From **Brunelli, Lina & Cicchitelli, Giuseppe (editors). Proceedings of the First Scientific Meeting (of the IASE).** Università di Perugia (Italy), 1994. Pages 367-375. This pdf file is from the IASE website at http://www.stat.auckland.nz/~iase/publications/proc1993. Copies of the complete Proceedings are available for 10 Euros from the ISI (International Statistical Institute). See http://isi.cbs.nl/sale-iase.htm for details.

SECRET CODES IN STATISTICAL EDUCATION

Erdinç Çakiroğlu and Tibor Nemetz
Middle East Technical University
Department of Sci. Education, 06531 Ankara, Turkey

1. Introduction

The utility of solving simple substitutional cryptograms in statistical education at school level has been widely demonstrated (see Nemetz, 1993). In England, for example, national, project type competition was organised in 1989 in this area and it has been proved very successful. For the goals and history of these competitions, the reader is referred to Hawkins (1989). There are other similar simple examples within cryptography which can be fruitfully used for demonstrating the power of statistical methods in "real world" problems. This paper discusses one of them, namely the transpositional ciphers. Their solution is a feasible task from middle school years up to university level. At pre-university level, instead of using likelihood ratios, we propose a different statistics, which matches the age of the children and "near optimal". Connection to computer programming is straightforward, applicable both to easy and to more involved tasks. By working through a few examples, the need for applying statistical methods and the nature of the statistical decisions will become evident.

At university level, this problem offers an excellent opportunity to introduce "real life" hypotheses testing problems and, at the same time, highly theoretical divergence measures probability distributions. For teacher training, the lesson to be learnt is that theoretically based procedures can be transformed to the level of understanding of school children. The relations to graph theory can also be pointed out. Such an abstract notion like the existence of the Hamiltonian circuit in a graph obtains clear meaning.

The primary goal of this presentation is to contribute to the collection of easily solvable project topics for teacher training, which could be accepted by the teachers to be included in the school practice.

2. Transpositional secret codes: an explanation

This encoding is applied to written texts. The message which is to be sent

over an unsafe channel is subjected to a coding procedure with the goal of hiding the secret information contained in it. To this end, the transpositional codes divide the message into blocks of fixed length, which we denote for simpler reference by the letter B. Then they transpose the positions within the given blocks by applying the same permutation for the positions in all blocks. The last block of the message is usually not complete. There are different rules to fill it in, e.g. randomly, which are not essential for the present purposes. The transformed blocks are sent over the channel, and the legal receiver simply applies the inverse permutation (known to him) to decode the original message. Usually, however, unwanted eavesdropping is possible, and a third, undesirable party can get hold of the cryptic text. In this paper we escort him on the way of breaking the secret code. For interested readers we recommend Kahn's "Bible of cryptographers" (1967), where the history of such crypto-codes is easy to follow. For your reference, we provide you with an example in the Appendix.

3. Breaking the code: a hypotheses testing problem

Before turning to educational aspects, let us formulate the codebreaker's task in the language of statistics. Say, he obtained N complete blocks, this is his sample. He has exactly $B!$ simple hypotheses, each of them is a possible candidate for the applied permutation. He knows for sure, that one of them was really applied. He also knows the habit of the correspondents, that the actual permutation was randomly chosen, i.e. with equal probabilities, from the set of all possible cases. Therefore, we have the Bayesian Hypothesis-Testing Problem with a large number of simple hypotheses.

At the beginnings, he does not know the underlying probability distributions, but we may suppose that he knows the source, at least in the sense, that he can, in principle, inspect a sample of arbitrary length to obtain a statistical estimation of the underlying probabilities with any desired precision at any given level. But now he runs into the conflict of theory and practice: memory and time constrains do not allow him to carry out this task.

