REMARKS ON THE PROBLEM OF
"HOMOGENIZATION OF BERNULLI TRIALS"

BY

STANLEY L. SCLOVE

TECHNICAL REPORT NO. 137
JUNE 15, 1968

THIS RESEARCH WAS SPONSORED BY THE ARMY RESEARCH OFFICE,
OFFICE OF NAVAL RESEARCH, AND AIR FORCE OFFICE OF
SCIENTIFIC RESEARCH BY CONTRACT NO.
Monr-225(52) (NR-342-022)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
REMARKS ON THE PROBLEM OF
"HOMOGENIZATION OF BERNULLI TRIALS"

by

Stanley L. Sclove

TECHNICAL REPORT NO. 137
June 15, 1968

PREPARED UNDER CONTRACT Nonr-225(52)
(NR-342-022) FOR
OFFICE OF NAVAL RESEARCH

Reproduction in Whole or in Part is Permitted for any Purpose of the United States Government

Department of Statistics
Stanford University
Stanford, California
1. The Problem. This technical report contains a statement of some preliminary results and some suggestions of possible methods for further analysis of the following problem.

Suppose that we have two coins with different values for the probability of heads. Let $p_1$, where $0 < p_1 < 1$, and $p_2$, where $0 < p_2 < 1$, be the respective probabilities of obtaining heads on a single toss of the coin. We assume that $p_1 \neq p_2$. Note that nothing is assumed about one or none of the coins being unbiased.

Now choose one of the two coins at random (with equal probability). Toss this coin an unknown number of times. Then toss the other coin an unknown number of times. Then toss the first coin an unknown number of times, and so on, alternating the coins for a very large total number of tosses. The numbers of times that each coin is tossed need not be the same, and in general will be different. All tosses are assumed to be independent in the probability sense.

We thus have available a very long sequence of observed heads and tails. We have no knowledge as to which coin has been used for a given subsequence of observed heads and tails. How do we divide the complete sequence of very many tosses into subsequences where we are "reasonably sure" that the same coin has been used for all the tosses in the subsequence? It is not crucial to detect exactly when a change of coin occurs. An appropriate loss function for the problem might be simply the total number of misclassifications. One might also wish to consider a more general loss function under which misclassifications of one type

*Prepared under Contract Nonr 225-52 (NR 342-022)
are more serious than misclassifications of the other type:

\[ \text{loss} = cn(2|1) + bn(1|2), \]

where \( n(1|j) \) is the number of errors made by saying Coin i was tossed when in fact Coin j was tossed.

This problem has been dubbed the problem of "homogenization" of Bernoulli trials.

We assume that a decision must be made regarding each and every toss; the decision rule must say "Coin 1" or "Coin 2". In some situations it may be reasonable to allow decision rules which say "Coin 1", "Coin 2", or "Don't Know".

There are two cases: \( P_1, P_2 \) known; \( P_1, P_2 \) unknown. From some preliminary analysis I have done it seems that the latter will be much more difficult than the former, which itself seems difficult.

2. Possible Applications. Let me mention some situations which come to mind.

1. **Quality Control.** Suppose we are willing to characterize a production process simply as either "in control" (Coin 1 being tossed) or "out of control" (Coin 2 being tossed). Then \( P_1 = \Pr(\text{defective item} | \text{process in control}) \) and \( P_2 = \Pr(\text{defective item} | \text{process out of control}) \).

   We wish to ascertain when (on which trials) the process was in control and when it was out of control. Then we might examine other variables measured concurrently in an attempt to find which variables are important for keeping the process in control.

2. **Epidemiology.** Let the tossing of Coin 2 correspond to the presence of an epidemic and the tossing of Coin 1 correspond to no epidemic. Take
\[ p_1 = \Pr\{\text{individual contracts disease, given no epidemic}\} \]
and
\[ p_2 = \Pr\{\text{individual contracts disease, given epidemic}\}. \]

3. **An Intuitive Approach.** Intuitively, it seems that the sum of squared deviations of a sequence of tosses should be smaller when one coin is tossed throughout the sequence than when both are tossed during the sequence. Hence, a moving sum of squared deviations might be used to indicate when shifts in coin occur.

