1 Introduction

The MURI Thrust 4 group has recently become somewhat interested in performing frequency analysis on the binary-valued ”Did you drink today?” time series. The binary nature of the values and the short duration of the time series both introduce some effects into the time series that require some care. In this document, I'll first summarize the major points to be somewhat concerned about, and then at the end I'll give proofs and explanations of these facts.

2 Features to be aware of

2.1 Fourier series is not interesting if there are only a small handful of drinking days

One can show that if there are zero drinking days, the amplitude of the Fourier frequency components for all frequencies will be zero (not surprising). One can also show that if there is only *one* drinking day, then the amplitude of each and every Fourier component will be exactly the same, and the value will be $2/N$, where N is the length of the time series (in our case 28 days). Thus, the bar of the histogram that contains the value $2/28 = 0.71$ will tend to be over-inflated if there are many individuals who only drank once in 28 days. This problem improves as the number of drinking days increases, but the takeaway is that for respondents with very few drinking days, the Fourier transform is never going to be worth looking at. Given this, for frequency analyses we should likely throw away any respondents with very few drinking days, since their Fourier components will just be confusing to any data analysis.

2.2 Some Fourier components are more multivalued than others

For a time series of 28 entries, there are 15 distinct frequency components in the Fourier series. These are indexed by an integer m, which ranges in value from $m = 0$, which is the 0-frequency or flat-line component of the signal, to $m = 1, 2, ..., 14$, which correspond to frequencies $\omega_m = 2\pi m/N$. Now it turns out that whenever N is divisible by m, the number of possible values for the m'th Fourier component is fewer. This is especially true the smaller the remainder is, so $m = 14$ and $m = 7$ take on fewer possible values than $m = 4$ or $m = 2$, which in turn take on fewer values than $m = 5$ which doesn't divide 28 evenly. This can be seen in Figure 1. We can also see the effect in all of the even plots. In truth, the “choppiness” increases with the number of shared factors in the prime factorizations of N and m. The prime factorization of $N = 28$ is $2 \times 2 \times 7$. So any even number will share a factor of 2, and therefore have some choppiness to it. There's no real way around this, so it's just something you have to be aware of when trying to interpret your data.

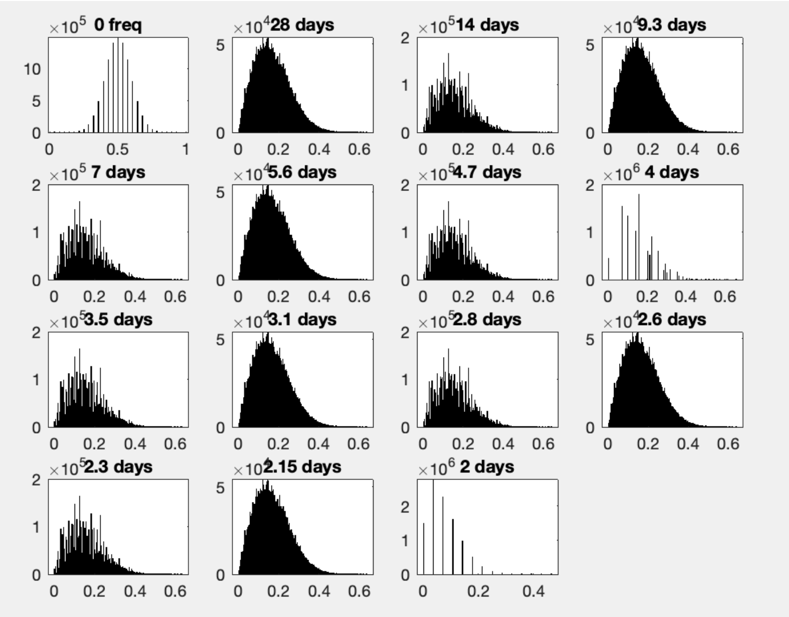

Figure 1: Example distribution of Fourier components for a Bernoulli sequence where each entry of the time series is either 1 or 0 with probability p=0.5. The histograms are of the Fourier components in increasing frequency as read from left-to-right and top-to-bottom. Pay special attention to the "choppiness" of the histogram for all even values, as well as for m=7 (the eight histogram), and m=14 (the last histogram), which are large divisors of N=28.