Again, theory gives the key to proceed further. It is well known that the Bayesian error probability for a multiple number of simple hypotheses can be bounded by a linear function of the maximum error probability for the pairs of the constituent hypotheses, moreover the error-exponents in these cases are the same. Therefore he can decompose the original testing problem into a number of not-so-complex problems, without loosing too much, in the sense of optimal error probability. A possible decomposition can be obtained in the following

way. To describe it, we introduce some formal notations:

- The original message is composed of blocks $M(i)$, $i = 1,2,...,N$, and the *jth* character of the *ith* block is denoted by $M(i,j)$, $j = 1,2,...,B$.

- The encoder uses the inverse permutation of P (we use this notation just to simplify formal expressions).

- The coded blocks are denoted by $C(i)$, $i = 1,2,...,N$, and *jth* character of the *ith* block is $C(i,j)$.
  Now we formulate B decision problems:

Decision problem $D(K)$, $K=1,2,...,B$:
Which is the true Hypothesis:
The *Kth* letters of the coded blocks are followed by their *Jth* letters in the corresponding message blocks, $J = 1,2,...,B$, $J<>K$ (i.e. the letters $B(I,P(J))$ are subsequent to $B(I,P(K))$ in $M(I)$),
or
The *Kth* letters of the coded blocks are the last letters in the message blocks.

It is easy to see, that a correct decision on P is equivalent to correct decisions for all $D(K)$. Also, for large sample size (i.e. for large N) the last Hypothesis in $D(K)$ is equivalent with large probability to the claim, that there is no position J, such that all the letters $B(I,P(J))$ are subsequent to $B(I,P(K))$ in $M(I)$.

Furthermore, since P has been chosen uniformly from all possible permutations, the hypotheses in all problems $D(K)$ have the same probability, none of them plays any special role. The maximum likelihood principle tells then that we have to accept the hypothesis J, for which the probability

(1) Pr { $M(I,P(K))$ is followed by $M(I,P(J))$, $I = 1,2,...,B$, $J<>K$ } is the maximum, unless the probability

(2) Pr { $M(I,P(K))$ is the last letter in all $M(I)$ } is larger, in which case this is the decision.

We have to turn to experiments to find out what the underlying probability distributions are, and to establish that they are different from the one-step transition distributions are, and this is not a difficult job. Under this condition the Bayesian error probabilities go to zero exponentially as N increases, and the error exponent is the same for all decision problems $D(K)$ (and is the same as

for the original Hypothesis testing problem on P); see e.g. Nemetz (1972).

Now we can propose the following decision procedure.

Decision procedure

Step 1. Prepare a conditional frequency distribution $F(Y|X)$ from consecutive characters of a "long" source sequence, which gives an "acceptable" estimation of the one step probability transition distribution, and use this approximation instead of the unknown real one.

Step 2. For any given K, $K = 1,2,...,B$ compute for all $J<>K$, $J = 1,2,...,B$ the negative of the log-likelihood functions

$$(3) - [\log F(C(1,J) / C(1,K)) + \log F(C(2,J) / C(2,K)) + ... + \log F(C(N,J) / C(N,K))]/N,$$

and take the value J that minimizes it. Note, that with large probability this is uniquely defined if N is large enough. If not, solve the breaks arbitrarily. Technically, one works with the negative of the likelihood function to get positive values, therefore we use the minimum.

Step 3. Define a directed graph with vertex-set as positions in the blocks, and connect position K to position J with a directed edge if J is the value which minimizes the likelihood functions for the given K.

Step 4. Find a maximum length Hamilton line in the graph above, and use it to define a permutation P which reconstructs the original message box. If such a line is not exhaustive, proceed arbitrarily.

We formulate a number of claims concerning this decision procedure in somewhat loose terms. Most of them is known or easy to prove.

