NASA TECHNICAL NOTE

NASA TN D-3616

c. 1

TECH LIBRARY KAFB, NM
0130311

LOAN COPY: RETU
AFWL (WLIL-2
KIRTLAND AFB, N

NASA TN D-3616

# A STATISTICAL DATA COMPRESSION TECHNIQUE

*by James W. Snively, Jr.*

*Goddard Space Flight Center*
*Greenbelt, Md.*

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • OCTOBER 1966

TECH LIBRARY KAFB, NM
0130311

# A STATISTICAL DATA COMPRESSION TECHNIQUE

By James W. Snively, Jr.

Goddard Space Flight Center
Greenbelt, Md.

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151 – Price $2.00

## ABSTRACT

Data contained in a histogram can be compressed by calculating and transmitting two quantities: The area of the histogram and the sum of the squares of each bar of the histogram. The paper presents a survey of what knowledge one has about the original histogram when given only these two quantities. A set of theorems is derived which indicates the magnitude limits of the individual bars of the histogram as a function of these two quantities. The technique results in a data compression factor of greater than 10 for certain scientific experiments where the only information required is the amplitude distribution of the individual histogram bars.

CONTENTS


Abstract . . . . . ii

INTRODUCTION . . . . . 1

TECHNIQUE . . . . . 1

Relationship between Area and Sum of Squares for a Histogram . . . . . 1

Hardware Use of this Relationship . . . . . 2

INFORMATION CONTENT OF THE AREA AND THE SUM OF SQUARES . . . . . 3

Definition of Ratio, r . . . . . 3

Peak Height . . . . . 3

Amplitude Distribution . . . . . 6

FURTHER HARDWARE CONSIDERATIONS . . . . . 10

SUMMARY AND CONCLUSIONS . . . . . 12

References . . . . . 12

Appendix—Mathematical Development . . . . . 13

# A STATISTICAL DATA COMPRESSION TECHNIQUE

by
James W. Snively, Jr.
*Goddard Space Flight Center*

## INTRODUCTION

Every person involved in spacecraft telemetry soon realizes that the amount of data desired or collected often exceeds the amount that can be transmitted. One method used to measure a phenomenon is to collect the data in the form of a histogram. This histogram can take a form such as particle detector counts versus the angular position of the sensor. It takes many bits to send back a complete histogram involving large amounts of counts — often more bits than are available. One way of solving this problem is to have equipment in the satellite to process incoming data and transmit only the *results* of its calculations. This paper presents the mathematical basis for a novel type of data compression that involves the transmission of only the area of a histogram and certain digits of the sum of the squares of the histogram bars. The amount of information about the histogram that can be recovered from these two quantities is surprising. For example, Figure 1 shows a pair of typical histograms expected in a plasma-measuring experiment. Knowledge of the total counts over all azimuthal angles combined with knowledge of the sum of the squares of each count in 22.5° intervals will readily distinguish interplanetary space curves from transition region curves and, in fact, allows one to make much finer distinctions.

## TECHNIQUE

### Relationship between Area and Sum of Squares for a Histogram

Assume that the area, $A$, of a histogram is known. The first theorem of the paper (Appendix) states that the sum of the squares of the histogram bars, $\Sigma$, must lie between two numbers which differ by a factor equal to the number of bars, $n$, in the histogram. The larger of these two limits for the sum of the squares corresponds to a histogram where all but one of the bars contain no counts. Therefore, this

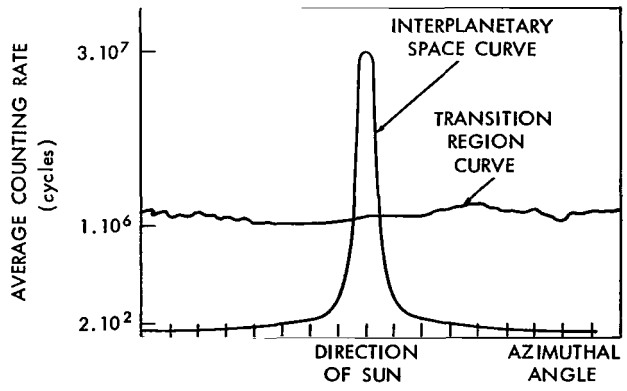


Figure 1 — Typical histograms expected in a plasma-measuring experiment.

larger limit for the sum of the squares is simply the square of the area, $A^2$. The smaller limit for the sum of the squares corresponds to a histogram where all of the bars have equal numbers of counts. In this case, the lower limit turns out to be $A^2/n$.

These facts illustrated in Figure 2 show the relationship between the area and the sum of the squares for histograms of sixteen bars. Sixteen bar histograms (with area less than $2^{19}$ counts) lie within the cross-hatched region between the two parallel lines. For example, a sixteen bar histogram with an area of $2^6$ must have its sum of squares between $2^{12}$ and $2^8$. This absolute constraint is taken advantage of in the telemetry technique we wish to describe.

## Hardware Use of this Relationship

If the quantity A is known, then according to the preceding section, the sum of squares, $\Sigma$, is known to within a factor of n. In other words, the location of the most significant bit of $\Sigma$ is known to within $\log_2 n$ bit positions. Specifically for a sixteen-bar histogram the most significant bit must lie within one of four positions. If, for this sixteen bar histogram, A is $2^6$ counts, then the most significant bit of $\Sigma$ is either the *9th, 10th, 11th* and *12th* bit of the word. The telemetry technique being expounded is simply the transmission of A and $\log_2 n$ bits of $\Sigma$. More bits of $\Sigma$ can be transmitted if still more accuracy is desired.

This study was conducted for use with a spacecraft bourne plasma measuring experiment (Reference 1) that will be flown on future Interplanetary Monitoring Platform (IMP) satellites. The histogram collected in this experiment contains sixteen bars, each containing the number of counts produced by a plasma detector during one-sixteenth of a spin of the satellite. Figure 3 shows a block diagram of the equipment designed to implement this telemetry technique in this particular application.