Consider a sequence of \( m \) tosses. Let

\[ X_i = \begin{cases} 
0 & \text{if the } i\text{-th toss is a Tail} \\
1 & \text{if the } i\text{-th toss is a Head} 
\end{cases} \]

Suppose \( n \) tosses have been made. Consider the sequence of the last \( m \) tosses \( X_n, X_{n-1}, \ldots, X_{n-m+1} \). Their sum of squared deviations is

\[ S_{n,m} = \sum_{i=1}^{m} X_{n-i}^2 - \frac{1}{m} \left( \sum_{i=1}^{m} X_{n-i} \right)^2. \]

Now, \( X_i^2 = X_i \), so

\[ S_{n,m} = \sum_{i=1}^{m} X_{n-i} - \frac{1}{m} \left( \sum_{i=1}^{m} X_{n-i} \right)^2 = T_{n,m} - \frac{1}{m} T_{n,m}^2, \]

where \( T_{n,m} = \sum_{i=1}^{m} X_{n-i} \). Thus, consideration of the sum of squared deviations reduces to consideration of the sum. Note that large values of \( S_{n,m} \) correspond to intermediate values of \( T_{n,m} \):

3.
$S_{n,m} \geq s$ if and only if $l_m < T_{n,m} < u_m$,

where $l_m = m[1-(1-4s/m)^{1/2}] / 2$ and $u_m = m[1+(1-4s/m)^{1/2}] / 2$. If a single coin is tossed for all $m$ tosses, then $S_{n,m}$ is small, i.e., $T_{n,m}$ is either small or large (depending upon which coin was being tossed).

If $m$ is chosen properly (relative to the frequency of coin switching), then a plot of $T_{n,m}$ vs. $n$ would look like a "square wave" (or rather a 'trapezoidal' wave, since the sides will tend to have a non-vertical slope; see graphs in Appendix).

If $p_1$ and $p_2$ are known, one might use the decision rule

\[
\begin{cases}
\text{Say } X_n \text{ came from Coin 2 if } T_{n,m} > m(p_1 + p_2) / 2 \\
\text{Otherwise, say } X_n \text{ came from Coin 1}
\end{cases}
\]

(See Figure L.). More generally, if the losses due to the two types of errors are not equal, e.g., if the loss is $c$ when we say $X_n$ arose from Coin 2 when in fact it came from Coin 1 and the loss is $b$ when we say $X_n$ arose from Coin 1 when in fact it came from Coin 2, then we might use the decision rule

\[
\begin{cases}
\text{Say } X_n \text{ came from Coin 2 if } T_{n,m} > \frac{c}{c+b} m(p_1 p_2) \\
\text{Otherwise, say } X_n \text{ came from Coin 1}
\end{cases}
\]

4. Moving Averages. We have discussed an intuitive approach which leads to consideration of the moving sum (or average) of observations, $T_{n,m}$. More generally, we could consider weighted moving averages.
\[ U_{n,m} = \sum_{i=1}^{m} c_{i,m} X_{n-i} . \]

The problem is how to design the moving average and how to evaluate the performance of the decision rule based on the moving average. By "designing" the moving average, I mean choosing the number \( m \) of observations to be included and choosing the coefficients \( c_{i,m} \).

(If we write the moving average as \( \sum_{i=0}^{\infty} d_{i} X_{n-i} \), then choosing \( m \) is equivalent to setting \( d_{i} = 0 \) for \( i \geq m \).) Use of a large \( m \) eliminates noise and use of a small \( m \) makes the moving average responsive; the optimal \( m \) must balance these desirable attributes.

