Edward Simpson: Bayes at Bletchley Park

Edward Simpson CB ceased being an active statistician in 1947, when he joined the Civil Service. But statistics owes him much. He is the Simpson of Simpson's index of diversity[1] and of Simpson's paradox[2], the bizarre apparent contradiction which he published in 1951 and which has puzzled students of statistics ever since. Perhaps more importantly, for the world as well as for statistics, from 1942 to 1945 he was a code breaker at Bletchley Park, where Alan Turing and others broke enemy ciphers and the world's first modern computer was developed. Here **Edward Simpson** tells the hitherto unpublished story of the part that Bayesian statistics played in breaking two of the enemy ciphers.

It is now widely, though not yet universally, understood that the world's first large-scale electronic digital computer was created at Bletchley Park during the Second World War. The introduction there of Colossus in late 1943 transformed the cryptanalytic attack on the German teleprinter cipher that the codebreakers called Tunny, and enabled it to be read.

Tunny was even more complex than the better-known Enigma. The machine that enciphered it was made by the Lorenz company. Its size meant that it was not a portable device like Enigma. It was used exclusively for the most important messages passing between the German High Command in Berlin and the Army Group commanders across Europe.

It took people who were conceptually and technically brilliant to break it. To name only three of them: Tunny's enciphering system was worked out, without anyone ever having seen the machine, by Bill Tutte; the concept and specification of high-speed electronic processing of the cryptanalysis and the leadership of its

application were due to Max Newman, both of these at Bletchley Park; Colossus itself was designed and built by Tommy Flowers, on his own initiative in the first instance, at the Post Office Research Station at Dollis Hill. Colossus first ran operationally at Bletchley Park on 5 February 1944[3].

Colossus itself is becoming better known and better documented. Scarcely known, on the other hand, is the use made of Bayes' theorem in several of Bletchley Park's areas of work, including the breaking of German Naval Enigma in Hut 8. Ralph Erskine writes in his *Action This Day* of "a highly sophisticated Bayesian system"[4]; Hugh Alexander, the chess champion who was a prominent member and later head of Hut 8, in his unpublished "Cryptographic History" passes it by[5].

I have been able to reconstruct what the Bayes contribution to the attack on Enigma must have been. Bayes was used on Tunny too, but I do not have the detail on that. But here, because I know it at first hand, I describe the use made of Bayes in the cryptanalytic attack on the main Japanese Naval cipher JN 25 in 1943–1945, by the team in Block B which I led.

Bayes and Japanese Naval 25

Enciphering by JN 25 was a two-stage process. Stage one was by a codebook, stage two by a printed set of numbers known as an enciphering table. The sender and the receiver of messages each had the codebook and the enciphering table. Using the codebook, the text of the message was transformed into numbers – in this case a series of five-digit code groups. Thus the word "maru" (ship) might be represented by the code group 70863. As a check against error, the codebook used only groups where the sum of the five digits was divisible by 3. Colloquially, such groups "scanned". The enciphering table was a vast array of random five-digit numbers called "additives". At the second stage the code groups were enciphered by placing alongside them a same-length series of consecutive additives taken from somewhere in the table; and then adding (non-carrying, i.e. $8+7=5$, not 15) each pair to produce the enciphered groups (see Figure 1).

Once an initial break had been achieved the task of the cryptanalysts at Bletchley Park (and their American and Australian counterparts) was to work out what the additives were – and to do it as quickly as possible so that intercepted messages could be deciphered. The parallel task of the book builders (not discussed here) was to work out the textual meanings of the code groups as in the Japanese codebook.

The cryptanalysts typically had on their table a "depth" of intercepted messages, anything from two to twenty, known to have been enciphered on the same stretch of additive, though not all starting at the same place. Correctly aligning the messages one below the other (which we could do – we knew the Japanese system that indicated it) produced "columns" of groups all enciphered by the same additives at their heads. The main method of recovering the additives started from the observation that the (non-carrying) difference between two enciphered groups in a column will be the same as the difference between the code groups underlying them, because the enciphering additive (common to the two) cancels out.

As a tool at their disposal the cryptanalysts had a body of "good groups" known to be used in the codebook because they had appeared in messages already successfully deciphered, together with their frequencies of occurrence. The best of these known good groups (say, the most frequent 100) were differenced one against the other, choosing each time the alternative below 55556. The resulting 4950 differences were sorted into numerical order, each accompanied by the two good groups that had produced it, and tabulated. For the production of this "index of differences" we relied on the massive Hollerith installation that served all Bletchley Park. This was sorting and tabulating machinery that worked by reading the holes punched in cards. Herman Hollerith had devised the system half a century earlier to process data from the US Census of 1890.