The area counter at the bottom of the diagram commutates four bits of the sum of squares counter to the telemetry system. Note that in this application a logarithmic counter (Reference 2) is used for the area determination. Therefore, although the total area of the

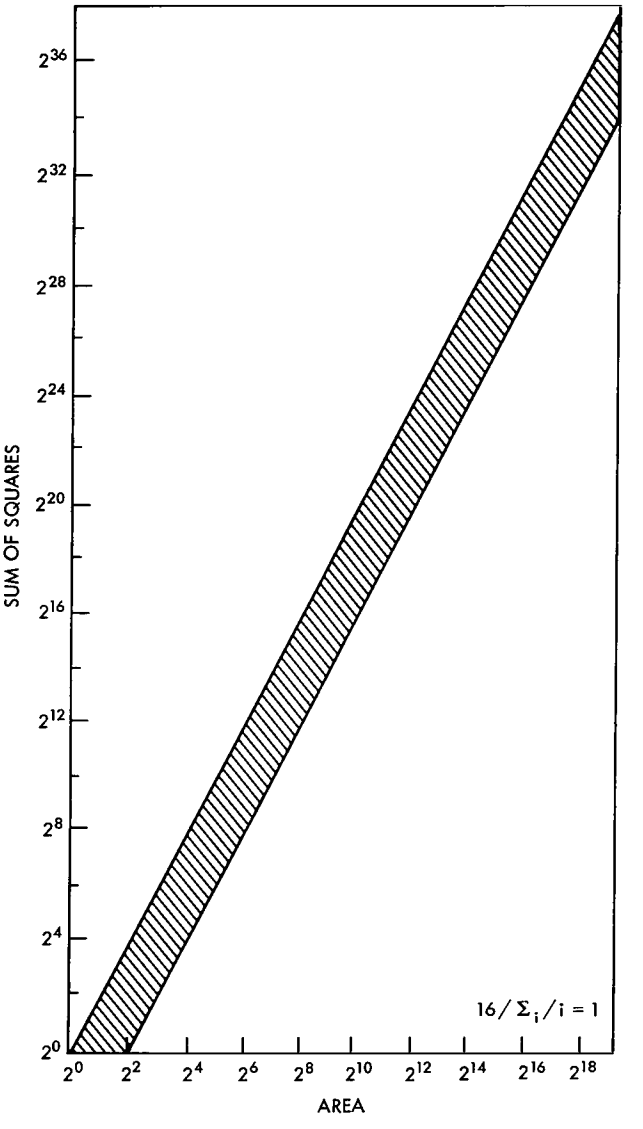


Figure 2 — Relationship between area and sum of squares for a histogram with 16 bars.

collected histogram can be as large as $2^{19}$ counts, only eight bits are required to represent this number to a ±3 percent accuracy.

Transmission of the twelve indicated bits allows one to determine the sum of the squares to at worst ±33 percent. This error occurs when the four commutated bits of the sum of squares counter are 0001. When these four bits are 1111 the worst error is ±3.3 percent.

The remainder of this paper is theoretical. It assumes that the area and sum of squares are known exactly. Consequences of such conditions will be discussed. The consequences obtained may be applied to cases where the area and sum of squares are not known exactly.

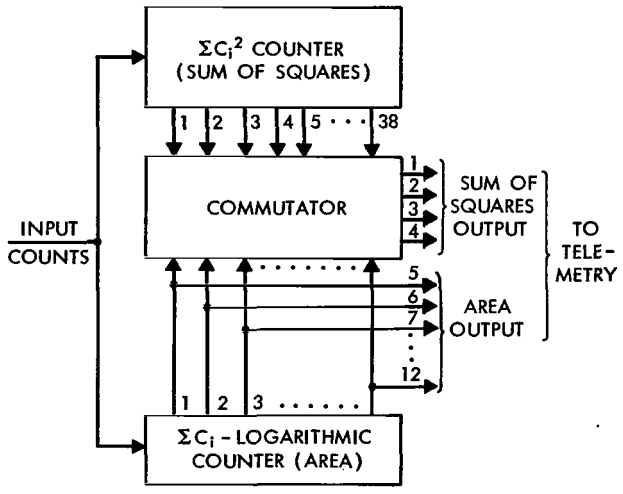


Figure 3 — Block diagram of processing equipment.

## INFORMATION CONTENT OF THE AREA AND THE SUM OF SQUARES

### Definition of Ratio, r

A quantity which we shall call the "ratio" denoted by the symbol, $r$, is defined as

$$r = \frac{n \sum c_i^2}{\left( \sum c_i \right)^2} \quad . \tag{1}$$

Here $n$ denotes the number of bars in the histogram and $c_i$ denotes the number of counts in the $i^{th}$ largest bar. This parameter is related to the mean, $\mu$, and the variance, $\sigma^2$, by the equation,

$$r = 1 + \frac{\sigma^2}{\mu^2} \quad , \tag{2}$$

but has the advantage of only assuming values between 1 and $n$. For this reason, it is a useful parameter for describing various properties of histograms.

### Peak Height

The value of the largest bar of a histogram, expressed as a fraction of the area of the histogram, cannot be greater than the area, $A$, nor smaller than an $n^{th}$ of the area, $A/n$, where $n$ is the number of bars in the histogram. If the largest bar has the value $A$ then all other bars must have the value

zero. If this largest bar has the value A/n all others must have this same value. These two cases correspond to ratios of n and of 1 respectively. These are the extreme cases.

For any given ratio the largest bar of a histogram can only assume values in a sub-interval of the interval between A/n and A. For a ratio of r the greatest possible value for the largest histogram bar is given by Theorem 2 in the appendix

$$\frac{A}{n}\left(1 + \sqrt{(n-1)(r-1)}\right) . \tag{3}$$

This expression reduces to A when $r = n$ and to A/n when $r = 1$. Therefore, the expression for the largest histogram bar agrees with the known histograms for the extreme values of r.

