SOME ELEMENTARY NOTES ON MEASURE AND INTEGRATION.

W.W.Sawyer.

Students may sometimes have the impression that mathematicians produce elaborate theories for no particular reason apart from a kind of professional etiquette. Such questions as "What do we mean by a number ?" or "What is area ?" do not strike the average citizen as being among the more urgent questions of the day. And indeed, as a matter of history, very little attention was given to such questions, even by mathematicians, until the 19th century, by which time mathematics had reached a stage of considerable elaboration. Sometimes, one must admit, a mathematician is led to investigate an apparently simple idea because he has formed the habit of systematic exposition ; after carefully developing some theory with precise definitions of every concept introduced, a lecturer or author may well be reluctant to change key and say,"Now we come to area and of course you all know what that is." But a more positive and compelling force is also at work. It may happen that, in the course of some investigation, a mathematician needs to determine the area of some region, the structure of which is so complicated that he begins to doubt whether it even has an area , and to wonder, if it has, however that area is to be defined and calculated. In such a situation, a theory of area is not a luxury, nor even simply a way of satisfying oneself that the work is logically sound ; it is a practical necessity,without which no further advance is possible. The Lebesgue theory of measure and integration is an excellent example of a branch of mathematics that has developed under the pressure of such necessity.

In the early stages of calculus a student is not aware of any such pressure. He accepts without hesitation that there is an area under the parabola $y = x^2$ between $x=0$ and $x=1$ and that $\int_0^1 x^2\,dx$ gives this area. Yet even in a first course of calculus, the first signs of complication begin to appear. Suppose that in Figure 1 he wishes to calculate the length of the arc ABC, which is half the circumference of the circle $x^2+y^2=1$. He knows he should obtain the answer π . The standard formula for the length of an arc leads him to write the integral

$$\int_{-1}^{1} (1-x^2)^{-1/2}\,dx ,$$

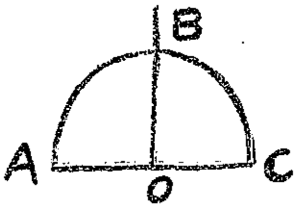


Figure 1.

the integrand of which is infinite at both ends of the interval, corresponding to the fact that the tangents at A and C are vertical.

If we interpreted this integral as an area, we would arrive at the diagram shown in Figure 2, and be led to conjecture that, although the shaded region extends to infinity, it yet has the finite area π .

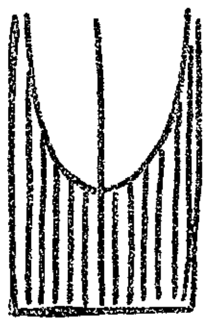


Figure 2.

We can give a formal definition of what we mean by the area of the shaded region by following Cauchy's ideas. We suppose a to be a little to the right of -1, b a little to the left of +1, and consider what happens to the area under the curve between a and b in the limit when a tends to -1 and b to +1.

If, as in Figure 3, a graph has a number of discontinuities, we can cope with the situation by evaluating the integral in the intervals between the discontinuities and adding the results.

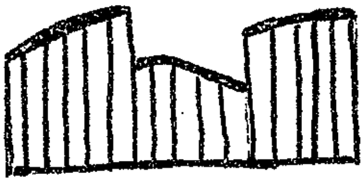


Figure 3

This approach, whether we are dealing with infinities or finite discontinuities, depends on there being intervals, between the exceptional points, in which the function is finite and continuous. But the development of mathematics led quite naturally to the study of functions whose discontinuities were <u>everywhere dense</u>, that is to say, so crowded together that you could never find space between them big enough to hold an interval. The first person to realize that sush a possibility existed was Riemann in 1854. "Riemann Theory of Integration" figures in most introductory courses in analysis, and it is a great pity that expositors rarely mention what led Riemann to work out his theory. He did not set out to investigate integration ; he was writing a paper on Fourier series,[1] a subject of great importance both in pure and applied mathematics, and one in which integration is a standard tool. In this work he arrived at some functions of an entirely novel kind, and had to pause to show that the usual procedures of Fourier series, involving integration, could in fact be applied to these functions.

Fourier series originally arose, around 1730, in connection with musical vibrations. The expression $b_1 \sin t$ represents a pure tone ; $b_2 \sin 2t$ corresponds to its first harmonic , $b_3 \sin 3t$ to the next harmonic, and so on. When a musical instrument is played, all the harmonies are liable to be produced simultaneously, corresponding to an expression of the form

$$b_1 \sin t + b_2 \sin 2t + b_3 \sin 3t + \ldots\ldots\ldots$$

The characteristic sound of an instrument depends on the ratios between the coefficients b_n in this series. Similar terms involving cosines may also appear. Such series are known as Fourier series, although they had been studied long before the time of Fourier.

