, using (2) and (3) again, but with n reduced by 1, and substituting a further time, this leads to

$$\begin{aligned} v_{n+4} &= pq\left(pH_{n+2} + qpH_{n+1}\right) + p^2q\left(pH_{n+1} + qpH_n\right) \\ &\qquad + pq\left(qT_{n+2} + qpT_{n+1}\right) + q^2p\left(qT_{n+1} + qpT_n\right) \\ &= pq\left(pH_{n+2} + qT_{n+2} + (qp+p^2)H_{n+1} + (qp+q^2)T_{n+1} + qp^2H_n + pq^2T_n\right) \end{aligned}$$

and noting that $qp + p^2 \equiv p$, and $qp + q^2 \equiv q$:

$$v_{n+4} = pq\left(v_{n+2} + v_{n+1} + qpv_n\right)$$

To get the initial conditions, with some slightly laborious counting, we can determine that :

$$\begin{aligned} v_1 &= 1 \\ v_2 &= 1 \\ v_3 &= 1 - p^3 - q^3 \equiv 3pq \\ v_4 &= 2p^3q + 6p^2q^2 + 2pq^3 \equiv 2pq(p^2 + 3pq + q^2) \equiv 2pq(1 + pq) \end{aligned}$$

Attempting to solve this we get the auxiliary equation is $\theta^4 - pq(\theta^2 + \theta + pq) = 0$ which does not factorise for θ in terms of p. We can still derive the values for any given values of p however.
Below is a table of probabilities for the first few values of n for $p = 0.4$ and $p = 0.5$ from a spreadsheet using the recurrence relationship above. As it is easy to make mistakes deriving formulae, it seems a good idea to verify results by using a separate mechanism - in this case, a simple Java program which simply 'makes', for each value of n, n tosses of a coin millions of times and counts the proportion which don't contain a run of three consecutive heads or tails.
As you can see, the results suggest that the recurrence relation and initial conditions matches the results from the 'real' sampling and are presumably correct!

(The first 4 values of v_n are 'hardcoded' from the initial conditions, the subsequent values are derived using the recurrence relation)
For $p = 0.5$, the auxiliary equation turns out to be $\theta^4 - \frac{1}{4}\theta^2 - \frac{1}{4}\theta - \frac{1}{16} = 0$ which has the four solutions, 0.8090169943749475, -0.30901699437494745, $-0.25 + 0.43301270189221935i$, $-0.25 - 0.43301270189221935i$.
The complex solutions would always have constants which are designed to cancel out any lingering imaginary components, but regardless, the "largest" solution , i.e. the one whose powers decreases slowest to 0 is 0.8090169943749475 which happens in this case to be $\frac{\sqrt{5}+1}{4}$.
We can see from the table that the successive ratios appear to be converging to this value.
Similar comments apply to other values of p.

	p = 0.5				**p = 0.4**		
n	v_n	**Ratio**	**Java**		v_n	**Ratio**	**Java**
1	1		1.0000000		1		1.0000000
2	1	1	1.0000000		1	1	1.0000000
3	0.75	0.75	0.7499893		0.72	0.72	0.7199326
4	0.625	0.83333333	0.6250490		0.5952	0.82666667	0.5951957
5	0.5	0.8	0.5000507		0.4704	0.79032258	0.4704234
6	0.40625	0.8125	0.4063904		0.373248	0.79346938	0.3732993
7	0.328125	0.80769230	0.3282143		0.297216	0.79629629	0.2971779
8	0.265625	0.80952381	0.2657503		0.23675904	0.79658914	0.2367888
9	0.21484375	0.80882352	0.2149047		0.1880064	0.79408330	0.1878943
10	0.173828125	0.80909090	0.1738321		0.149653094	0.796	0.1496198
11	0.140625	0.80898876	0.1406627		0.119063347	0.79559562	0.1191090
12	0.113769531	0.80902777	0.1138066		0.094675599	0.7951699	0.0946223
13	0.092041016	0.80901287	0.0920047		0.075321115	0.79557050	0.0753025
14	0.074462891	0.80901856	0.0745270		0.059917365	0.79549228	0.0599458
15	0.060241699	0.80901639	0.0602841		0.04765726	0.79538310	0.0476805