3 Proofs

First, we have to define how we're even computing the Fourier components. The idea of a Fourier series is that you take your data $y(t)$, which are observed at times $t = t_1, t_2, ..., t_N$, and you want to instead express this series as a sum of oscillating functions with some phase offsets. So, for example, if your data were taken at $t = 0, \pi/2, \pi, 3\pi/2$ seconds, and they took on the values $y = 0, 1, 0, -1$, then it's clear that, rather than represent your data via the ordered pairs $(0, 0), (\pi/2, 1), (\pi, 0), (3\pi/2, -1)$, you could instead represent your data by the function $\sin(t)$, and the time points $t = 0, \pi/2, \pi, 3\pi/2$. It turns out this can generally be done for *any* finite time series, even if the points aren't drawn directly from a single oscillatory function as above. The general idea is as follows: suppose you have a time series $y(t)$ which takes on values at different times t ranging from $t = 0$ to $t = T$. We generally want to reexpress $y(t)$ in terms of a sum of oscillatory functions like sines or cosines. There are a few mathematically equivalent ways of writing this down. The most straightforward is to write

$$y(t) = \sum_{n=0}^{\infty} a_n \cos\left(\frac{2\pi n}{T} t + \phi_n\right), \tag{1}$$

where a_n are positive numbers representing the amplitude of the given frequency component, the increasing values of n correspond to faster and faster oscillating trig functions, and the phase shift ϕ_n allows the trig functions to be properly lined up with one another. The integer values of n have to do with the fact that the functions must all be periodic on the interval from 0 to T. Note that there are a few ways we can rewrite this expression. For example, using the trigonometric identity $\cos(x + y) = \cos(x)\cos(y) - \sin(x)\sin(y)$, we can rewrite the above expression as

$$y(t) = \sum_{n=0}^{\infty} a_n \cos(\phi_n) \cos\left(\frac{2\pi n}{T} t\right) - a_n \sin(\phi_n) \sin\left(\frac{2\pi n}{T} t\right), \tag{2}$$

which can be written more simply as

$$y(t) = \sum_{n=0}^{\infty} A_n \cos\left(\frac{2\pi n}{T} t\right) + B_n \sin\left(\frac{2\pi n}{T} t\right). \tag{3}$$

Note, however, that now the coefficients A_n and B_n may be positive or negative, whereas a_n was assumed positive. A third possible way of writing this same expression is using complex numbers. To do this, note the Euler identities $\cos(x) = (e^{ix} + e^{-ix})/2$, and $\sin(x) = -i(e^{ix} - e^{-ix})/(2)$. Plugging this into the above expression gives

$$y(t) = \sum_{n=0}^{\infty} \frac{A_n}{2}\left(e^{i\frac{2\pi n}{T}t} + e^{-i\frac{2\pi n}{T}t}\right) + \frac{B_n}{2i}\left(e^{i\frac{2\pi n}{T}t} - e^{-i\frac{2\pi n}{T}t}\right). \tag{4}$$

Now we can group these in the following way:

$$y(t) = \sum_{n=0}^{\infty} \frac{A_n - iB_n}{2} e^{i\frac{2\pi n}{T}t} + \frac{A_n + iB_n}{2} e^{-i\frac{2\pi n}{T}t} \tag{5}$$

Now let $C_n = (A_n - iB_n)/2$, and let $C_{-n} = C_n^*$, where x^* is the complex conjugate of x, obtained by flipping the sign of the imaginary part of x. In this notation, we can rewrite the above expression as

$$y(t) = \sum_{n=-\infty}^{\infty} C_n e^{i\frac{2\pi n}{T}t}. \tag{6}$$

All of these expressions are mathematically equivalent, and all appear on occasion, so all are presented here for clarity. That said, we will move forward using expression (3), which is simpler to work with than expression (1), but doesn't require familiarity with complex numbers as does expression (6).