If the largest bar of a histogram with ratio r is equal to the value given by Expression 3, that histogram in this special case is completely determined. In fact, the theorem states that all of the remaining $n - 1$ bars are equal. Therefore, these bars must be equal to the value of the expression,

$$\frac{a}{n}\left(1 - \sqrt{\frac{r-1}{n-1}}\right) , \tag{4}$$

which is merely the area not included in the large bar divided by $n - 1$. Figure 4 shows several five bar histograms. Each histogram has the greatest largest bar possible for the indicated r. Note that the largest bar decreases as r decreases and that the base of the histogram increases as r decreases.

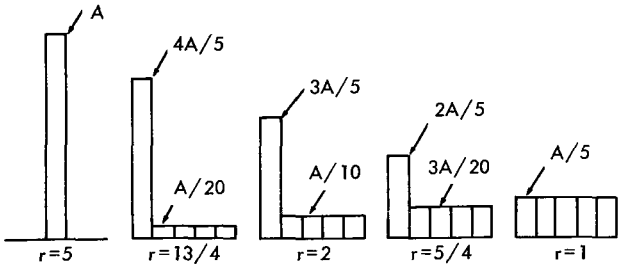


Figure 4 — Histograms (n = 5) with the largest peak height for the indicated ratios.

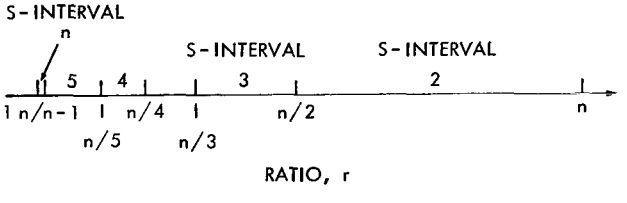


Figure 5 — A segment of the real number line illustrating the definition of s-intervals.

The ratio r of any histogram must lie in the interval between 1 and n. Divide this interval into $n - 1$ sub-intervals (or s-intervals) and label these intervals from 2 to n as shown in Figure 5. Any sub-interval s contains ratios between n/s and $n/(s - 1)$. The third theorem in the appendix states that this integer s is the fewest number of non-zero bars possible for histograms with ratios in the s-interval. For example, if $r = 1$ then s equals n since "one" lies in the interval bounded by 1 and $n/n - 1$. Therefore, n bars must have non-zero values for a ratio of 1. Another example is that if $n = 16$ and the ratio equals 5, then since 5 lies between 16/4 and 16/3, s equals 4. Therefore this sixteen bar histogram must have *at least* four non-zero bars. There is no restriction on *how many* non-zero bars a histogram may have.

The smallest possible value for the largest bar is shown in Theorem 4 in the appendix, to be

$$\frac{A}{s}\left(1 + \sqrt{\frac{rs - n}{n(s-1)}}\right) \quad . \tag{5}$$

When $r = n$ we have $s = 2$, and the expression reduces to A. When $r = 1$ we have $s = n$ and the expression reduces to $A/n$. Therefore, this expression agrees with the known histograms for the extreme values of r.

If the largest bar of a histogram with ratio r is equal to the value given by Expression 5, that histogram in this special case is completely determined. From the proof of Theorem 4 we discover that it is a histogram with s bars, $s - 1$ of which are equal to the value of Expression 5. The remaining bar has the value

$$\frac{A}{s}\left(1 - \sqrt{\frac{(s-1)(rs-n)}{n}}\right) \quad . \tag{6}$$

Figure 6 shows several five bar histograms. Each histogram has the smallest largest bar possible for the indicated r. Note that the largest bars decrease as r decreases and that the smaller bar increases as r decreases until it equals the larger bars. Then as r continues to decrease a new bar increases from zero until it equals the larger bars again. This process continues until we have n equal bars.

Figure 7 illustrates for sixteen bar histograms how the bounds on the largest histogram bar vary with the ratio. All histograms must lie within the cross-hatched region between the two bounds.

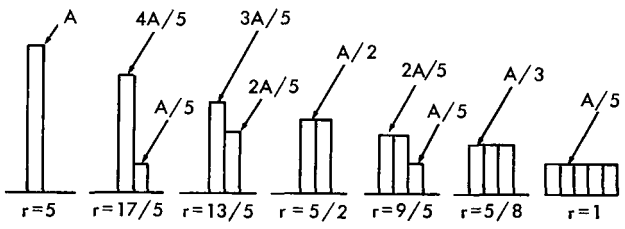


Figure 6—Histograms (n = s) with the smallest peak height for indicated ratios.

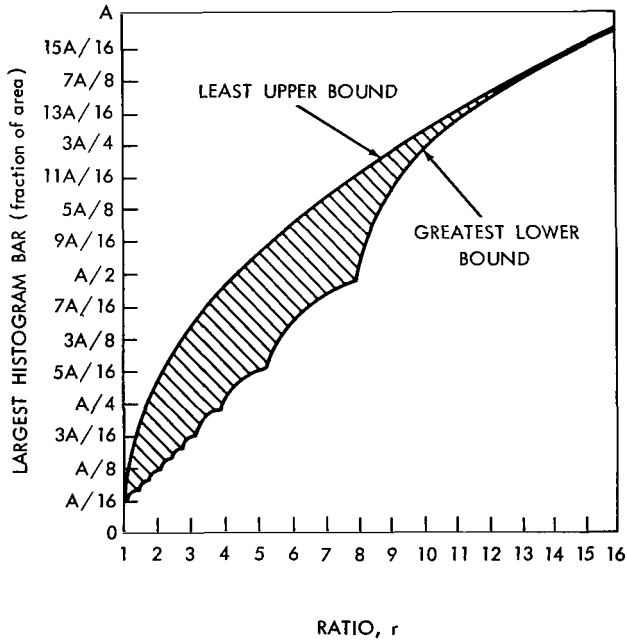


Figure 7—Bounds for the largest histogram bar (n = 16).

Figure 8 is a plot for sixteen-bar histograms of the largest possible plus-or-minus percent error to which the largest histogram bar is known as a function of the ratio. Note that for the sharp peaked curves the percentage error is quite small. The maximum percentage error of ±43 percent occurs for a histogram with a ratio of slightly less than two. This corresponds to a flat curve where the peak height is of lesser significance than for the very peaked curves of the higher ratios.