[1] Paper XII in "Collected Works of Bernhard Riemann" (Dover). This paper is in German.

One of the great controversies of the 18th century was concerned with the question of what functions could be represented by Fourier series. This question was vital for applications of Fourier series made at that time and since then. For instance, could a Fourier series represent a graph made up of broken lines, of the kind shown in Figure 4 ? Eminent mathematicians, notably Euler and d'Alembert, refused to believe it could. Yet in fact Fourier series can be found which represent this and other much more complicated functions. In Figure 5 we see pieces of a line, a parabola and a circle, and these pieces do not even join together to give a continuous function. Yet a Fourier series exists which has the graph shown in Figure 5.

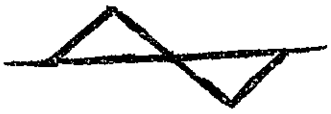


Figure 4.

It will be noticed that Figure 5 possesses a certain symmetry. This symmetry is due to the fact that, for purposes of illustration, it is quite sufficient to consider series consisting of sines alone. As $\sin (2\pi - k) = - \sin k$ for every k, a function representable by sines alone is bound to have this type of symmetry.

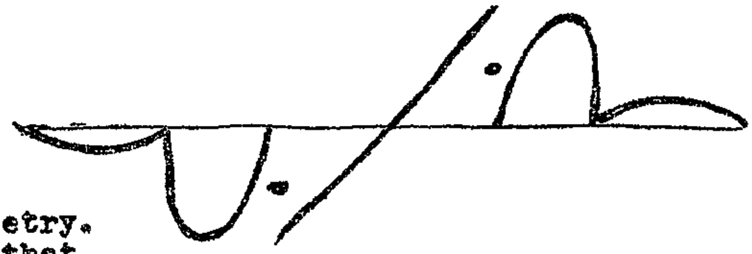


Figure 5.

In Figure 5 a dot will be noticed in the middle of each of the jumps. This is typical of the way Fourier series behave ; when they make a jump, they put a stepping stone half way across.

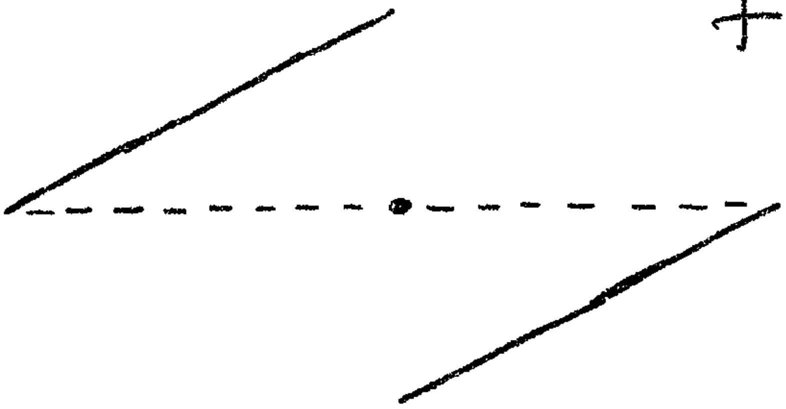
The importance of Fourier series forced mathematicians to study graphs much wilder and more irregular than anything ever considered before. Riemann was the first person to realize that a Fourier series could have a graph that could not even be drawn, because its discontinuities were so crowded that the pencil never had a chance to move along the surface of the paper. He gave an example of such a function in the paper specified in the footnote on page 2. The example we shall now give is in essentials the same as Riemann's. We have modified Riemann's example in such a way as to make the work arithmetically simpler.

As a first step towards constructing our example, we use a standard result of Fourier series. Consider f(x) where

$$f(x) = \sin x - (1/2) \sin 2x + (1/3) \sin 3x - (1/4) \sin 4x + \dots$$

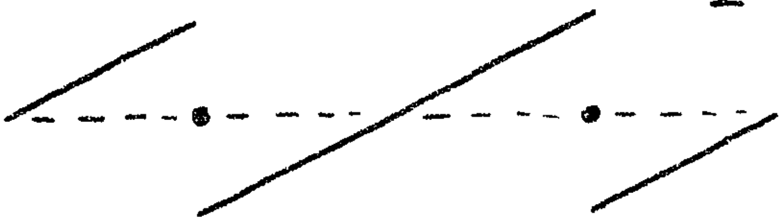
It is well known that the graph of f(x) in the interval $(0,2\pi)$ is as shown in Figure 6 (a) on page 4 of these notes. Outside the interval $(0,2\pi)$ this graph is repeated indefinitely, since the function has period 2π. Figure 6(b) shows the graph of $(1/2) f(2x)$, with both the vertical scale and the period half of what they were for f(x). Similarly 6(c) shows the graph of $(1/4) f(4x)$.