Now we've established this goal of rewriting $y(t)$ in terms of cosines and sines, but we still have all of these unknown coefficients A_n and B_n. In order to calculate these given a particular function $y(t)$, we make use of a clever trigonometric trick. Beginning with expression (3), we multiply both sides by $\cos(\frac{2\pi m}{T}t)$ and integrate both sides from $t = 0$ to $t = T$. That is,

$$\begin{aligned}\int_0^T dt\, y(t) \cos\left(\frac{2\pi m}{T}t\right) = \int_0^T dt \sum_{n=0}^{\infty} & A_n \cos\left(\frac{2\pi n}{T}t\right) \cos\left(\frac{2\pi m}{T}t\right) \\ & + B_n \sin\left(\frac{2\pi n}{T}t\right) \cos\left(\frac{2\pi m}{T}t\right)\end{aligned} \tag{7}$$

Now the order of the integral and sum can be flipped, and the coefficients A_n and B_n don't depend on t, so we really have to evaluate the integrals

$$\int_0^T dt \, \cos\left(\frac{2\pi n}{T}t\right) \cos\left(\frac{2\pi m}{T}t\right), \text{ and} \tag{8}$$

$$\int_0^T dt \, \sin\left(\frac{2\pi n}{T}t\right) \cos\left(\frac{2\pi m}{T}t\right) \tag{9}$$

Now we use the trigonometric identities $\cos(x)\cos(y) = (\cos(x+y)+\cos(x-y))/2$ and $\sin(x)\cos(y) = (\sin(x+y)+\sin(x-y))/2$. These turn the above expressions into

$$\int_0^T dt\, \frac{1}{2}\left(\cos\left(\frac{2\pi(n+m)}{T}t\right) + \cos\left(\frac{2\pi(n-m)}{T}t\right)\right), \text{ and} \tag{10}$$

$$\int_0^T dt\, \frac{1}{2}\left(\sin\left(\frac{2\pi(n+m)}{T}t\right) + \sin\left(\frac{2\pi(n-m)}{T}t\right)\right) \tag{11}$$

Since the integration can be split over addition, there are really only two integrals to consider here. These are

$$\int_0^T dt \cos\left(\frac{2\pi k}{T}t\right), \text{ and} \tag{12}$$

$$\int_0^T dt \sin\left(\frac{2\pi k}{T}t\right), \tag{13}$$

where k is some integer, positive, negative, or zero. Now since these integrals range over a whole number of cycles of sine or cosine, and since sine and cosine are positive and negative in equal portion over a full cycle, these integrals will always be zero, with only one exception. The exception is when $k = 0$ in the cosine integral, in which case the integrand becomes 1, and so the integral evaluates to T. Now $k = 0$ only when $m = n$, which means that in our original expression (7), the only term that survives the integration is the cosine term where $n = m$. Thus, equation (7) becomes

$$\int_0^T dt\, y(t) \cos\left(\frac{2\pi m}{T}t\right) = A_m \frac{T}{2} \tag{14}$$

So we find that

$$A_m = \frac{2}{T}\int_0^T dt\, y(t) \cos\left(\frac{2\pi m}{T}t\right) \tag{15}$$

We could also have multiplied both sides by $\sin(\frac{2\pi m}{T}t)$, in which case we could replicate the above arguments and find

$$B_m = \frac{2}{T}\int_0^T dt\, y(t) \sin\left(\frac{2\pi m}{T}t\right) \tag{16}$$

This essentially completes the necessary theory of Fourier transforms, but there is one more detail to consider. The above expressions assume $y(t)$ is a continuous function of t. In fact, $y(t)$ represents our data vector. Suppose our data vector is obtained by measuring y at even intervals of time Δt from $t = 0$ to $t = T$. We'd rather refer to things using a discrete index n than a continuous time t, so let $t_n = n\Delta t$, where $n = 0, 1, 2, ..., N$, and $N = T/\Delta t$. Also let $y(t) = y(t_n) = y_n$. In this way, the integrals (15) and (16) can be replaced by summations over the index n. That is,

$$A_m = \frac{2}{T}\sum_{n=0}^{N} y_n \cos\left(\frac{2\pi m}{N}n\right), \text{ and} \tag{17}$$

$$B_m = \frac{2}{T}\sum_{n=0}^{N} y_n \sin\left(\frac{2\pi m}{N}n\right). \tag{18}$$