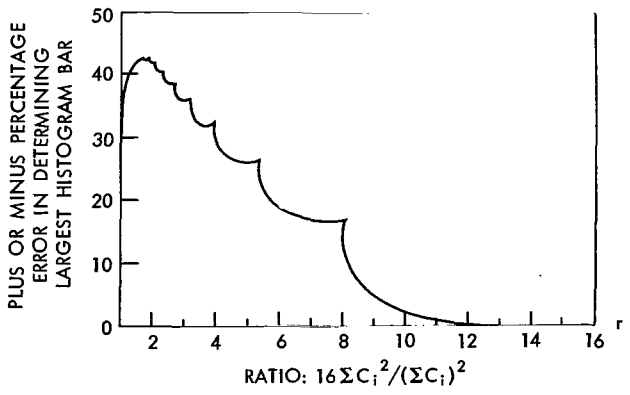


Figure 8—Percentage error in the largest histogram bar vs. ratio.

## Amplitude Distribution

Another quantity which can be deduced from a knowledge of the area and the sum of squares of a histogram is amplitude distribution. In fact, Theorems 5 and 6 give bounds for each bar of a histogram that are similar to those for the largest one. Let $c_1$ refer to the largest bar of a histogram, $c_2$ to the second largest bar, etc., so that $c_n$ refers to the $n^{th}$ largest bar of the histogram. In this notation for the case of a histogram with n bars $c_n$ will be the smallest bar of the histogram.

Theorem 5 in the appendix gives the largest possible value for the $p^{th}$ largest bar of a histogram with area, A, and ratio, r. This value must be computed in one of two ways depending on the value of the ratio. If the ratio is between n/p and n, the largest possible value for the $p^{th}$ largest bar is given by

$$\frac{A}{p}\left(1 - \sqrt{\frac{rp - n}{n(p - 1)}}\right) . \tag{7}$$

If the ratio is between 1 and n/p, the largest possible value for the $p^{th}$ largest bar is given by

$$\frac{A}{n}\left(1 + \sqrt{\frac{(n - p)(r - 1)}{p}}\right) . \tag{8}$$

If $p = 1$, the entire range of ratios is covered by Expression 8 which reduces to Expression 3 when 1 is substituted for p. Thus it follows that Theorem 2 is a corollary to Theorem 5.

When the ratio is equal to n/p (except for $p = 1$) both Expressions 7 and 8 give identical results, namely, that the largest possible value for the $p^{th}$ largest histogram bar is exactly A/p. This value is the largest possible value that the $p^{th}$ largest bar of a histogram can have for any ratio. It corresponds to a histogram with exactly p, equal, non-zero bars.

Theorem 6 in the appendix gives the smallest possible value for the $p^{th}$ largest histogram bar of a histogram with area, A, and ratio, r, for these cases not covered by previous theorems.

(Remember that Theorem 4 gives the smallest possible value for the largest bar, as in Expression 5, and that when the ratio is between $n/p - 1$ and $n$, the smallest possible value for the $p^{th}$ bar is zero from Theorem 3.) According to Theorem 6 if the ratio is between 1 and $n/(p - 1)$, the smallest possible value for the $p^{th}$ largest histogram bar ($p \geq 2$) is

$$\frac{A}{n}\left(1 - \sqrt{\frac{(p-1)(r-1)}{n-p+1}}\right) \quad . \tag{9}$$

When the ratio is 1, Expression 9 reduces to $A/n$. This corresponds to the known extreme histogram for this ratio, namely, the one with $n$ equal bars.

When the ratio is $n/(p - 1)$ Expression 9 reduces to zero. This is in agreement with the result of Theorem 3.

Figure 9 illustrates how the bounds on some of the histogram bars vary with ratio for the case where $n = 16$. Note that bounds on the largest bar were already illustrated in Figure 7. For bars