$f(x)$



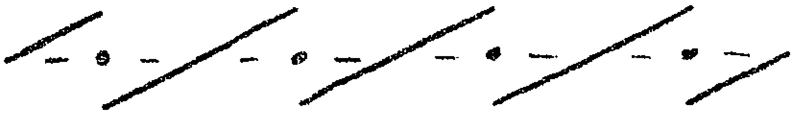
6(a)

$\frac{1}{2}f(2x)$



6(b)

$\frac{1}{4}f(4x)$



6(c)

As Figure 6(a) shows, $f(x)$ has a discontinuity in the middle of the interval. In 6(b) discontinuities are seen 1/4 and 3/4 of the way across ; in 6(c) the jumps occur at 1/8, 3/8, 5/8 and 7/8 of the way across. For the function

$$2^{-n} f(2^n x)$$

discontinuities correspond to each fraction of the interval with the form $(2k+1)/2^n$ where k is a whole number.

We now consider the function $F(x)$ defined by means of a uniformly convergent series ; -

$$F(x) = f(x) + (1/2)f(2x) + (1/4)f(4x) + ... = \sum_0^\infty 2^{-n} f(2^n x)$$

It is easily seen that $F(x)$ has a discontinuity at every fraction of the interval for which the denominator is a power of 2 and the numerator an odd number. This is the example we promised of a function whose graph cannot be drawn with a pencil, though we can plot as many individual points on it as we may wish.

Now each term in the above sum can be expanded as a series of sines. It is suggested that $F(x)$ itself may have such an expansion. The argument is put somewhat conservatively, since adding an infinite number of infinite series is a process in which paradoxes can arise.

The standard procedure for finding the Fourier series $\sum b_n$ sin nx which corresponds to (and in suitable cases converges to) a given function $\phi(x)$ is to use the formula

$$\pi b_n = \int_0^{2\pi} \phi(x) \sin nx \, dx$$

We are thus naturally led to ask whether this standard procedure can be applied to $F(x)$. The discontinuities of $F(x)$ will cause discontinuities in the product $F(x)$ sin nx which appears as the integrand in the above formula, when $F(x)$ is substituted for $\phi(x)$. This is the problem Riemann attacked and solved - to find a definition of integral that remained valid for such a function, with discontinuities everywhere dense. Riemann's definition in fact copes successfully with $F(x)$ but not with every discontinuous function, and not even with every function that can arise in the study of Fourier series.

Riemann's paper thus had two effects. First, it called attention to the fact that a harmless-looking series such as $\sum b_n$ sin nx, which any physicist or engineer would accept as a reasonable expression (not something cooked up by a pure mathematician just to show how complicated things could be) - that such a series could behave in a much more irregular way than one would expect, and that we should not restrict our investigations to smooth, continuous functions only. Secondly, since Riemann's definition eventually proved not even sufficient for Fourier theory, it stimulated a search for even more-embracing definitions of integral.

In elementary calculus $\int_a^b f(x)\,dx$ is thought of as measuring the area under the graph of f(x) between x=a and x=b. The region under the graph may be regarded as a set of points ; a point (x,y) belongs to this set if it satisfies the two conditions $a \leqslant x \leqslant b$, $0 \leqslant y \leqslant f(x)$. This set is called the <u>ordinate set</u> of the function f(x) for the interval (a,b). In beginning calculus the upper boundary of this region is usually a smooth curve. However the specification just given in no way requires this to be so. Given any function, we can construct the ordinate set by taking each point (x,0) in turn and placing on it an upright line of length f(x). (For the present, it will be sufficient to consider functions that do not take negative values.) For a very discontinuous function, such as that having f(x) =1 for rational x, f(x) = 0 for irrational x , we cannot draw the ordinate set but we can still imagine the construction being carried out. We shall be able to define the integral of such a function if we can find some way to explain what we mean by the area of its ordinate set.

