THE VARIABLE WORD AND RECORD LENGTH PROBLEM AND THE COMBINED RECORD APPROACH ON ELECTRONIC DATA PROCESSING SYSTEMS

Neal J. Dean

The Ramo-Wooldridge Corporation

Presented at the Western Joint Computer Conference
Los Angeles, California, February 28, 1957

THE VARIABLE WORD AND RECORD LENGTH PROBLEM AND THE COMBINED RECORD APPROACH ON ELECTRONIC DATA PROCESSING SYSTEMS

Neal J. Dean
The Ramo-Wooldridge Corporation
Los Angeles, California

In the literature concerning electronic data processing systems there has been much discussion of the advantages of the variable word length feature as opposed to the fixed word length restriction. It might be well at the outset to specify what is meant by a "word." A word is defined by the IRE Committee as "an ordered set of symbols which is the normal unit in which information may be stored, transmitted, or operated upon within the computer." It is characterized by the fact that it is usually a single unit of information about the record, such as the "balance on hand" in an inventory application or the employee's current weekly salary in payroll. In some cases it should be pointed out that this restriction does not strictly apply, because there can be a combination of several independent items of information in a single word. This might be referred to as a hybrid word and is frequently resorted to in order to increase the efficiency of storage, when the individual items are short -- such as a yes-no condition. However, this is an exception and in general the comments made will apply even when this hybrid technique is used.

The Word Length Problem

If a machine utilizes a fixed word length it means that all of the items of information within a record and from record to record must be of the same size. This is a rather stringent restriction and one which, in general, results in wasted storage space. Consider, for example, the restriction applied to a payroll application where in a given employee's record is stored several items of information including his annual salary and his hourly rate. In a typical case, the annual salary may require six or seven decimal digits (including the cents digits) and the hourly rate would be typically three decimal digits. If the same size word had to be used for both of these items of information, this word length would have to be at least seven decimal digits long. If it were seven decimal digits, the hourly salary would be using less than one-half of the assigned space. Hence, we see a relatively inefficient storage situation resulting from the fixed word length.

We might also consider the variation of word length for the same word from record to record. Of course, if all of the words within a machine were a fixed word length there would be the same space penalty, not only within the record, but also from record to record. However, there is a degree of variability which has been built into some commercial machines which can be described as follows: The individual words within a record can be of different size but must be pre-set by the programmer for a given application. Then they must be of the same size from record to record. For example, if the annual salary for an individual were word number one and it were assigned seven decimal digits in Record 1 (for a certain employee), it would have to be seven decimal digits for every record (every employee). This would not result in as severe a loss in storage efficiency, however, since the degree of variability for a given word from individual to individual is not as great (probably only the variation from seven to six decimal digits for annual salary). Similarly the hourly rate word might be assigned three decimal digits which would probably accommodate the entire range involved in the payroll.

The Record Length Problem

Now let us turn our attention to the variability in a record size. A record might be defined as all of the individual items of information (or words) about a given file unit (for example, the employee in a payroll application, the part number in an inventory application, the depositor's account information in a commercial check handling application for a bank, etc.).* It is obvious the degree of variability in a record can be greater than that in a word since it can vary not only in the length of the individual words but also in the number of words that make up the record. The later variability can be a much more serious one even than the variability of the word length, in cases where individual transaction detail is to be stored on an account.

For example, in a commercial deposit accounting application in a bank, the number of checks drawn on a given account in a given month, may vary from tens of thousands for a large corporate payroll account to even zero for some individuals' accounts or inactive business accounts (where the business restricts the use of the account to rare entries). In fact, there are quite a few "dormant" accounts in the typical system, which have no activity month after month.

*It is defined by the IRE Committee as "a unit of correlated information relating to a single person or article." However, in many machines a record refers to the largest block of information which can be directly transferred as a unit.