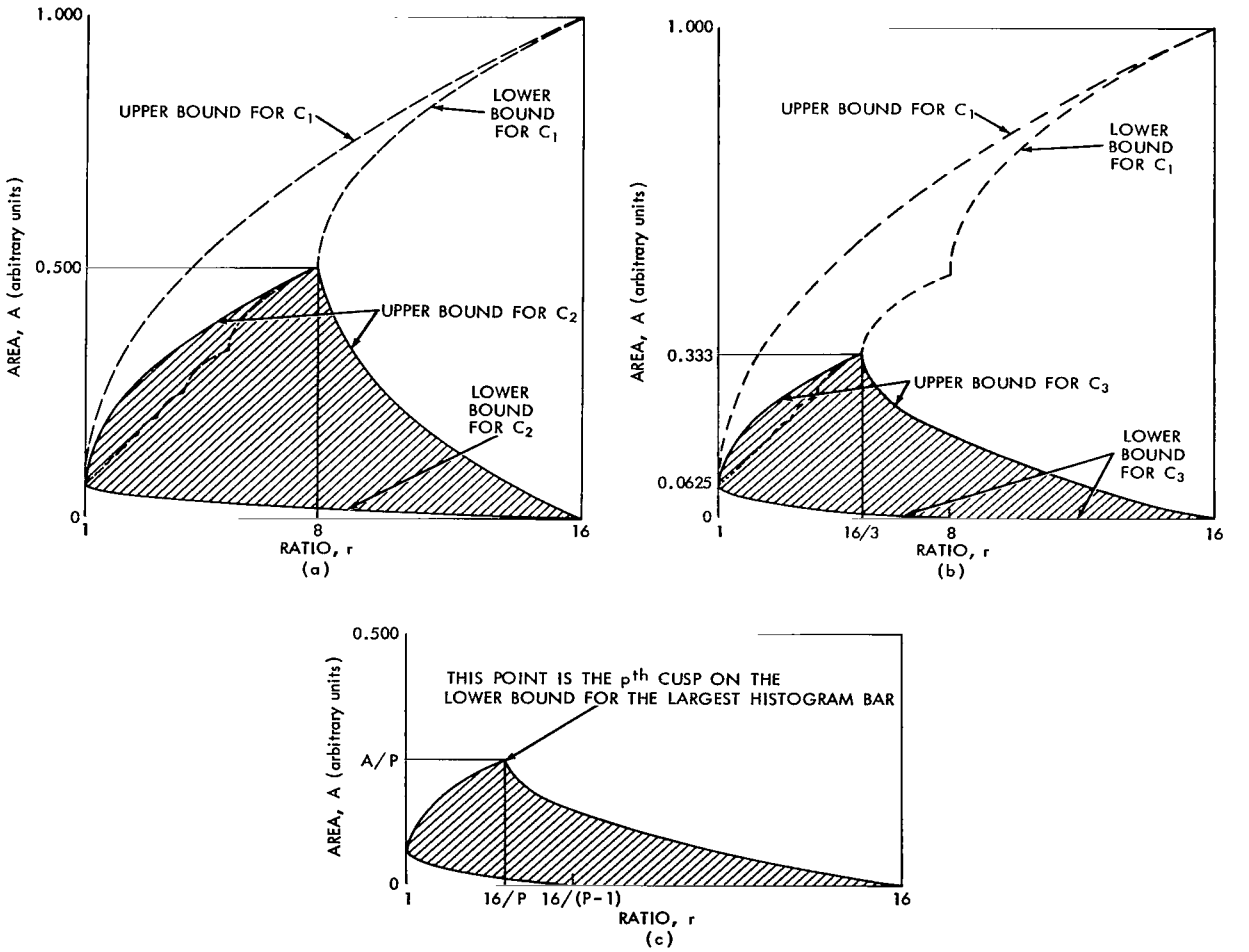


Figure 9—Bounds for (a) the second largest histogram bar, $C_2$, ($n = 16$), (b) the third largest histogram bar, $C_3$, ($n = 16$), and (c) the $p^{th}$ largest histogram bar $C_p$, ($n = 16$).

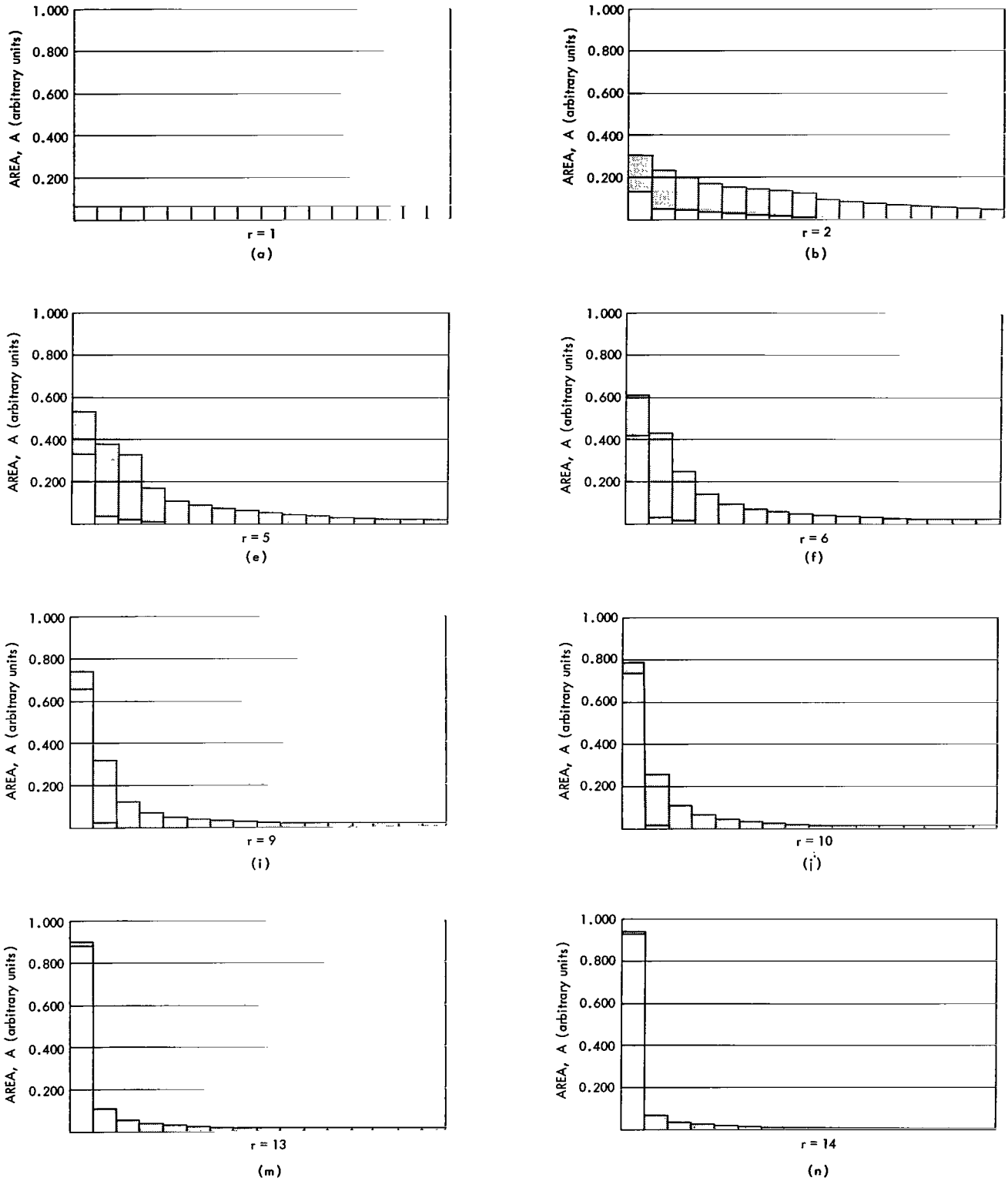


Figure 10 (a through p) — Largest and smallest bars for histograms with the indicated ratios. (The bars of any histogram with the indicated ratio must lie in the shaded region.)