Claim 1. Due to the law of large numbers, the likelihood functions in (3) converge against their expected values. If J yields the correct hypothesis, then these are the differences of the conditional Kullback-Leibler Informational Divergencies $D(F(2,1)||P(2,1))$ and the first order entropy $H(1)$ of the source. Here $P(2,1)$ denotes the "true" probability distribution of the consecutive characters of the (stationary) source sequences. For $D(.||.)$ and its educational oriented introduction we refer to Nemetz (1972). Under mild conditions this difference goes to zero if the size of the sample used to derive the approximation $F(.,.)$ increases.

Remark: The expected value is infinite, if $F(Y|X) = 0$ whenever $P(Y|X)$ is not. The usual trick to overcome this difficulty is to modify the frequency distributions a little bit by adding 1 to all possible frequency entries before forming the relative frequencies. In what follows we always use $F(.,.)$ with this meaning.

Claim 2: The conditional distributions of the *Kth* successors of a source symbol are monotonically "nearer" to the (stationary) one dimensional distributions. This implies, that if J is not the proper hypothesis, the expected value is larger than in the case of the true J. Therefore the Bayesian error probability of the proposed decision rule goes to zero when N goes to infinity.

Remark: If K is large, practically there is no observable dependence. For our present purposes $K > 5$ is large enough. Our empirical studies show that in case of Hungarian and Turkish written text sources, for blocklength $B = 20$ and message length of $N = 50$ blocks, the error probability is under 1 per cent.

Remark: In case of written language sources, an acceptable approximation $F(.,.)$ can be obtained by analysing a text of around 50.000 characters. Such a job was a slaves' work two decades ago, but now is within the range of middle schools. Historically, language statistics have a long standing, they were used in the last century (see Kaeding, 1898).

4. Tasks for teacher training college students

The appropriate place in the curriculum for solving transpositional codes is the time of discussing hypotheses testing problems. The students may observe the necessity of collecting data, they can feel that the problem is not artificial, and sense the power and the nature of statistical decisions. Also, they can practise the way of applying theoretical results. Performing the tasks in this section may be instructive to social science, business and electrical engineering students, as well.

Task 4.1. Construct a file of a source sequence of about 50.000 characters, and another smaller one playing the usual role of the control group. The first one is to be used to construct a frequency table, the second one is for checking the goodness of the decision rule.

In constructing the source files, work-sharing should be applied. In a class of 25-30 students this means of inputing 2.500 characters per students, which

is a realistic demand. It is to be noted, that the output data files can be used for many statistical illustrations, as well.

Task 4.2. Prepare frequency tables of the k-step transitions of the source data file, and extend them to contingency tables. (Here k-step transition relates to bigrams of the source sequence where the first and the second characters are separated by exactly $k-1$ symbols inbetween. Care should be taken to explain this). Note, that a single program will do if k is treated as a parameter.

Task 4.3. Prepare the log-likelihood tables. For technical reasons, outlined in the previous section, add +1 to all entries in the frequency tables, and divide the entries by their row-sums to get the "forward" k-step transition relative frequencies, by column-sums to get backwards transition relative frequencies, and by the total-sum to get the absolute relative frequency tables. Note that all relative frequencies appearing here are strictly positive. Thus it is meaningful to take logarithms of the entries. The resulting tables all called log-likelihood tables.

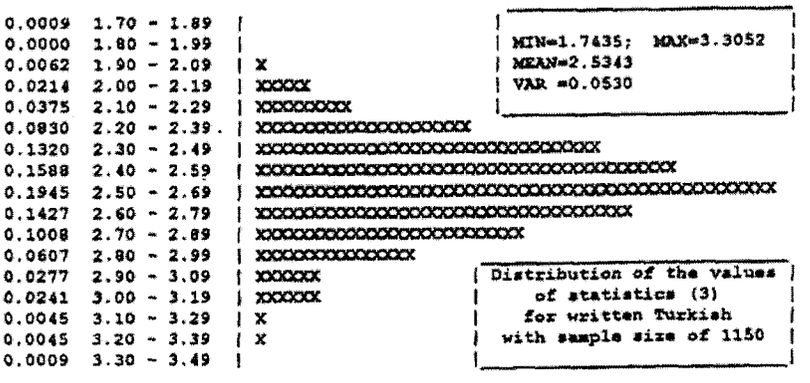

Figure 1. *Distribution of statistics (3) under the null hypothesis*

Task 4.4. Prove that the goodness-characteristics of the decision procedure do not depend on the choice of the permutation P. Conclude that therefore any performance analysis can be carried out using the identity permutation, only.