Chernoff and Zacks [1] used a weighted moving sum which in terms of our problem would be

\[ U_{n,m} = \sum_{i=0}^{m-1} (m-i-1) X_{n-i} . \]

\( (X_{n-m+1} \) has zero weight, so \( U_{n,m} \) contains only \( m-1 \) observations.) This sum, which weights recent observations more heavily than old ones, may be more responsive to changes in coin than the ordinary unweighted average. For testing for a single change in mean [1], such a statistic was shown to be Bayes against a uniform prior on the time of the change in mean. A moving sum with exponential weights might prove easy to analyze; e.g., we might consider

\[ U_{n,m} = \sum_{i=0}^{m-1} a^i X_{n-i} \] where \( 0 < a < 1 \). The quantity \( a \) may be taken to be a function of \( p_1 \) and \( p_2 \) when they are known.
The papers of Chernoff and Zacks [1] and Kander and Zacks [2] deal with the problems of testing for a change in mean in a sequence of random variables $X_1, \ldots, X_m$. Say $E[X_i] = \mu_i$. The problem was to test

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_m = \mu \text{ (known)}$$

against

$$H_1: \mu_1 = \mu_2 = \cdots = \mu_r = \mu$$

$$\mu_{r+1} = \cdots = \mu_m = \mu + \delta \text{ (} \delta > 0 \text{ unknown)}.$$

In Chernoff and Zacks' work the $X_i$ are normal. Kander and Zacks considered other exponential families as well; they showed that the statistic

$$\frac{m-1}{\sum_{i=1}^{m-1} i X_{i+1}}$$

is Bayes against a uniform prior for the point of shift, $r$, when the $X_i$ are Bernoulli. The analog of this statistic in the present problem is (3). Hence use of this statistic in the problem at hand can be justified by postulating a coin-switching mechanism satisfying the following:

**Assumption 1.** Each coin is always tossed at least $m-1$ times.

**Assumption 2.** The distribution of the point of change in coin is uniform over any sequence of $m$ tosses.

Assumption 1 implies that in any sequence of $m$ tosses at most one change in coin occurs.
5. **Computer Simulation.** The coin tossing process was simulated using the Burroughs B5500 and the moving averages $T_{n,m}$ and $U_{n,m}$ were plotted vs. $n$ using the Calcomp Plotter. Each coin was tossed for 100 tosses. (If further study of this problem is deemed necessary, more simulation could be done using a random mechanism to generate the number of tosses. On the other hand, mathematical analysis may obviate the necessity for further simulation. See Section 7.) Various values of $p_1$, $p_2$ and $n$ were used. Some of the results (for $p_1 = .1$, $p_2 = .8$) are given graphically in the Appendix. The results indicate that a good value of $m$ is important and that $U_{n,m}$ seems more responsive than $T_{n,m}$; both of these findings were expected.

6. **A procedure for the case of known $p_1$, $p_2$.** In this problem, the computer and graphical analysis may play a role in the actual data analysis. I can envisage implementation of the following two-step procedure: first, plots of moving average vs. $n$ are made via computer to determine a good value of $m$ (one which gives a plot which looks most nearly like a square wave); secondly, the decision rule (1) mentioned above is used to assign each toss to Coin 1 or Coin 2. This can be used as a working rule until further analysis provides other, more precise, rules.

7. **Suggested Methods for further analysis.** In thinking about this problem it is helpful to conceive of two processes; one a coin-switching process by which Nature chooses at each trial which coin is to be tossed on the next trial and the other the actual coin-
tossing process. Introduction of a suitable model for the coin-switching process may provide a structure for the problem and make the analysis more fruitful. It would be useful to model the coin-switching process as a Markov chain with transition probabilities,

\[ t_{ij} = \Pr(\text{Coin } j \text{ is tossed on next trial}|\text{Coin } i \text{ is tossed on current trial}), \quad i,j = 1,2 \]

(These transition probabilities are assumed to be independent of the trial number \( n \), i.e., are assumed stationary: \( t_{ij}(n) = t_{ij} \) for all \( n \).) The symmetric case \( t_{12} = t_{21} \) may be of special interest. Under the Markov chain model for coin-switching, the random variables \( N_i \) = number of times Coin \( i \) is tossed before a switch to the other coin occurs, \( i = 1,2 \), have a geometric distribution with parameter \( p = 1 - t_{ii} \), e.g.,

\[ \Pr(N_i = k) = (1-t_{ii})t_{ii}^k = t_{12}t_{11}^k, \quad k = 0, 1, 2, \ldots \]

In the beginning, Nature is supposed to choose a coin at random for the first toss; this part of the procedure is modeled by specifying the initial distribution vector \( (t_1, t_2) \) of the Markov chain \( (0 < t_1, t_2 < 1, t_1 + t_2 = 1) \). In the statement of the problem we specify \( t_1 = t_2 = 1/2 \); we might wish to allow general \( (t_1, t_2) \).