Taking from the depth a column of enciphered groups (say, six deep) and differencing them one against the other produced a set of differences (15 of them in this example). Each was looked up in the index. If it was there – a "click" – inserting back into the column the two good groups tabulated with it in the index led by subtraction to the identification of a speculative additive to go at the head of that column (see Figure 2). The speculative additives thus generated then had to be tested to winnow the genuine from the false.

The first test was a simple one. Take A as the speculative additive to be tested. "Stripping" down the column meant subtracting A (non-carrying) from each enciphered group in turn, to produce speculative deciphered code groups P, Q, R, …, including of course the two from the index. If all of P, Q, R, … scanned, speculative additive A survived into the next test.

Among P, Q, R, … some would be known as good groups and others not. The essential judgement to be made was whether the collective evidence of the good groups appearing amongst P, Q, R, … was sufficiently convincing for A to be accepted as genuine and written in as "column solved". This judgement was as subtle as the scanning test was simple.

Those doing this work, a mixture of civilians and Wrens, were quick and accurate but not mathematically trained. Means had to be found of enabling them to make the judgements quickly and objectively, of standardising the judgements across the team and of initiating new recruits (for the team was expanding rapidly) with the least delay.

The mathematicians in the JN 25 team were Ian Cassels (later Fellow of Trinity and Sadleirian Professor at Cambridge), Jimmy Whitworth and myself. Our solution for the required judgement process started with an application of Bayes' theorem.

The hypothesis to be tested was that A was true – that that this speculatively deduced additive was a correct one. The events were the speculatively deciphered code groups P, Q, R, … (all scanning.) Suppose that Q was a good

Encoding using JN 25

Message To Be Sent (Message One):	MARU	GOOD	WEATHER
Code groups (from codebook; one five-digit group for each word; each group "scans", i.e. is divisible by 3)	70863	34131	30525
Additives (random five-digit numbers from printed encoding table)	58304	68035	91107
Enciphered groups (code groups plus additives, non-carrying)	**28167**	**92166**	**21622**
This is the encoded Message One, as sent by the Japanese radio operator			

Message Two:	SUPPLIES	STOP	YOUR
Code groups (scanning)	83010	50418	29931
Additives (this part of message is from same part of encoding table)	58304	68035	91107
Enciphered groups (code groups plus additives, non-carrying)	**31314**	**18443**	**10038**
This is the encoded Message Two			

Figure 1. Encoding using JN 25

Breaking code JN 25

We know from previous decipherments that groups "34131" and "50418" are used by the Japanese in their codebook – they are "good", i.e. known, groups. (They may or may not have had their meanings identified, as "Good" and "Stop", but that is another issue.) The difference between them (non-carrying) is 26387 – recorded as one of many in our code-breaking "index of differences", together with the groups 34131 and 50418 that generated it.

To identify an additive, set the messages as intercepted, one above the other, and subtract (non-carrying):

31314	**18443**	**10038**
28167	**92166**	**21622**
13257	26387	99416

(The last number is greater than 55554. We therefore replace it with 21622 – 10038 = 11694.)

Look for a result that appears in our "index of differences" – here the red number. We speculate therefore that the messages may contain groups 34131 and 50418 – in which case 18443 – 50418 = 68035 (or 92166 – 34131 = 68035) will be an additive in the Japanese encoding table. But this is only a speculation. Bayesian analysis will be needed to support or refute it.

Figure 2. First stage of decipherment: identify an additive

group. The probability of Q occurring if A was true, p_t, was derived from the assembled body of good groups and their frequencies of occurrence. If A was false, Q was just a random scanning five-digit number with probability p_f of 3:100 000. The weight of the evidence provided by Q in favour of A being true was the Bayes factor p_t divided by p_f.

For practical purposes there was no need to agonise over the prior odds to be assigned to the hypothesis of A's truth. The evidence quantified in the factor was sufficient. Similarly, if (say) R was not a good group, this event would have a probability a little less than random if A were true and a resulting factor a little below 1. But at a stage when only a small fraction of the groups in the codebook were known, the deviation below 1 was going to be small; so, trading this small degree of accuracy for speed and simplicity, R's factor was taken as 1 and its appearance effectively ignored.