2.5 Answer for $R = 4$

This is only slightly more complex than the $R = 3$ case, mainly in the areas of larger formulae and more effort getting the initial conditions. Feel free to skip this section if uninterested, as there's nothing essentially new here apart from perhaps confirming some of the patterns we might have seen from the previous two solutions!
Let A_j represent the event that, in j tosses, no run of four heads or tails occur successively and let $v_j := \mathbf{P}(A_j)$.
Also, define H_j as $\mathbf{P}(A_j \mid \text{first toss is a head})$ and T_j as $\mathbf{P}(A_j \mid \text{first toss is a tail})$.
Conditioning on the result of the first toss for v_{n+6},

$$v_{n+6} = \mathbf{P}(A_{n+6}) = H_{n+6}p + T_{n+6}q$$

But, H_{n+6} is the probability of no run of four occurring in $n+6$ tosses given a head on the first toss, which means that either :

(i) The next toss is a tail (with probability q) and the 'state' is then reset to that in which there are $n+5$ tosses remaining with the "first" toss being a tail

(ii) The next toss is a head, followed by a tail (with probability pq) and the 'state' is then reset to that in which there are $n+4$ tosses remaining with the "first" toss being a tail

(iii) The next two tosses are heads, followed by a tail (with probability p^2q) and the 'state' is then reset to that in which there are $n+3$ tosses remaining with the "first" toss being a tail

So

$$H_{n+6} = qT_{n+5} + pqT_{n+4} + p^2qt_{n+3} \tag{4}$$

Similarly :

$$T_{n+6} = pH_{n+5} + qpH_{n+4} + q^2pH_{n+3} \tag{5}$$

which gives after substituting :

$$v_{n+6} = pqT_{n+5} + p^2qT_{n+4} + p^3qT_{n+3} + pqH_{n+5} + q^2pH_{n+4} + q^3pH_{n+3}$$

And again, using (4) and (5), but with n reduced by 1, and substituting a further time, (and noting that $qp + p^2 \equiv p$ and $pq + q^2 \equiv q$) this leads to :

$$\begin{aligned}
v_{n+4} &= pq\left(pH_{n+4} + qpH_{n+3} + q^2pH_{n+2}\right) + p^2q\left(pH_{n+3} + qpH_{n+2} + q^2pH_{n+1}\right) \\
&\quad + p^3q\left(pH_{n+2} + qpH_{n+1} + q^2pH_n\right) + pq\left(qT_{n+4} + pqT_{n+3} + p^2qT_{n+2}\right) \\
&\quad + pq^2\left(qT_{n+3} + pqT_{n+2} + p^2qT_{n+1}\right) + pq^3\left(qT_{n+2} + pqT_{n+1} + p^2qT_n\right) \\
&= pq\Big\{pH_{n+4} + qT_{n+4} + (qp + p^2)H_{n+3} + (pq + q^2)T_{n+3} \\
&\quad + (q^2p + qp^2 + p^3)H_{n+2} + (p^2q + pq^2 + q^3)T_{n+2} \\
&\quad + (q^2p^2 + qp^3)H_{n+1} + (p^2q^2 + pq^3)T_{n+1} + q^2p^3H_n + p^2q^3T_n\Big\} \\
&= pq\left(v_{n+4} + v_{n+3} + (1 - pq)v_{n+2} + pqv_{n+1} + p^2q^2v_n\right)
\end{aligned}$$