Obviously if a record of fixed size were to be assigned in the electronic data processing system to accommodate all of the depositor's accounts for the bank, it would either be much too large for the inactive accounts resulting in a ridiculous waste in storage space or the more active accounts would exceed the capacity assigned and would overflow. One might now consider assigning different length records to the different accounts based upon past experience or predicted activities. This would certainly result in increased efficiency; but there is also the degree of variability from account to account for a given month. However, even if this technique of assigning a fixed space dependent upon past experience with an account is used, either a much larger capacity than the account needs on the average would have to be assigned or the frequency of overflow would be large. In addition, the procedure involved in assigning a specified space to each individual account in a commercial bank may prove quite unwieldy. This is particularly so, since the bank in general is not aware of how active an account will be when it opens, indeed, the individual depositors may not accurately know -- particularly where there are several special purpose accounts for a business. The controller's office for that business frequently shifts the significance of these accounts and the activity on the individual accounts changes radically.

Variable Vs. Adjustable

Thus, we can see from the above discussion that there is a tremendous advantage in at least being able to set a word length and a record length in advance to different sizes depending upon the application. The author submits that this should not be referred to as a "variable" word and record length technique, but an "adjustable" word and record length. The word and record lengths are set in advance, not necessarily all the same, of course, but of a length which must persist throughout the application. In the case of word lengths, they must be the same for the same word from record to record. In the case of the records, the individual record lengths have been pre-set and must maintain this length during the operating period.

However, even this restricted degree of variability which we have referred to as "adjustability" is of great value in improving the efficiency of storage as we can easily see considering the above two examples of the variable word length (between the annual salary and the hourly rate) and the variable record length (between active corporate accounts and inactive individual accounts). In fact the casual observer might feel that the combination of the adjustable record and the adjustable word length features goes so far toward optimizing record storage efficiency that it would be adequate.

However, let us consider in a little more detail the commercial banking application. The specific dollar amount on individual checks might conceivable vary from even one or two digits to a maximum (in regular

checking accounts for commercial banks) of about ten decimal digits. A study made by the author in a large commercial bank indicated that the average is about 4.5 decimal digits. Hence, if the word length--even if adjustable--assigned to each individual item was ten decimal digits, the efficiency of storage for this information would be less than 50%. Since the programmer or systems designer will not know in advance how many digits the individual transactions charged against the account will be, the adjustable feature is of no assistance in reducing this waste storage space. If, however, the system were able to accommodate a truly "variable" word length in which the individual items would be only as long as required to store the information in the item and would be placed densely in the storage space, then a truly efficient storage system would result. [1]

"Expandable" Record Lengths

If the data processing system, then, accommodates a truly variable word length, the loss in efficiency that would result from requiring that each item have the same word length would be eliminated. However, the wasted space due to the fact that the record length cannot be predicted in advance would still remain. If the record length were adjustable and set on the basis of experience, the wasted space would only be that resulting from fluctuating activity from month to month, but that can be quite large. There is a technique which can eliminate even the wasted space due to the variation in record length from month to month. This technique might be referred to as an "expandable" record length in which there is no fixed space assigned to the record but the entire file of records is constantly rewritten whenever the file is updated. This might be likened to an expandable file drawer whereby the space is signed to any particular account is truly expandable and simply pushes back the rear end of the drawer when necessary.

This system is actually afforded in most magnetic tape file systems, where the record length is not fixed and where the entire tape file is updated; the entire file is rewritten on an output tape and the new items to be entered are merely inserted into the individual account storage and written together with the previously accumulated file for each account on the output tape. Thus, we have an expanding tape file as new activity

---

[1] This has some important implications for the data processing system addressing scheme which is beyond the scope of this paper to discuss, but those readers who are familiar with the problem will recognize it as requiring a word addressing technique rather than a character addressing technique for locating information in storage.

is introduced. In the case of the deposit accounting application for a bank, the tape file length would be a minimum at the beginning of the month and expand to a maximum at the end of the month. Presumably at that time the conventional printed statements would be prepared, and the magnetic tape file for that account wiped clean with the new balance being that obtained at the end of the just concluded month.