Task 4.5. Take a sample $(U(i), V(i))$ of pairs of characters from the "control"

text, which are separated by exactly $k-1$ letters, for $k = 1, ..., 20$. Determine the empirical distribution function of the statistic $-logF(V,U)$. Find a graphical representation that can convincingly show, why the statistics (3) works. Fig. 1 shows what kind of graphs are to be expected.

Task 4.6. Take $B = 20$, $N = 50$, P = identity, and use the control text to analyse the performance of the decision rule empirically.

At that point we stop listing further tasks, but, obviously, every experienced instructor can supplement them in a number of ways. Now we formulate two research problems.

Open problem 1. Construct an appropriate source distribution model (e.g. Markov chain) and determine the probability, that the decision rules above define a complete Hamilton line.

Open problem 2. Generalize the decision rule to include the first-step backwards transition probabilities, which could define "backwards edges" in the position graph, and investigate if it improves the performance characteristics.

5. Tasks at school level

Unlike at university level, here the theoretical aspects of hypothesis testing are not dealt with. Here an important issue is to try to make children understand the notion of dependence, to convince them that statistics could be and should be applied in real world problems and to develop such abilities through discovery and activity approach. Our present example offers a possibility to accomplish this job. We have to transform the tasks posed to college students, since to explain why we use the logarithmic-sum function (3) would take generally much more time than available in schools. The raw material is obviously needed, thus we have to include tasks 4.1 and 4.2. Using the output of Task 4.2, we can construct succession-tables, showing which block-letters are positioned after a given block-letter in the clear-text.

The Succession Index $SI(Y|X)$ is defined as the index-number of the letter Y in the conditional frequency order of Y among the letters following X in the source sequence. E.g. $S(Y|X) = 1$ iff Y is the most frequent letter to follow up the letter X, it is 2, if Y is the second most frequent letter, and so on. It is the teacher's job to explain and demonstrate the utility of the index numbers. Refer to Nemetz (1991) for an explanation of this. The matrix S with entries $SI(Y|X)$ is called the succession

matrix. The statistics (3) used for making the decision should be replaced accordingly by

(4) $[S(C(1,J) / C(1,K)) + S(C(2,J) / C(2,K)) + \ldots + S(C(N,J) / C(N,K))] / N.$

The decision algorithm remains the same with this exception.

Now we formally list the tasks in accordance with section 4.

Task 5.1 is identical to Task 4.1.

Task 5.2 is the same as Task 4.2., with the possible restriction to the 1-step transitions.

Task 5.3. Prepare the succession matrix S of Succession Indices based on the output frequency table(s) of the Task 5.3.

Task 5.4. Give arguments/run a discussion in the classroom with the aim to show that the goodness-characteristics of the decision procedure do not depend on the choice of the permutation P. Conclude that therefore any performance analysis can be carried out using the identity permutation, only.

Task 5.5. Take a sample $(U(i), V(i))$ of pairs of characters from the "control" text, which are separated by exactly $k-1$ letters, for $k = 1, \ldots, 5$. Determine the empirical distribution of the statistic $SI(V|U)$. (In this case it ranges from 1 to the alphabetsize through integers). Find a graphical representation that can convincingly show, why the statistics (4) works.

Task 5.6. Same as Task 4.6., with the difference in the statistics used.