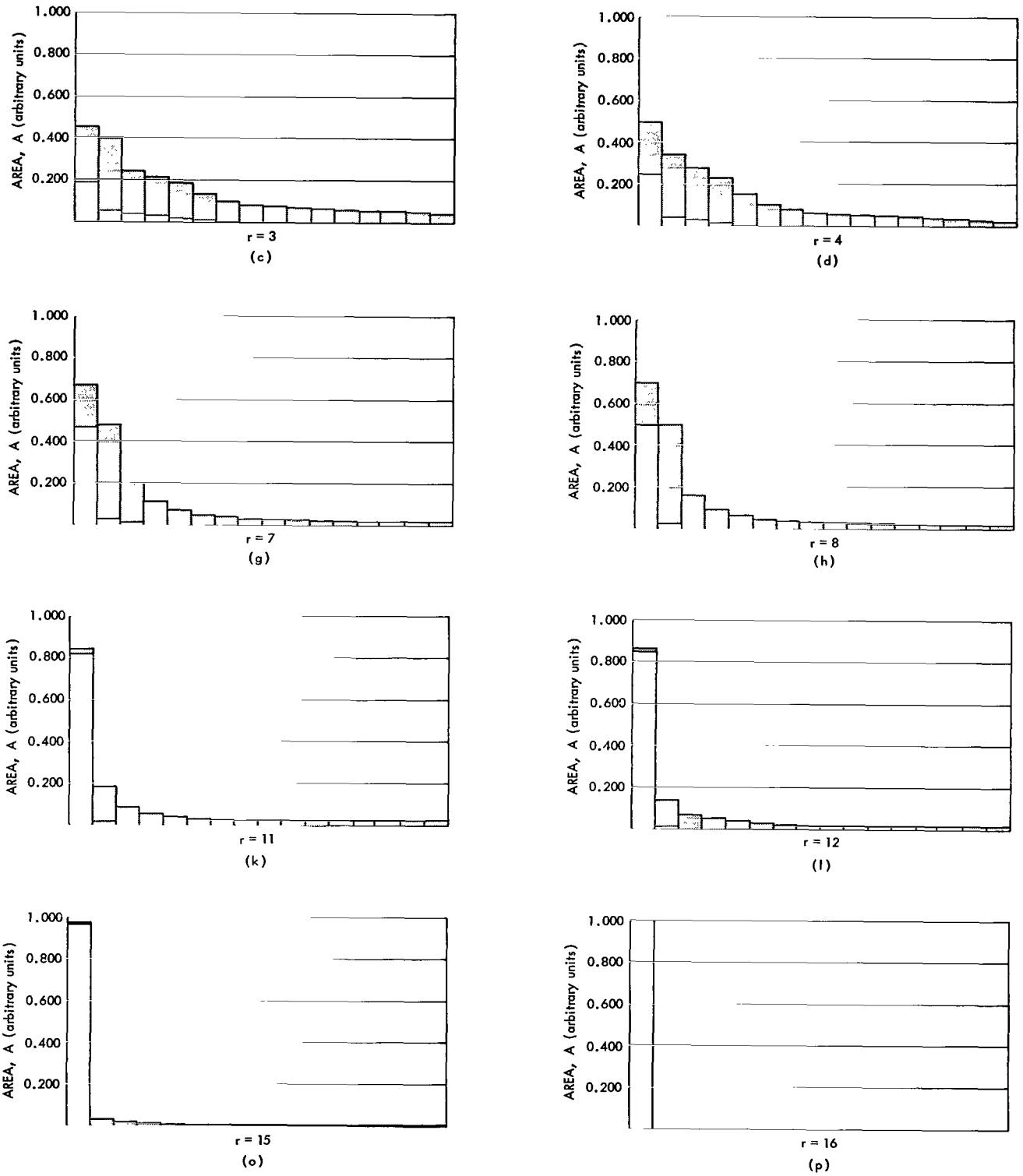


Figure 10 (a through p) — Largest and smallest bars for histograms with the indicated ratios. (The bars of any histogram with the indicated ratio must lie in the shaded region.)

other than the largest, the least upper bound for the $p^{th}$ largest bar increases from A/n when $r = 1$, to A/p when $r = n/p$, and then decreases to zero when $r = n$, and the greatest lower bound decreases from A/n when $r = 1$ to zero when $r = n/(p - 1)$. The two extreme cases $r = 1$ and $r = n$ are the only completely determined ones.

Figure 10 illustrates the consequences of these results for the case where $n = 16$. For ratios with integer values, the bounds for each histogram bar are drawn with the largest bar on the left and the smallest bar on the right. In this form one can see at a glance how the shape of the resulting histograms varies as the ratio changes. Note that histograms with larger ratios are much steeper and narrower than those with lower ratios.

## FURTHER HARDWARE CONSIDERATIONS

In the preceding section the information content of the area and sum of squares of a histogram was discussed. The discussion was based on these quantities' being known exactly. In practice, however, as in the IMP plasma experiment, these quantities will not be known exactly, but for a given set of bits these quantities will be known to lie between certain known bounds. For example, suppose the area, A, is known to lie between $A_{max}$ and $A_{min}$ and the sum of squares, $\Sigma$, is known to lie between $\Sigma_{max}$ and $\Sigma_{min}$. Then the ratio, r, must satisfy the following bounds:

$$\frac{n \sum_{min}}{A_{max}^2} \leq r \leq \frac{n \sum_{max}}{A_{min}^2} \quad . \tag{10}$$

Similarly, bounds for any bar of the histogram could be found by choosing the largest upper bound and the smallest lower bound for the bar from among those for all ratios in the range of Expression 10.

Using the flight hardware for the IMP plasma experiment designed for a maximum number of counts in any histogram bar of $2^{17}$ and a maximum total area over 16 bars of $2^{19}$ counts, the ratio, r, is always determined to better than ±40.0 percent of its range, but on the average it is determined to about ±12.0 percent.