Combined Record Technique

Certain types of storage media do not lend themselves to this expanding record length technique. They have, however, other operational advantages which sometimes make it desirable to incorporate these media in a data processing system. For example, a magnetic drum file would normally not be rewritten during each processing, since one of the advantages is that of random access and only the accounts which have been active need to be posted. This reduces the time for updating the file and makes it more feasible to post activity in an "on-line" fashion, in random, and more promptly. In a magnetic tape file, of the type we described, all of the accounts would have to be rewritten on the output tape regardless of the activity ratio.

On a magnetic drum which is not contstantly rewritten,[2] a certain space must be assigned to each record when the data processing application has been established. Of course, it might be changed from month to month based upon experience, but we would still have this difficulty of the variation in the actual activity from the predicted activity either resulting in a low storage efficiency or a high probability of overflow. In order to reduce this in a commercial checking account application, the author investigated the possibility of a "combined record" technique.

If one were to investigate the degree of variability on an individual checking account over a long period of time one would find that this variation was considerably greater than the variation in the combined activity for a large number of similar checking accounts over the same

---

[2]Incidentally, the same conclusions would apply to a magnetic tape file in which the entire file was not rewritten but the new information inserted in the account records storage on the tape.

period of time. This is very familiar to statisticians, and others who have considered the implication of an averaging process.[3]

We might think of this averaging effect which reduces the variability of a group of accounts as follows: If ten similar accounts were combined in a single record storage it would be expected that the variation in the space required for this kind of combined record would be less than the variation required for the ten accounts if they were all kept separately, for the same degree of overflow. This result might be anticipated since the probability of member A of this group of ten being active at precisely the same time as member B is not very great except for such common activity increases as seasonal peaks. A given individual's activity for such personal transactions as buying a house, moving, purchasing an automobile, etc., would not be correlated with other individuals' activities of the same type. (We are not attempting to make the thesis here that there is no correlation, but simply that the correlation is considerably less than one.)

Specifically if we wish to reduce the probability of overflow to some specified value, the amount of storage space required for the ten combined accounts detail lumped into one storage area (i. e., a record) would be considerably less than the amount of storage space required for the total of ten individual accounts for the same probability of overflow. The author became quite interested in the possibilities of this technique and investigated in detail the special checking account application at a large commercial bank.

Since it was considered too difficult to acquire a large amount of data over a long period on individual accounts, the approach was tried on a slightly different problem under the assumption that the same general conclusions would obtain in the case of the temporal variations as for the variations from account to account for a given time period. Hence, a significant sample of special checking accounts was examined for a given month. The distribution of the number of accounts versus the amount of activity in the account was obtained and is shown in Figure 1.

---

[3]It can be proven with the use of mathematical statistics that, regardless of what the individual distribution of activity might be over a period of time, if a sufficiently large number of them were to be combined, the distribution is of a type referred to as a normal or Gaussian distribution. The variability (referred to as standard deviation) of the normal distribution would be less than that of the more radical individual variations.