Bibliography

Hawkins A. (1989), The annual United Kingdom Statistical Prize, in M. Morris (ed.), *The teaching of statistics, Studies in mathematics education,* Vol. 7, Unesco.

Kaeding F. W. (1898), *Haufigkeitswoerterbuch der deutschen Sprache,* Akademic-Verlag, Berlin.

Kahn D. (1967), *The codebreakers,* MacMillan Co., New York.

Nemetz T. (1972), *Information type measures and finite decision problems,* Lecture Notes, No. 2/1972, Carleton University, Ottawa.

Nemetz T. (1991), Automatic Discrimination of Written Languages from Random Strings with an Application to Error Detection, in M. Niss (ed.), *Teaching Mathematical Modelling and Applications,* Ellis Horwood, New York, pp. 230-241.

Nemetz T. (1993), Cryptology: a Rich Source of Applications Offering Entertaining Mathematical Instruction, in J. de Lange *et al.* (eds.), *Innovation in Maths Education by Modelling and Applications,* Ellis Horwood, New York, pp. 93-102.

Appendix

An illustration to tranpositional secret codes. This example provided our students with valuable experience in a statistics course for student teachers at METU. Should you decide to use it in your own course, we would be glad to share your experience. The following text was distributed:

```
We have used capital letters TO WRITE AN ENGLISH TEXT. THEN ALL SPACES
HAVE BEEN REPLACED BY THE LETTER X,
ANDXTHEXRE  SULTINGXTE  XTXHASXBEE  NXDEVIDEDX  INTOXBLOCK  SXOFXLENGT
HXTENXWHIL  EXALLXPUNC  TUATIONXMA  RKSXHAVEXB  EENXOMITTE  DXTHENXWEX
HAVEXAPPLI  EDXTHEXSAM  EXPERMUTAT  IONXTOXALL  XBLOCKSXBE  LOWXWEXGIV
EXTHEXRESU  LTXYOUXARE  XINVITEDXT  OXRECONSTR  UCTXTHEXOR  IGINALXTEX
TXTHEXFIRS  TXCORRECTX  SOLUTIONXI  SXWORTHXOF  XFIVEXPOIN  TSXTIESXWI
LLXBEXDISS  OLVEDXBYXR  ANDOMXCHOI  CEXINDEPEN  DENTXESSAY  SXLEADINGX
TOXGENERAL XSOLUTIONX METHODSXWI LLXBEXHONO UREDXBYXTE   NXPOINTS
**************** an encoded message follows ****************
OJTCHTXEBE  XOTXVISRFE  AISUEHVORX  ERXHRXSCAE  VTSECAIIIT  XENHAXDCTN
XEIVCILRSA  CNCUECXDOS  BXEHETXTYD  VRTIUXISEN  NBSXXYAEXC  RZXDMUAEIM
OLSWSAFOLX  CNIRTXXTOO  XOHTUBEXTT  VNMEXEDCAA  OXICNEXSFT  XNDXNEEDAC
OMTNVELEPE  TCONOXXHEF  XOOCOLYXTG  BTTXTNIEUR  TOLAXOANIN  EOEMDXVPLE
YMIBTNBOXX  NXDNILIAGZ  ETEVEXFICF  SNTXYLUGIX  EOCREHRUSX  FTXESEOHXX
ESYTNUVIRI  MNUPTXXIAO  XPILALEPAT  EECRDERASX  XXAEXHSMAA  FSMISNOTXX
IGNALUTXNA  SINIXDNRPI  SCERXGAXIB  CXXOESROHA  AEHTRPPXRE  CSRAXEESEN
CGUOXYARKB  OXTNDNFIRX  SIILREIPCD  XPORANYPAR  SAXDCAENXH  EAOIOCPTRO
XEAEXNNSRI  ADOTCRXXNH  DTXELXAHXE  LXXFERUOTS  ACTXEREHRS  ESGAXOHTXT
XPILXEFPAO OXXXACIXNT
```