Multiplying together the several factors derived from the whole series P, Q, R, … now gave a composite factor quantifying the whole column's evidence for or against the truth of A. With these factors, objective comparisons could be made: whether one speculative additive that produced one very strong good group but little else was more or less plausible than another that produced a string of middling ones. Thresholds were then set, empirically, and varied in the light of experience, for the size of composite factor that would justify confirming a speculative additive as true or sending it for more detailed study. A balance was struck between confirming enough additives to make fast progress and confirming so many that too many errors occurred.

The next stage was to replace each good group's factor by that factor's logarithm; and, again trading a degree of precision for speed and simplicity, to scale and round the logarithms to a set of two-digit whole-number "scores". A "score book" was tabulated, giving the score for each known good group. Time-consuming multiplication was thus replaced by addition simple enough to be done mentally or by pencil jotting. And the thresholds were similarly transformed into their scaled logarithmic equivalents.

From this preparatory work a simple procedure emerged. The streamlined job, as performed by the civilians and Wrens, was to:

- strip down a column by subtracting the speculative additive from each enciphered group in turn;
- check whether the resulting deciphered groups scanned; if they all did, look each one up in the score book and note its score if it was there;
- add the scores and, if the total reached the threshold, rejoice, write in that additive as confirmed and move on to the next.

The system performed very satisfactorily its intended purpose of quickly and systematically testing masses of speculative additives and picking out those that were probably true. It was never seen as doing the whole job. Borderline cases, or messages of particular concern, could be handed over for more intensive study elsewhere in the team. An observant eye and a keen memory could often find significance where the arithmetic alone failed to. And, as always in cryptanalysis, the imaginative hunch grounded in experience could sometimes make the most important contribution of all.

Diversity at Bletchley Park

A huge diversity of minds was engaged at Bletchley Park. It would be a mistake to suppose that the cryptanalysts were all mathematicians.

A German Enigma coding machine. Courtesy of Bletchley Park Museum

If the frequencies of the 26 letters of the alphabet in plain language are $f_1, \ldots, f_{26}$, the probability that two letters picked independently will be the same – that is, will match – is

$$\frac{\sum_1^{26} f^2}{\left(\sum_1^{26} f\right)^2}$$

This is the ratio of repetition or repeat rate: the obverse of diversity. It has much in common with Udny Yule's 'characteristic' based on word frequencies which he devised when addressing the question of Thomas à Kempis's authorship of *De Imitatio Christi*[7]. I called on Udny Yule in St John's College, Cambridge in 1946: a tiny figure in a skull-cap in a huge and lofty Fellow's Room. We talked about his characteristic as a means of analysing language. Of course Bletchley Park was not mentioned. But, given how many Cambridge mathematicians had been at work there, I wondered whether he knew more than he was supposed to.

Using letter frequencies quoted in Wikipedia, we get repeat rates of 1 in 15.3 for English and 1 in 13.1 for German. The Hut 8 team used 1 in 17, presumably derived from a large sample of German Naval messages. (Naval language will differ from normal language.) Two letters picked independently at random have a repeat rate of 1 in 26. Thus the event of a single letter matching between the two messages will have a probability of 1/17 if the hypothesis of a correct alignment is true, and a probability of 1/26 if it is false. Bayes' theorem then tells us that the prior odds on the hypothesis will be multiplied by a factor of 26/17 or 1.53 to give the posterior odds after that event. The alternative event of letters not matching between the two messages will have a probability 16/17 if the hypothesis is true and a probability 25/26 if it is not; giving a factor of (16×26)/(17×25) = 0.979, slightly reducing the prior odds on the hypothesis. Moreover, because successive events along the overlap of the two messages are for practical purposes independent, their factors can be multiplied together to give a composite factor for the alignment as a whole.

Take as an example the testing of an alignment of two messages with an overlap of 32 letters, which yielded 7 pairs that matched and 25 that did not. The composite Bayes factor for this is $1.53^7 \times 0.979^{25} = 11.5$ in favour of this alignment being correct. Prior odds of 1 to 49 (because there were 50 equally likely alignments before the event) become posterior odds of 1 to 4.3.

To simplify the handling of the many Bayes factors produced, at the speed required, Turing brought over the decibel unit[6]. This is familiarly used for measuring loudness, but is not confined to that. Its generalised definition is one-tenth of the base-10 logarithm of the ratio of two measures of any quantity that can be measured. Using the decibel unit brings the simplification that multiplying the ratios gives way to more simply adding their decibels. So it was precisely apt for handling the Bayes factor, the ratio of two probabilities. Extending the language adopted from Banbury (see main text), Turing changed the unit's name from decibel to deci*Ban*. Later, trading a degree of precision for facility and speed, the unit used was changed to the half-deciBan or hdB and the measures were rounded to whole numbers.