This does, however, introduce a peculiar problem. Namely, suppose $m = N + k$ for some positive integer k. In that case,

$$\cos\left(\frac{2\pi(N-k)}{N}n\right) = \cos\left(2\pi n + \frac{2\pi k}{N}n\right) = \cos\left(\frac{2\pi k}{N}n\right) \tag{19}$$

So why does this matter? Well, it means that the coefficients aren't all unique. The coefficient A_{N+k} will be the same as A_k, as will the coefficients A_{2N+k}, A_{3N+k}, and so on. The solution, it turns out, is to simply ignore $m > N$, so rather than being unbounded, m will range from 0 to N. Actually, there's an even more subtle issue. Namely, suppose $m = N - k$ for some positive integer k. Then,

$$\cos\left(\frac{2\pi(N-k)}{N}n\right) = \cos\left(2\pi n - \frac{2\pi k}{N}n\right) = \cos\left(-\frac{2\pi k}{N}n\right) = \cos\left(\frac{2\pi k}{N}n\right) \tag{20}$$

and for sine,

$$\sin\left(\frac{2\pi(N-k)}{N}n\right) = -\sin\left(\frac{2\pi k}{N}n\right) \tag{21}$$

So there's another redundancy as well. Namely, $A_{N-k} = A_k$ and $B_{N-k} = -B_k$. This redundancy means we will further restrict m to range from 0 to $N/2$ (or $N/2 + 1$ if N is odd). Now, finally, we have all the tools in place to address the concerns raised above about Fourier transforms of our binary, 28-point time series.

3.1 Fourier series is not interesting if there are only a small handful of drinking days

First, let's bring our primary equations back to the front. Namely, the Fourier coefficients are defined by

$$A_m = \frac{2}{T}\sum_{n=0}^{N} y_n \cos\left(\frac{2\pi m}{N}n\right), \text{ and} \tag{22}$$

$$B_m = \frac{2}{T}\sum_{n=0}^{N} y_n \sin\left(\frac{2\pi m}{N}n\right), \tag{23}$$

where $m = 0, 1, ..., N$.

Since y_n is always equal to either 0 or 1, when computing the coefficients A_m and B_m, it essentially just serves to tell us which terms to include in the summation. If no terms are included (i.e. $y_n = 0$ for every n), then it's quite clear that the summation will be zero for every value of m. Next suppose there is exactly one non-zero entry in y_n, and suppose that non-zero entry occurs at $n = k$ for some integer k. Then,

$$A_m = \frac{2}{T}\cos\left(\frac{2\pi m}{N}k\right), \text{ and} \tag{24}$$

$$B_m = \frac{2}{T}\sin\left(\frac{2\pi m}{N}k\right), \tag{25}$$

Note, however, that we are generally interested in the *total power* in a given frequency band. Since sine and cosine can both separately contribute to the same frequency band, we want to combine A_m and B_m into one *total* amplitude. The clue for how to do this comes from our original expression for the Fourier series, expression (1), which gave a_n as a total amplitude of the frequency component with a phase shift ϕ_n. To convert from A_m and B_m to a_m and ϕ_m, refer to expressions (2) and (3). There we see that

$$A_m = a_m \cos(\phi_m), \text{ and} \tag{26}$$

$$B_m = -a_m \sin(\phi_m)\,. \tag{27}$$

We can solve these expressions for a_m and ϕ_m in terms of A_m and B_m, and we find

$$\phi_m = -\frac{B_m}{A_m}, \text{ and} \tag{28}$$

$$a_m = \sqrt{A_m^2 + B_m^2}\,. \tag{29}$$

So we see that the total amplitude of a given frequency component is given by the square root of the squared sum of the sine and cosine amplitudes (often called addition in quadrature). Performing this summation on expressions (24) and (25), we find

$$a_m = \frac{2}{T}\sqrt{\cos^2\left(\frac{2\pi m}{N}k\right) + \sin^2\left(\frac{2\pi m}{N}k\right)}\,. \tag{30}$$