We are proposing decision rules of the form

\[ \text{Say Coin 2 was tossed on trial } n \text{ if } U_{n,m} > k, \]

\[ 8. \]
where $U_{n,m}$ is of the form (2). The optimal constant $k$ will presumably be a function of the losses ($c$ and $b$) and the $p_i$. The dependence of $k$ upon the $p_i$ is what will make the problem so much more difficult when the $p_i$ are unknown.

Analysis may proceed by first solving some simpler problems. We simplify by removing some of the randomness. It may be helpful to consider the case in which the coin-switching process is degenerate; i.e., $N_i = n_i$ with probability one, $i = 1, 2$, where the $n_i$ are unknown (but large). For the symmetric case, take $n_1 = n_2$. Thus we assume that each coin is always tossed the same amount of times, but we do not know how many times. After determining the optimal $m$, $c_{i,m}$ and $k$ as functions of $c, b, p_1$ and $p_2$ for the simplified problem we can reintroduce the Markov chain model for the coin-switching process.

A different simplification is to take the randomness out of the coin-tossing process; that is, assume that

$$X_n = p_i \text{ with probability one if Coin } i \text{ is tossed on trial } n.$$  

This simplification of the problem is artificial, because we would make our decision on the basis of the $X_n$ themselves, not a moving average; that is, we would take $m = 1$. Perhaps we could add to assumption (6) some other assumption that would create a problem that would give insight on the original problem; e.g., suppose we assume that we have to use a moving average; say we require $m \geq m_0 > 2$. Since the only randomness is in the coin-switching,
the problem seems simple. For, if Coin i has been tossed for at least
the last m consecutive trials, then the value of the moving sum
would be exactly \( mp_i \), and as soon as its value changed we would
know that a change in coin had occurred, and where. The problem is
that two or more changes in coin would have occurred during the last
m tosses. The number m must be small enough so that there is only
a small probability of more than one change in m tosses.

Remark. The homogenization problem has the aspects of an empirical
Bayes decision problem. Specifically, we are dealing with a process
which generates pairs \((X_n, \theta_n)\), where \(X_n\) is observable, \(\theta_n\) is not
observable and \(\theta_n\) is the parameter of the distribution of \(X_n\):

\[
Pr(X_n = x|\theta_n = \theta) = P_\theta(x)
\]

where \(\{P_\theta(\cdot), \theta \in \Theta\}\) is a set of probability functions.
In the present case, \(\theta_n\) can have only two possible values, \(P_1\) or
\(P_2\). What is more, there is a lot of structure built into the \(\theta_n\-
sequence, since we are assuming that each coin is tossed for a long
time. If we use (4) to model the coin-switching process, then we
have an empirical Bayes situation in which the parameter sequence is
a Markov process: \(Pr(\theta_n = p_j|\theta_{n-1} = p_i) = t_{ij}\).

Acknowledgments. Thanks are due Herman Chernoff for discussing this
problem with me and Susan Boyle for computer programming.
8. References


[2] Kander, Z., and Zacks, S. (1966). "Test Procedures for Possible Changes in Parameters of Statistical Distributions Occurring at Unknown Time Points." Ann. Math. Statist., 37, 1196-1210. The statistic suggested in (1) yields a generalization to the case of the one-parameter exponential family. The problem is to test the hypothesis that all \( n \) observations have the same known mean against the composite alternative that at some unknown trial \( m \) the mean increases by an unknown amount. Operating characteristics of the test statistic are studied. The binomial case (\(-1, +1\) random variables) is one of the examples. (The transformation from \( 0, 1 \) random variables to \(-1, +1\) random variables is \( u = 2t - 1 \); this can be used for relating the results to this memorandum.)

[3] Page, E.S. (1955). "A Test for a Change in a Parameter Occurring at an Unknown Point." Biometrika, 42, 523-526. Use of the cumulative sum in a similar context is suggested. In (1) it is shown that the power function values of the statistic proposed there are higher, for most points of shift and sizes of shift, than those of Page's test statistic. In (2) it is stated that a comparative study in which the effectiveness of test procedures based on the Chernoff-Kander-Zacks test statistics relative to those based on Page's and other procedures will be given 'elsewhere' for the exponential case, and other distributions of practical interest.