Continuing the example above of 7 matches in a 32-letter overlap, the base-10 logarithm of the 11.5 factor was 1.061, so it measured as 21.2 hdB which rounded to 21. This was called the 'score' of the alignment.

Classicists abounded too. Dilly Knox, the star cryptanalyst of Room 40 in the First World War and of Bletchley Park in the Second, was first (and last) a palaeographer, specialising in ancient handwriting. The whole staff (over 7000 at peak) ranged from the thousands operating teleprinters, the Hollerith installation, the key-breaking Enigma bombes and the several Colossus computers round the clock to a Dilly working alone with pencil and paper. Civilians on the Foreign Office payroll mixed in teams with men and women of the three services. Age and rank took second place to those with the gift. And the gift came from diverse sources. The joint head of the JN 11 team next door to us, Army Intelligence Captain Brian Augarde, was a professional jazz clarinettist. The story that Geoffrey Tandy was recruited through a misreading of his expertise in cryptogams (ferns, lichen and fungi) may be apocryphal but it illustrates a truth.

At work we were tightly compartmented for security reasons and never compared notes with other teams. Off-duty we mixed freely. With almost no contact with the people of Bletchley and the surrounding villages, and most of us far from our families, we were a very inward-looking society. All the civilian men (except for some of the most senior) had to serve in the Bletchley Park Company of the Home Guard: this was a great mixer and leveller. On the intellectual side, chess was probably the most glittering circle, with Hugh Alexander, Harry Golombek and Stuart Milner-Barry at its centre. These three had been together in the British team at the Chess Olympiad in Argentina when the war started in September 1939. There was music of high quality. Myra Hess visited to give a recital. Performances mounted from within the staff included "Dido and Aeneas", Brian Augarde's jazz quintet and several satirical revues. A group of us went often by train and bicycle to the Shakespeare Memorial Theatre at Stratford-upon-Avon. Scottish country dancing flourished. The Hall, which was built outside the perimeter security fence so that Bletchley people could use it too, provided for dances as well as the performances and a cinema. One memorable occasion was the showing of *Munchhausen*, in colour (probably the first colour film that most of us had seen) and in German without subtitles. I heard no explanation of how it came to be at Bletchley Park. I doubt that it was through the normal distribution channels.

Bayes and German Naval Enigma

The cryptanalysis of the German Naval Enigma was significantly more complex than that of JN 25. Greatly simplifying, the Enigma machine contained three (later four) 26-lettered wheels on a single axis, and a 26-letter plugboard; each wheel could be in any one of 26 different positions relative to its neighbours, resulting in an astronomical number of possible combinations – and each combination was used to encode just a single letter of the message. Recovering these wheel settings, which changed daily, was the heart of the deciphering process. Again I mention only three of the many people involved: Alan Turing, who needs no introduction; Hugh Alexander, who besides being British chess champion was both a master cryptanalyst and a master manager of cryptanalysts; and Jack Good, mathematician and later Professor of Statistics at Virginia Tech. Because of the compartmenting of all the various teams at Bletchley Park, I knew nothing of their work at the time.

Unlike JN 25, Enigma encipherment did not start by transforming message texts into code groups. Messages were enciphered by the machine, letter by letter, and with each letter's encipherment the machine's wheel settings automatically changed before enciphering the next. As with JN 25, an important early stage of the cryptanalysis was to establish a depth.

The analysis starts with two messages known from their indicators to have two of the three wheel settings in common. The immediate objective is to identify the third. The messages

can be slid one against the other, up to 25 places to the left or to the right; and at each of these 50 alignments there will be a stretch of letters one above the other. This is the "overlap". The messages already have two wheel settings in common: one of the 50 alignments must correspond to their having the third in common too. At that position all the pairs of letters in the overlap will have been enciphered exactly alike. The messages are then in "true depth". Turing noticed, possibly as early as late 1939, that he could establish which of the 50 alignments was likely to give the true depth by the statistical exploitation of two simple observations:

- if two letters, one above the other in the enciphered messages, were the same – that is, if they "matched" – and the alignment was correct, then (because the same machine setting had enciphered them) the corresponding letters in the original plain texts must also match;
- because the letter frequencies in language are distributed unevenly, matching will occur more frequently in true depth than in false.