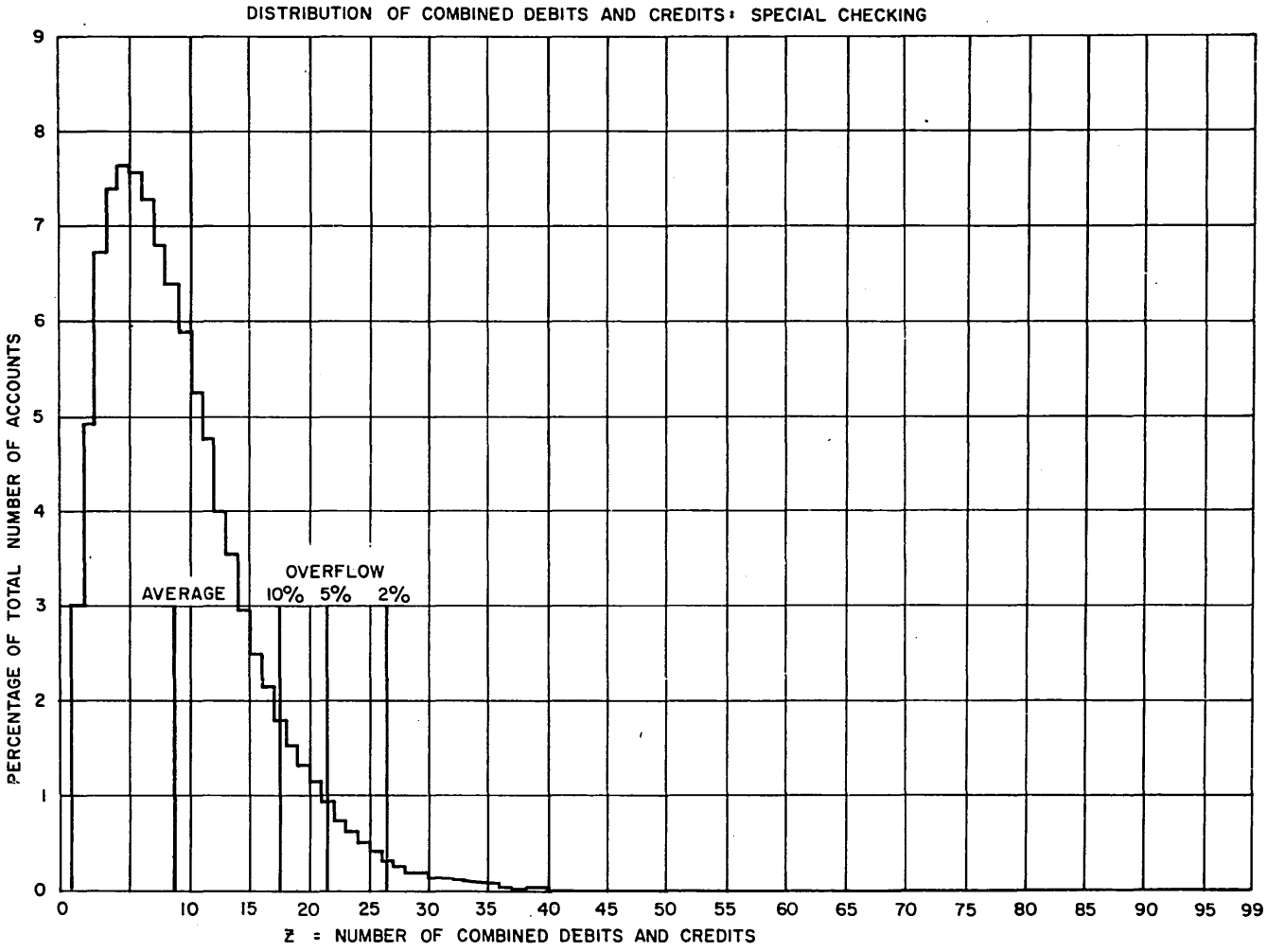
DISTRIBUTION OF COMBINED DEBITS AND CREDITS: SPECIAL CHECKING
PERCENTAGE OF TOTAL NUMBER OF ACCOUNTS
0
1
2
3
4
5
6
7
8
9
AVERAGE
OVERFLOW
10%
5%
2%
0
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
99
Z = NUMBER OF COMBINED DEBITS AND CREDITS

Figure 1

It is seen that, although the average activity for the special checking account was about ten items, in order to reduce the probability of overflow to two per cent (have two per cent of the accounts overflowing), about 26 or 27 transactions would have to be accommodated. The same results are shown in a cumulative distribution in Figure 2. If, however, 16 (in this case 16 was selected because it is a power of two which made the calculations somewhat simpler) accounts were combined, the cumulative distribution shown in Figure 3 results. Here, for a two per cent overflow, the number of transactions to be stored could be reduced to 12. This resulted in a better than 2 to 1 reduction in storage space required.