However, the expression inside the square root is simply equal to 1, and so

$$a_m = \frac{2}{T}\,. \tag{31}$$

In our case $T = 28$ days, and so $a_m = 2/28 \approx 0.0714$ days^{-1}. So whenever we have a time series with only a single drinking day, we will find all the frequency components in the Fourier spectrum have amplitude 0.0714. Since this is a mathematical feature, rather than something interesting about our time series, it's something we should be aware of moving forward. Fortunately, this makes some sense, since if we see an event only one time, it's impossible to answer the question "How frequently does the event occur?" It is perhaps comforting to see that our Fourier decomposition assigns equal weight to all frequencies in such cases.

3.2 Some Fourier components are more multivalued than others

As mentioned in the last section, we can think of y_n simply as a filter which decides which terms of the summation we will keep when computing the Fourier components. However, it's important to consider how many unique values the terms in the summation can actually take. For example, consider the $m = N/2$

Fourier component. In this case,

$$A_{N/2} = \frac{2}{T}\sum_{n=0}^{N} y_n \cos(\pi n) = \frac{2}{T}\sum_{n=0}^{N} y_n(-1)^n \tag{32}$$

From this expression, it's clear that $A_{N/2}$ can only take on a fairly limited number of possible values. Namely, $A_{N/2}$ must be an integer, since it's a sum of $+1$'s and -1's.

On the other hand, consider $m = 1$. In this case

$$A_1 = \frac{2}{T}\sum_{n=0}^{N} y_n \cos\left(\frac{2\pi n}{N}\right) \tag{33}$$

In this case, the terms of the sum can take on around $N/2$ unique values (recall that the $n = N - k$ symmetry means roughly half the values are equal to one another). Because the terms of the sum can take on many more possible values, the possible outcomes are the summation are also much more numerous, and so a histogram of the Fourier components for different time series will be much more smoothed out for A_1, whereas for $A_{N/2}$, it will contain several discrete possible values with large gaps in between these. In general, the more unique values there are in the terms of the summation, the smoother the histograms will look. So how many unique values are there for a given m and N value?

Consider $\cos(\frac{2\pi m}{N}n)$. We want to count how many unique values this quantity takes on as n ranges from 0 to N, for a given choice of m and N. First, suppose m and N have no common factors at all. In this case, the only symmetry is the $n = N - k$ symmetry discussed above, so there will either be $N/2$ unique values if N is even, or $N/2 + 1$ unique values if N is odd. Next suppose m and N share a common integer factor of k. That is, $m = ak$ and $N = bk$ for some integers a and b. Then $\cos(\frac{2\pi m}{N}n) = \cos(\frac{2\pi a}{b}n)$. This introduces a new symmetry. Namely, if $n = b + k$, then $\cos(\frac{2\pi a}{b}n) = \cos(\frac{2\pi a}{b}(b + k)) = \cos(2\pi a + \frac{2\pi a}{b}k) = \cos(\frac{2\pi a}{b}k)$. This means that, in terms of unique values, n may as well only range from 0 to b, since whenever n is greater than b, it just revisits a value that has already been seen previously. Since the $n = b - k$ symmetry also still holds, this means that the total number of unique values will be $b/2$ if b is even, or $b/2 + 1$ if b is odd.

Thus, the total number of unique terms in the summation can be found in the following way: beginning with the fraction m/N, simplify it to a form in which m and N have no more common factors. Call this simplified fraction a/b. Having done this, the number of unique terms in the summation will be either $b/2$ or $b/2 + 1$, depending on whether b is even or odd. The larger the number of unique values, the smoother will be the resultant histograms. If $N = 28$, as it does in our data set, this means that any of the odd Fourier components will be maximally smooth, while Fourier components which are divisible by 2, 4, or 7 will be increasingly discretized in the histograms.