The problem of <u>measure</u> is to find definitions of volume,area and length that still work for complicated and irregular collections of points. As an integral corresponds to a particular question about area, solving the general problem of measure automatically solves the problem of defining integration.

Even in elementary school the problem of measure arises, implicitly if not explicitly, when we talk about the area of a country, an oak leaf or a country. For area is measured in square inches,square centimetres or square miles ; Canada, a leaf and a circle are equally unsuitable shapes for dissection into square pieces. Of course there is no difficulty in finding the area of a region that can be broken into a number of rectangles. Both in practice and in theory, the area of a region with curved boundaries is found by means of over- and under-estimates based on shapes built from rectangles. In Figure 7, the region bounded by heavy lines gives us an over-estimate of the area of a quarter circle, while the shaded region gives us an underestimate. We define as the area of the quarter circle the common limit to which the over-estimates and the under-estimates tend ; it is the infimum of the over-estimates and the supremum of the under-estimates. Figure 7, in essentials, indicates the basic idea in Riemann's definition of an integral. Riemann's contribution lay in showing that this simple idea could cope with unexpectedly complicated cases, such as that of F(x) discussed earlier.

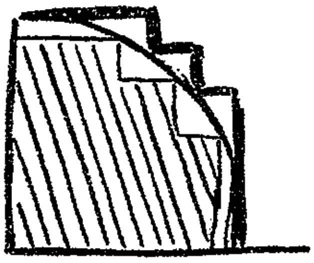


Figure 7 .

We can give an example of a function, for which the Riemann theory of integration is inadequate, without using any advanced idea. Let B(x) = 1 when $x = (2k+1)/2^n$ where k and n are whole numbers, B(x) = 0 for all other x. We now try to define $\int_0^1 B(x)\,dx$. To get the ordinate set we put vertical lines of unit length for x = 1/2, 1/4, 3/4, ... Figure 8 is meant to suggest the result of doing this. To get an over-estimate of the area of this ordinate set, we have to cover all these lines with rectangles. It is fairly clear in doing this we are bound to cover the square with corners (0,0),(1,0),(0,1),(1,1). This square by itself, in fact, gives the most economical over-estimate, namely area 1. How are we to get an under-estimate ? In Figure 7, we got an under-estimate by putting a number of rectangles inside the unit circle. However we cannot put any rectangle at all inside the ordinate set in Figure 8. A situation of this kind frequently arises, and a device has been found to cope with this difficulty. We consider the points of the unit square that do not belong to the ordinate set ; by covering these with rectangles we find an over-estimate for their area ; we then subtract this over-estimate from the area of the unit square. This gives us an under-estimate for the area of the points that are in the ordinate set.

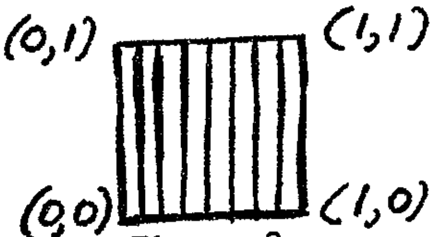


Figure 8.

Now the points not in the ordinate set correspond to those values of x which are either rational numbers with denominators not powers of 2, or irrational. These points also lie on vertical lines which effectively give a shading of the unit square. The lowest over-estimate we can find for them is also 1. As 1-1=0, the best under-estimate we can make for the area of the ordinate set is 0.

Combining these results, all we can say about the ordinate set is that its area is not less than 0 and not more than 1, which is neither a surprising nor a helpful conclusion. We are unable to attach any particular number to $\int_0^1 B(x)\,dx$ by this procedure ; this integral is in fact, as we promised earlier, an example of an integral for which Riemann's definition is inadequate. The definition of integral however can be extended to make this integral meaningful ; this is achieved by Lebesgue's definition of integration, published in the years 1901 and 1902. Some idea of the difficulty of the problem may be gathered by noting that nearly half a century separates the work of Riemann from that of Lebesgue. In the intervening years many good mathematicians worked on the theory of integration but failed to find the key that would unlock the problem.

The contribution of Borel.

The decisive new idea that this problem demanded appeared for the first time in Borel's book "Leçons sur la Théorie des Fonctions", published in 1898. The Russian historian, I.N. Pyesin, in his book on the development of the concept of the integral, comments on the almost casual way Borel presented this new idea. Borel took only 3 pages to sketch his theory of measure ; the thoughts are presented clearly but without detailed proofs, in section 3 of Chapter 3 of his book.