The results are even more dramatic if one considers probabilities of overflow to be allowed to be considerably less. For example, if it were .1% the number of transactions which must be provided for per account if the accounts are not grouped was 47. If 16 accounts were grouped, this number could be reduced to 14 per cent per account thus resulting in a better than 3 to 1 reduction in storage requirements.

If one were to plot the entire graph indicating the storage space required as a function of the number of accounts grouped for various probabilities of overflow, the results shown in Figure 4 would be obtained.

Of course, it is obvious that there is a disadvantage to this system as opposed to having each individual account stored separately: i.e., the fact that all of the transactions on the combined ten accounts have been lumped together. However, in at least one commercially available data processing system this disadvantage did not prove operationally serious due to the ability to rapidly sort individual transactions from a group of transactions. The technique is based upon a single digit added to each item to indicate which one of the ten accounts the item referred to (in the case where ten transactions were grouped together).

Looking at Figure 4 we can see the signigicance of grouping ten accounts if a 1% overflow figure were to be tolerated. If the transactions had not been combined, a storage space of 31 transactions per account would have been required; but, with a combined record approach, a storage space sufficient to accommodate a little over 13 items is adequate. Thus the storage space required is reduced by about 60%. The technique for selecting which of the ten accounts a given item belongs to on the basis of this single digit is beyond the scope of this paper and would depend upon the specific data processor utilized.

The author feels that this technique of combining similar accounts in a single record storage where the data processor can accommodate the sorting required is a very powerful one, indeed, in reducing the storage space required, particularly where storage space is at a premium as it is in magnetic drum or core storage. As pointed out above, in many applications the advantages of this more expensive storage in the terms of more immediate random access are essential.