In other words, a message can be distinguished from a random jumble because the letter frequencies are different. This difference will be hidden in the coded messages, but will re-appear, when coded messages are correctly aligned, in the frequencies of the matching pairs (see box).

Practical steps were taken to convert this statistical statement into a procedure which could be followed by non-mathematical staff on the scale and at the speed required. Counting the number of matching pairs involved recording the messages on heavy paper sheets printed with up to 250 vertical A–Z alphabets: this was done by punching a hole at each letter's position in successive columns. The two sheets were then slid one against the other above a lit background to reproduce successively all 50 alignments. At each position, matching letters showed up as visibly matching holes, to be counted and recorded for that alignment. As the sheets were printed in Banbury they were called Banburies, and the whole process thus begun was Banburismus. Further details of the breaking of Enigma encipherments are given in the website version of this article, at `www.interscience.wiley.com/journal/significance`.

It is piquant to observe that both the simple dexterous and visual procedure with Banburies and the high-speed mechanised one of the Hollerith machine exploited holes punched in cards. Both were derived from the punched-card device of the 1801 Jacquard loom, which attracted Ada Lovelace as she visualised the modern computer in 1843. Her vision was realised in Colossus. Bletchley Park used both the first computer and the device that had inspired it 100 years before.

Hut 1 at Bletchley Park. It was from buildings like these in the grounds of the manor house that codebreaking took place. Bletchley Park is now a museum and heritage site open daily to visitors. Details can be found at www.bletchleypark.org.uk. (Photograph by Toby Oxborrow.)

All the above is a simple, stripped-down description of the procedure, but it is enough to demonstrate the part played by Bayes' theorem. To put it in perspective, it was only the beginning of the daily breaking of the Naval Enigma wheel settings. With possible alignments for a pair of messages scored, it was still often a matter of judgement to pick the correct one: other techniques, cryptanalytic rather than statistical, brought experience to bear. Bayes provided only the platform from which the recovery of the daily wheel settings was launched.

When the war ended, Hugh Alexander and many others stayed with the organisation when it moved to Cheltenham as the Government Communications Headquarters. The dons returned to Cambridge and Oxford. Stuart Milner-Barry returned to the Treasury and was later the Government's Ceremonial Officer. Max Newman and Alan Turing went on developing computers at Manchester University, but Tommy Flowers returned to Dollis Hill with no recognition that he had done anything out of the ordinary.

Others went in less foreseeable directions. Edward Boyle and Roy Jenkins went into politics and government. Henry Reed, a Japanese linguist while in our JN 25 team, went back to poetry, radio plays and the BBC. His "The Naming of Parts", which has been called "the best-loved and most anthologised poem of the Second World War", voices his reflections while serving in the Bletchley Park Home Guard. Angus Wilson became an acclaimed novelist. Colin Thompson became Director of the National Galleries of Scotland, and Peter Benenson founded Amnesty International. Peter Laslett became a Cambridge social historian and initiated the Dawn University, out of which grew the Open University and the University of the Third Age.

In Jack Good's obituary *The Times* wrote: "To statisticians, Good is one of the founding fathers of Bayesian statistics. This approach was little used before Good's work but was given an important boost by his 1950 book *Probability and the Weighing of Evidence*[6]. Good's importance outside academia rests on his having been a key figure in the mathematical team at Bletchley Park."[8] So in relation to Bayes, as in the development of computers, Bletchley Park's contribution continued from wartime into peace.

References

1. Simpson, E. H. (1949) Measurement of Diversity. *Nature*, **163**, 688.
2. Simpson, E.H. (1951) The interpretation of interaction in contingency table. *Journal of the Royal Statistical Society* (Series B), **13**, 238–241.
3. Copeland, B. J. (2006) *Colossus: The Secrets of Bletchley Park's Codebreaking Computers*. Oxford: Oxford University Press.
4. Smith, M. and Erskine, R. (2001) *Action This Day*, p. 185. London: Bantam Press.
5. Alexander, C. H. O'D. (*c*.1945) Cryptographic history of work on the German Naval Enigma, PRO HW 25/1 and `www.alanturing.net/turning-archive`.
6. Good, I.J. (1950) Probability and the Weighing of Evidence. London: Charles Griffin.
7. Yule, G. U. (1944) *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press.
8. *The Times*, April 16th, 2009.