Phew! To get the initial conditions, we need careful counting after which we can determine that (after some simplification) :

$$\begin{aligned}
v_1 &= 1 \\
v_2 &= 1 \\
v_3 &= 1 \\
v_4 &= 1 - p^4 - q^4 \equiv 4p^3q + 6p^2q^2 + 4pq^3 \equiv 4pq - 2p^2q^2 \\
v_5 &= 3p^4q + 10p^3q^2 + 10p^2q^3 + 3pq^4 \equiv pq(3p^3 + 3q^3 + 10(p^2q + q^2p)) \equiv 3pq + p^2q^2 \\
v_6 &= 2p^5q + 12p^4q^2 + 20p^3q^3 + 12p^2q^4 + 2pq^5 \equiv 2pq + 4p^2q^2
\end{aligned}$$

As before, below is a table of probabilities for the first few values of n for $p = 0.4$ and $p = 0.5$ from a spreadsheet using the recurrence relationship above including the results from a 'brute force' Java program to lend some validity to the formulae.
Again, the results suggest that the recurrence relation and initial conditions matches the results from the 'real' sampling and are presumably correct!

	p = 0.5				**p = 0.4**		
n	v_n	**Ratio**	**Java**		v_n	**Ratio**	**Java**
1	1		1.0000000		1		1.0000000
2	1	1	1.0000000		1	1	1.0000000
3	1	1	1.0000000		1	1	1.0000000
4	0.875	0.875	0.8749609		0.8448	0.8448	0.8447761
5	0.8125	0.92857142	0.8126014		0.7776	0.92045454	0.7776422
6	0.75	0.92307692	0.7500013		0.7104	0.91358024	0.7105294
7	0.6875	0.91666666	0.6876102		0.6432	0.90540540	0.6431889
8	0.6328125	0.92045454	0.6329641		0.58263552	0.90583880	0.5825617
9	0.58203125	0.91975308	0.5820986		0.52918272	0.90825688	0.5290458
10	0.53515625	0.91946308	0.5351430		0.48024576	0.90752351	0.4801654
11	0.4921875	0.91970802	0.4922160		0.43582464	0.90750335	0.4357943
12	0.452636719	0.91964285	0.4526049		0.395404444	0.90725582	0.3954265
13	0.416259766	0.91963322	0.4162215		0.358831227	0.90750428	0.3588773
14	0.3828125	0.91964809	0.3828548		0.325627085	0.90746585	0.3255918
15	0.352050781	0.91964285	0.3521412		0.295488553	0.90744464	0.2955003

(The first 6 values of v_n are 'hardcoded' from the initial conditions, the subsequent values are derived using the recurrence relation)
For $p = 0.5$, the auxiliary equation turns out to be $\theta^6 - \frac{1}{4}\theta^4 - \frac{1}{4}\theta^3 - \frac{3}{16}\theta^2 - \frac{1}{16}\theta - \frac{1}{64} = 0$ which has the six solutions, $\pm 0.5i$, -0.5, 0.9196433776070806, $-0.2098216888035 \pm 0.3031453646036i$.
Again, the "largest" solution , i.e. the one whose powers decreases slowest to 0 is 0.9196433776070806 and we can see from the above table that the successive ratios appear to be converging to this value. Similar comments apply to other values of p.

2.6 Comments on solutions

There are some features which are worth noting and probably apply to all values of R.
Firstly, for a given value of R, the final recurrence relationship is of the form $v_{n+2R} = C_{n+2R-1}v_{n+2R-1} + C_{n+2R-2}v_{n+2R-2} + \cdots + C_n v_n$. I suspect that C_{n+2R-1} is always 0.
Secondly, each approach seems to require the double substitution of the formulae for H_i and T_i after which 'magically' allows us to rewrite in terms of v_i.
Thirdly, it is interesting that every recurrence relation and initial condition can be formulated in terms of pq and no other functions of p or q. This is probably because of (a) the obvious symmetry between p and q in the problem, and (b) that many symmetrical functions of 2 variables can often be rewritten as functions of the product and sum of those variables. Since $p + q \equiv 1$, we are only left with the product pq.