Figure 11 shows a segment of the output of a computer program which relates the output of the IMP flight hardware to the input data which produced the specified output. For example, if the log counter output is 189 (275 octal) the area of the input histogram lies between 29,696 counts and 30,719 counts. If, furthermore, the squarer output is 12, the ratio of the input histogram must be between 13.64 and 15.85 and the largest bar of this input histogram is between 27,312 counts and 30,571 counts. For each of these quantities the harmonic mean (H.M.) of the range and the maximum ± percentage error (P.E.) are also listed.

| ABITS | AREA.. | MIN. | MAX. | H.M. | P.E. |
|---|---|---|---|---|---|
| 189 (275) | | 29696.00 | 30719.00 | 30198.84 | 1.69 |

| SBITS | RATIO.. | MIN. | MAX. | H.M. | P.E. | C( 1).. | MIN. | MAX. | H.M. | P.E. |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | 1.00 | 1.24 | 1.11 | 10.53 | | 1856.00 | 5527.94 | 2778.97 | 49.73 |
| 1 | | 1.12 | 2.45 | 1.54 | 37.26 | | 2099.35 | 10883.28 | 3519.75 | 67.66 |
| 2 | | 2.26 | 3.67 | 2.80 | 23.81 | | 4216.92 | 14071.71 | 6489.20 | 53.88 |
| 3 | | 3.40 | 4.89 | 4.01 | 18.00 | | 6675.80 | 16582.46 | 9519.30 | 42.59 |
| 4 | | 4.53 | 6.11 | 5.20 | 14.77 | | 8991.18 | 18722.12 | 12148.24 | 35.11 |
| 5 | | 5.67 | 7.32 | 6.39 | 12.70 | | 11664.00 | 20618.53 | 14899.36 | 27.74 |
| 6 | | 6.81 | 8.54 | 7.58 | 11.27 | | 13582.24 | 22339.56 | 16893.43 | 24.38 |
| 7 | | 7.95 | 9.76 | 8.76 | 10.22 | | 14799.78 | 23926.42 | 18287.66 | 23.57 |
| 8 | | 9.09 | 10.98 | 9.94 | 9.42 | | 20319.01 | 25406.30 | 22579.65 | 11.13 |
| 9 | | 10.22 | 12.19 | 11.12 | 8.79 | | 22676.71 | 26798.30 | 24565.83 | 8.33 |
| 10 | | 11.36 | 13.41 | 12.30 | 8.27 | | 24473.25 | 28116.44 | 26168.66 | 6.93 |
| 11 | | 12.50 | 14.63 | 13.48 | 7.85 | | 25983.63 | 29371.36 | 27573.83 | 6.12 |
| 12 | | 13.64 | 15.85 | 14.66 | 7.49 | | 27312.31 | 30571.37 | 28850.09 | 5.63 |
| 13 | | 14.78 | 16.00 | 15.36 | 3.98 | | 28512.40 | 30719.00 | 29574.60 | 3.73 |
| 14 | | 15.91 | 16.00 | 15.96 | 0.27 | | 29615.28 | 30719.00 | 30157.05 | 1.83 |

| ABITS | AREA.. | MIN. | MAX. | H.M. | P.E. |
|---|---|---|---|---|---|
| 190 (276) | | 30720.00 | 31743.00 | 31223.12 | 1.64 |

| SBITS | RATIO.. | MIN. | MAX. | H.M. | P.E. | C( 1).. | MIN. | MAX. | H.M. | P.E. |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | 1.00 | 1.15 | 1.07 | 7.19 | | 1920.00 | 5009.03 | 2775.95 | 44.58 |
| 1 | | 1.05 | 2.29 | 1.44 | 37.20 | | 2030.29 | 10720.40 | 3414.01 | 68.15 |
| 2 | | 2.12 | 3.43 | 2.62 | 23.72 | | 4188.21 | 13963.10 | 6443.65 | 53.85 |
| 3 | | 3.18 | 4.57 | 3.75 | 17.91 | | 6125.34 | 16498.58 | 8933.86 | 45.85 |
| 4 | | 4.25 | 5.71 | 4.87 | 14.67 | | 8780.43 | 18652.75 | 11940.22 | 35.99 |
| 5 | | 5.31 | 6.84 | 5.98 | 12.60 | | 10219.43 | 20558.76 | 13652.44 | 33.59 |
| 6 | | 6.38 | 7.98 | 7.09 | 11.17 | | 13444.01 | 22286.61 | 16771.13 | 24.75 |
| 7 | | 7.44 | 9.12 | 8.20 | 10.12 | | 14794.25 | 23878.52 | 18269.43 | 23.49 |
| 8 | | 8.51 | 10.26 | 9.30 | 9.32 | | 19233.88 | 25362.29 | 21877.00 | 13.74 |
| 9 | | 9.57 | 11.39 | 10.41 | 8.68 | | 22174.21 | 26757.35 | 24251.14 | 9.37 |
| 10 | | 10.64 | 12.53 | 11.51 | 8.17 | | 24183.83 | 28077.93 | 25985.80 | 7.45 |
| 11 | | 11.71 | 13.67 | 12.61 | 7.74 | | 25814.02 | 29334.83 | 27462.03 | 6.38 |
| 12 | | 12.77 | 14.81 | 13.71 | 7.39 | | 27222.25 | 30536.44 | 28784.26 | 5.74 |
| 13 | | 13.84 | 15.95 | 14.82 | 7.08 | | 28480.20 | 31689.49 | 29999.26 | 5.33 |
| 14 | | 14.90 | 16.00 | 15.43 | 3.55 | | 29627.66 | 31743.00 | 30648.87 | 3.45 |
| 15 | | 15.97 | 16.00 | 15.98 | 0.10 | | 30689.47 | 31743.00 | 31207.34 | 1.69 |

Figure 11—Segment of the output of a computer program relating output of IMP hardware to the input data which produced it.

## SUMMARY AND CONCLUSIONS

This paper has shown how transmission of the area of a histogram and certain bits of the sum of squares enables one to recover much of the original histogram. It has shown how knowledge of the area of the histogram implies a restriction on the sum of squares of the histogram. This fact is the backbone of the entire process for it enables the sum of squares to be transmitted with a smaller number of bits if the area is also transmitted than if it is not. The paper has also shown how knowledge of both the area and sum of squares of a histogram implies restrictions on each of the bars of the histogram.