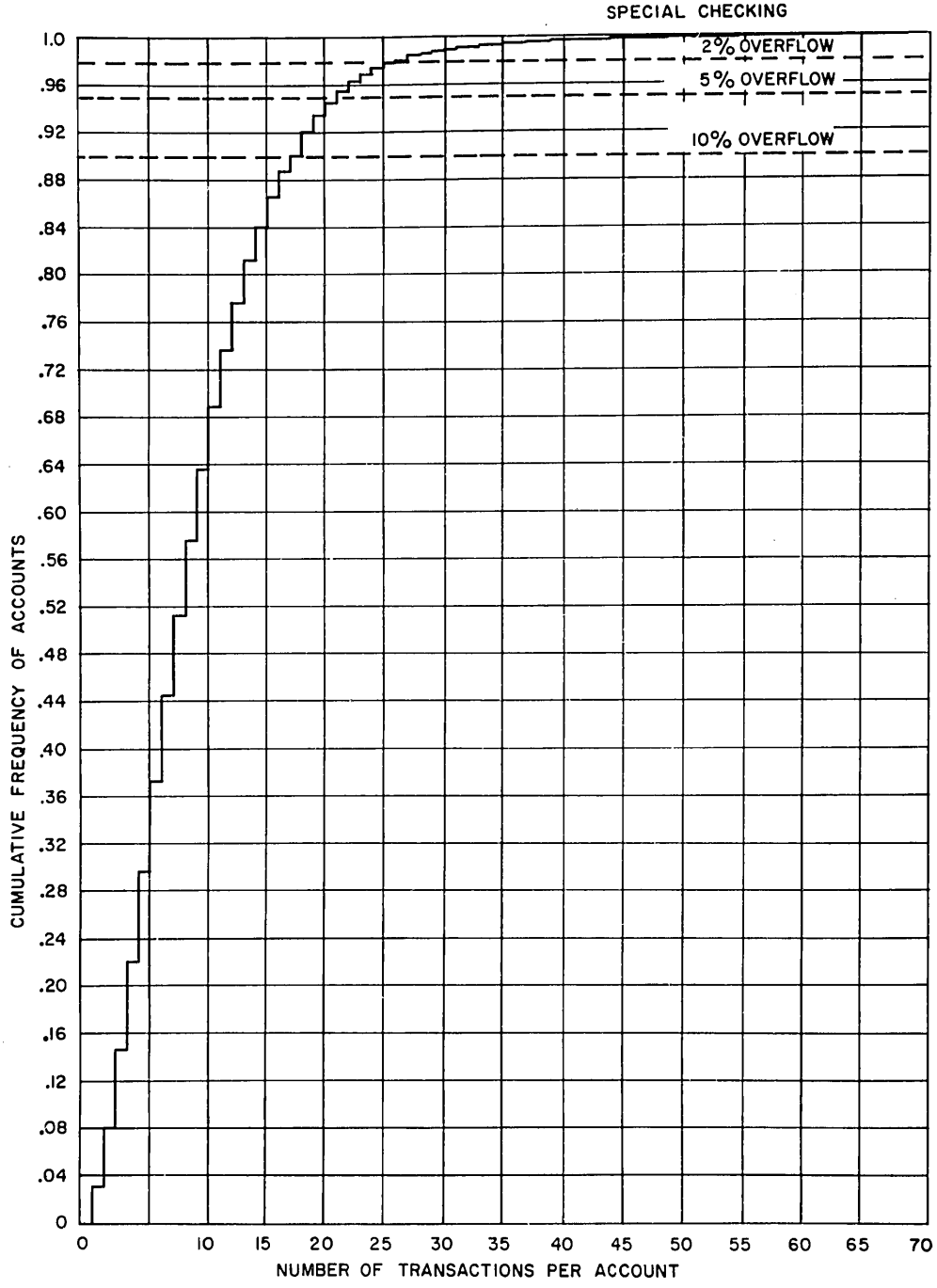
SPECIAL CHECKING
2% OVERFLOW
5% OVERFLOW
10% OVERFLOW
CUMULATIVE FREQUENCY OF ACCOUNTS
1.0
.96
.92
.88
.84
.80
.76
.72
.68
.64
.60
.56
.52
.48
.44
.40
.36
.32
.28
.24
.20
.16
.12
.08
.04
0
0
10
15
20
25
30
35
40
45
50
55
60
65
70
NUMBER OF TRANSACTIONS PER ACCOUNT