[5] Silvey, S.D. (1958). "The Lindisfarne Scribes' Problem." Jour. Royal Stat. Soc., 20, 93-10. This paper treats the problem of estimating, from observations on an ordered set of random variables, the number of changes which occur in the distribution functions of these random variables, when the probability of over-estimating the number of changes is to be controlled.
APPENDIX

Some results of the computer simulation.

Plots of $T_{n,m}$ and $U_{n,m}$ vs. $n$ for $m = 5, 10, 25, 50$ are given. The units on the n-axis (abscissa) are tens of trials.

(The plots rise at the beginning until $n > m$; before $n > m$ there are not enough trials to fill out the moving average.) Each coin was given 100 tosses, 4 times (800 tosses in all). $m = 10$ appears to give the best results. $m = 5$ clearly separates Coins 1 and 2, but the peaks and valleys due only to noise are too pronounced.

With $m = 25$ and 50 too many observations from both coins are present in the moving averages to make clear discrimination possible.
The diagram shows two lines plotted against the variable $N$. The upper line starts at a value of 0 and increases to a peak just below 50. The lower line starts at a value of 0 and increases to a peak around 30. The variable $m$ is labeled as 50 at the top of the diagram.
**DOCUMENT CONTROL DATA - R&D**

1. **ORIGINATING ACTIVITY (Corporate author)**
   Department of Statistics
   Stanford University
   Stanford, California

2a. **REPORT SECURITY CLASSIFICATION**
   
   2b. **GROUP**
   

3. **REPORT TITLE**
   REMARKS ON THE PROBLEM OF "HOMOGENIZATION OF BERNOULLI TRIALS"

4. **DESCRIPTIVE NOTES (Type of report and inclusive dates)**
   Technical Report

5. **AUTHOR(S) (Last name, first name, initial)**
   Stanley L., Sclove

6. **REPORT DATE**
   June 15, 1968

7a. **TOTAL NO. OF PAGES**
   12

7b. **NO. OF REFs**
   5

8a. **CONTRACT OR GRANT NO.**
   Nom-225(52)

8b. **PROJECT NO.**
   NR-542-022

9a. **ORIGINATOR'S REPORT NUMBER(S)**
   Technical Report No. 137

9b. **OTHER REPORT NO(S) (Any other numbers that may be assigned this report)**

10. **AVAILABILITY/LIMITATION NOTICES**
    Distribution of this document is unlimited.

11. **SUPPLEMENTARY NOTES**

12. **SPONSORING MILITARY ACTIVITY**
   Logistics and Mathematical Sciences Branch
   Office of Naval Research
   Washington, D.C.

13. **ABSTRACT**

This report presents some preliminary results and suggestions of possible methods for further analysis of the following problem. We have two coins with different values for the probability of Heads. One of the two coins is chosen at random and tossed an unknown number of times. Then the other coin is tossed an unknown number of times. Then the first coin is tossed again, and so on, the coins being alternated for a very large total number of tosses. The numbers of times that each coin is tossed will in general be different. All tosses are assumed to be statistically independent. We thus have available a very long sequence of observed Heads and Tails. The problem is to divide the complete sequence of very many tosses into subsequences where we are "reasonably sure" that the same coin has been used for all the tosses in the subsequence. It is not crucial to detect exactly when a change of coin occurs, and an appropriate loss function for the problem might be simply the total number of misclassifications.

Possible fields of application are quality control and epidemiology, among others.

It is suggested that a moving average be used for dividing the sequence of tosses into homogeneous subsequences. Some suggestions for further analysis are made. It is noted that the problem has the aspects of an empirical Bayes problem.
Quality control
Epidemiology
Moving average
Empirical Bayes theory

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization (corporate author) issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Markings to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year, or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b. & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers (either by the originator or by the sponsor), also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through"

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through"

(5) "All distribution of this report is controlled. Qualified DDC users shall request through"

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring (paying for) the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as (TS), (S), (C), or (U).

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.