Borel seems to have arrived at this new idea in the course of attacking a problem that had nothing to do with integration. An irrational number can be approximated by a rational number. For instance 1, 2/3, 5/7, 12/17, 29/41 are increasingly good approximations to $1/\sqrt{2}$. Liouville in 1844 investigated how good such approximations were. From his very general theory we select a very special result, namely that $1/\sqrt{2}$ differs from 2/3 by more than $(1/2)/3^3$, from 5/7 by more than $(1/2)/7^3$, from 12/17 by more than $(1/2)/17^3$ and so on ; in fact, any fraction p/q differs from $1/\sqrt{2}$ by more than $(1/2)/q^3$, which we may write more conveniently as $1/2q^3$.

Liouville's results were obtained by detailed considerations of continued fractions. Borel found a simpler way of looking at certain broad aspects of Liouville's work. The result given in the previous paragraph could be formulated as follows. Suppose we start with the interval [0,1], and for each rational number p/q we chop out all the numbers that differ from p/q by $1/(2q^3)$ or less. Thus we would remove the interval [7/16, 9/16] consisting of all the points within 1/16 of 1/2. Similarly we would remove the interval [17/54, 19/54] centred on 1/3, and so on. In the process every rational number would be removed, together with an interval surrounding it. One might well guess that nothing would remain. However Liouville's theorems show that $1/\sqrt{2}$ and various other numbers would survive this operation. Borel calculated the total length of the intervals removed. There are q-1 fractions with denominator q, so we remove q-1 intervals of length $1/q^3$. Together these amount to less than $1/q^2$. As q runs through the numbers 2,3,4..... in turn, the total length removed is less than

$$1/2^2 + 1/3^2 + 1/4^2 + \ldots\ldots\ldots$$

which is known to be $(\pi^2/6) -1$ or approximately 0.645. As this number is less than 1, it would indeed be remarkable if nothing remained of the original interval, of length 1, after removing the intervals. Borel thus found a simple argument showing that results <u>of the kind</u> Liouville obtained ought to be expected.

Borel went further. He showed that the set of points remaining after the removal of all the intervals was actually uncountable. The proof is by reductio ad absurdum. Suppose the opposite to be true. Then the set of points is countable, and the points can be arranged in a sequence a_1 , a_2 , a_3 , a_4 , ... Imagine these points marked on a line. We take a short piece of thread, say of length 0.1 . We cut it in halves and glue one piece to the line so that it covers a_1 . The remaining piece we again bisect, and use one of the halves to cover a_2 . We continue in this way ; at each stage we use only half of the balance in hand to cover the next point. Thus we never use up all our thread at any finite stage of the process. When the process is complete - if one can concede the completion of an infinite process - every point a_n has been covered with thread. Thus all the points surviving the removal of the intervals would have been covered with a length of only 0.1. But the intervals removed had a total length of only 0.645. Thus it would seem that the whole of the unit interval can be covered with a length of only 0.745. One naturally assumes this to be impossible, and we have reached the required contradiction. Borel gave a proof that the argument used here was in fact perfectly rigorous ; his proof used the principle often known as the Heine-Borel Theorem.

Borel's argument above is entirely based on the idea of <u>length</u> ; it is thus relevant to measure theory. The departure it makes from all previous work is that it envisages covering a set with <u>an infinite number</u> of pieces of material, whereas earlier theories had allowed only a finite number of pieces, - in the line, a finite number of intervals ; for areas, a finite number of rectangles. Borel saw that fruitful results came from arguments, such as that just sketched, in which coverings by an infinity of pieces were accepted. He analyzed the properties of length that he had implicitly assumed in such arguments, and devised a definition of measure that would ensure the existence of these properties. (See footnote 1, page 48, of the 1950 edition of Borel's book cited above.)

In the attitudes of mathematicians to infinity one can discern a swing of the pendulum. The ancient Greeks, logically cautious, avoided its use. The 17th and 18th century mathematicians, in the excitement of developing the calculus, used infinity with reckless abandon. In reaction against this, 19th century mathematicians developed a horror of infinity, and built analysis on the basis of statements about finite numbers. It was probably this dread of infinity that inhibited mathematicians from anticipating Borel's ideas. Cantor's work, bitterly opposed, represented the beginning of the swing towards a more naive and trusting use of infinity ; in the main we are still under the influence of Cantor. Future generations may well see further oscillations of this pendulum. However that may be, it is clear that the procedure of Borel, developed and worked out in detail by Lebesgue, gave mathematicians new methods that were both powerful and extremely convenient.