Figure 2

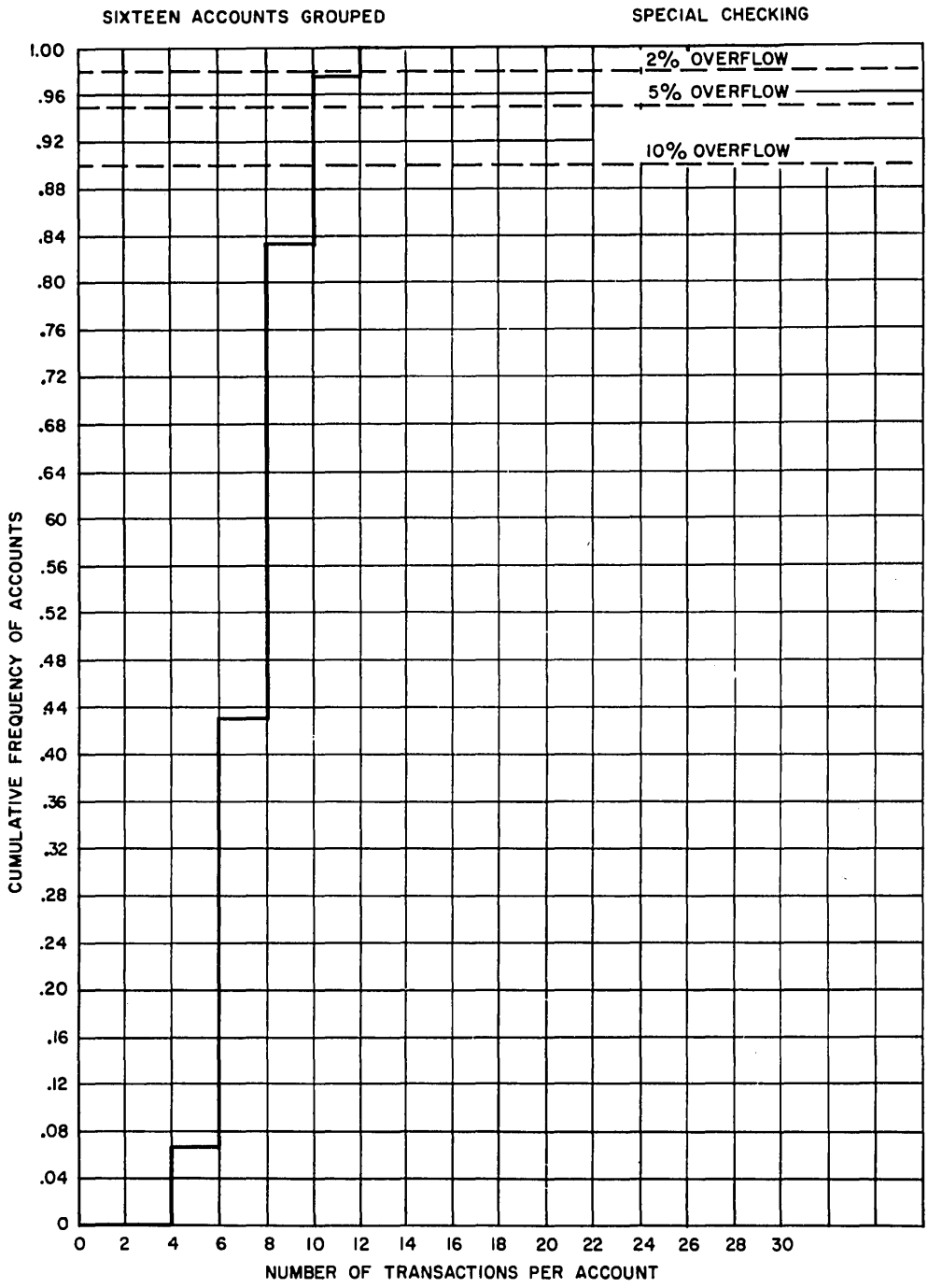
SIXTEEN ACCOUNTS GROUPED
SPECIAL CHECKING
2% OVERFLOW
5% OVERFLOW
10% OVERFLOW
CUMULATIVE FREQUENCY OF ACCOUNTS
1.00
.96
.92
.88
.84
.80
.76
.72
.68
.64
60
.56
.52
.48
.44
.40
.36
.32
.28
.24
.20
.16
.12
.08
.04
0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
NUMBER OF TRANSACTIONS PER ACCOUNT

Figure 3

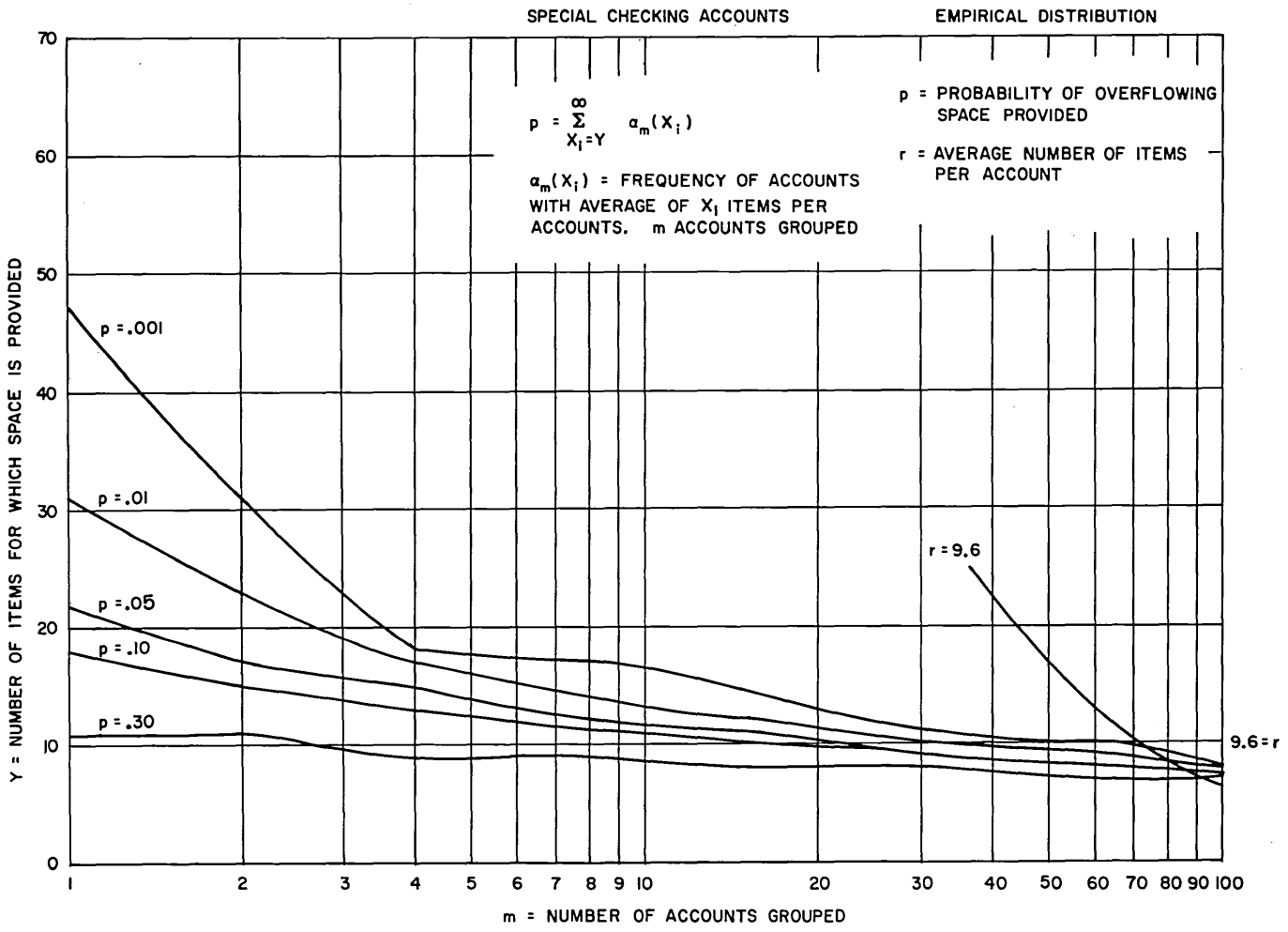
SPECIAL CHECKING ACCOUNTS
EMPIRICAL DISTRIBUTION
$p = \sum_{X_i=Y}^{\infty} \alpha_m(X_i)$
$\alpha_m(X_i)$ = FREQUENCY OF ACCOUNTS WITH AVERAGE OF X_i ITEMS PER ACCOUNTS. m ACCOUNTS GROUPED
p = PROBABILITY OF OVERFLOWING SPACE PROVIDED
r = AVERAGE NUMBER OF ITEMS PER ACCOUNT
p = .001
p = .01
p = .05
p = .10
p = .30
r = 9.6
9.6 = r
Y = NUMBER OF ITEMS FOR WHICH SPACE IS PROVIDED
m = NUMBER OF ACCOUNTS GROUPED
0
10
20
30
40
50
60
70
1
2
3
4
5
6
7
8
9
10
20
30
40
50
60
70
80
90
100

Figure 4