As spacecraft travel farther away from the earth, the need for on-board processing of data will increase. The IMP plasma experiment is an example of a situation where onboard processing is necessary in order to be able to accomplish desired goals. The computation discussed in this paper has been able to increase the effective amount of information that can be transmitted to Earth from this experiment by more than an order of magnitude.

(Manuscript received October 7, 1965)

REFERENCES

1. Wilkerson, T. D. and Ogilvie, K. W., "Plasma Experiments for Space Vehicles," University of Maryland TN BN-378, October, 1964.

2. Schaefer, D. H., "Logarithmic Compression of Binary Numbers," Proceedings of the IRE 49(7); 1219, July, 1961.

Appendix A

# Mathematical Development

## Theorem 1

Let $n$ be a positive integer and let $c_i$ $(i=1, 2, \ldots, n)$ be non-negative real numbers. Then

$$\frac{1}{n}\left(\sum_{i=1}^{n} c_i\right)^2 \leqq \sum_{i=1}^{n} c_i^2 \leqq \left(\sum_{i=1}^{n} c_i\right)^2 . \tag{A1}$$

*Proof:* One form of the well known Schwartz inequality is:

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leqq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) ,$$

for any arbitrary real numbers $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$. Choosing $a_1 = a_2 = \ldots = a_n = 1$ and $b_i = c_i$ $(i=1, 2, \ldots, n)$ leads to the following result:

$$\left(\sum_{i=1}^{n} c_i\right)^2 \leqq n \sum_{i=1}^{n} c_i^2 ,$$

from which the left hand inequality of Equation A1 follows by division by $n$. Next the following equality shall be established for $n \geqq 2$ by using incomplete induction.

$$\left(\sum_{i=1}^{n} c_i\right)^2 = \sum_{i=1}^{n} c_i^2 + 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} c_i c_j . \tag{A2}$$

If $n = 2$ Equation A2 becomes: $(c_1 + c_2)^2 = c_1^2 + 2c_1c_2 + c_2^2$ which is clearly true. Assume Equation A2 holds for $n$. Then

$$\left(\sum_{i=1}^{n+1} c_i\right)^2 = \left(\sum_{i=1}^{n} c_i + c_{n+1}\right)^2 = \left(\sum_{i=1}^{n} c_i\right)^2 + 2c_{n+1}\left(\sum_{i=1}^{n} c_i\right) + c_{n+1}^2 .$$

Apply the induction hypothesis to the first term of the last expression:

$$\left(\sum_{i=1}^{n+1} c_i\right)^2 = \sum_{i=1}^{n} c_i^2 + 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} c_i c_j + 2c_{n+1} \left(\sum_{i=1}^{n} c_i\right) + c_{n+1}^2 ,$$

whence

$$\left(\sum_{i=1}^{n+1} c_i\right)^2 = \sum_{i=1}^{n+1} c_i^2 + 2 \sum_{i=2}^{n+1} \sum_{j=1}^{i-1} c_i c_j .$$

Thus, if Equation A2 is true for $n$ it is also true for $n+1$, since it is true for 2, it must be true for $n \geq 2$. Now by hypothesis $c_i \geq 0$, therefore $c_i c_j \geq 0$ and

$$2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} c_i c_j \geq 0 .$$

Combining this last inequality with Equation A2 yields the right hand inequality of Equation A1.

## Theorem 2

$$\text{Let} \quad \sum_{i=1}^{n} c_i = A \quad \text{and} \quad \frac{n \sum_{i=1}^{n} c_i^2}{\left(\sum_{i=1}^{n} c_i\right)^2} = r ,$$

then $c_1 \leq \frac{A}{n} \left[1 + \sqrt{(n-1)(r-1)}\right]$.

*Proof:* The second hypothesis may be rewritten

$$n \left(c_1^2 + \sum_{i=2}^{n} c_i^2\right) = r \left(c_1 + \sum_{i=2}^{n} c_i\right)^2 .$$

From Theorem 1 it follows that

$$\frac{1}{n-1}\left(\sum_{i=2}^{n} c_i\right)^2 \leqq \sum_{i=2}^{n} c_i^2 \ .$$

Combining the last two relations and transposing to one side gives

$$(n-r)\,c_1^2 - 2rc_1\left(\sum_{i=1}^{n} c_i\right) + \left(\frac{n}{n-1} - r\right)\left(\sum_{i=2}^{n} c_i\right)^2 \leqq 0 \ .$$

Since

$$A = \sum_{i=1}^{n} c_i, \quad \sum_{i=2}^{n} c_i = A - c_1$$

and the last expression becomes

$$(n-r)\,c_1^2 - 2rc_1\,(A-c_1) + \left(\frac{n}{n-1} - r\right)\,(A-c_1)^2 \leqq 0 \ ,$$

$$\frac{n^2}{n-1}\,c_1^2 - \frac{2n}{n-1}\,c_1 A + \left(\frac{n}{n-1} - r\right)\,A^2 \leqq 0 \ ,$$

$$n^2 c_1^2 - 2nc_1 A + [n - r\,(n-1)]\,A^2 \leqq 0 \ ,$$

$$n^2 c_1^2 - 2nc_1 A + A^2 \leqq (n-1)\,(r-1)\,A^2 \ ,$$

$$(nc_1 - A)^2 \leqq (n-1)\,(r-1)\,A^2 \ ,$$

$$nc_1 - A \leqq \sqrt{(n-1)\,(r-1)}\ A \ ,$$

and

$$c_1 \leqq \frac{A}{n}\left[1 + \sqrt{(n-1)\,(r-1)}\right] \ .$$

## Theorem 3

Let s denote the number of non-zero bars in a histogram. Then for this histogram the ratio

$$r = \frac{n\sum_{i=1}^{n} c_i^2}{\left(\sum_{i=1}^{n} c_i\right)^2} \geqq \frac{n}{s} \ .$$

In particular it follows that

$$\frac{n}{r} \leqq s \leqq n \ .$$

*Proof:* Assume without loss of generality that $c_{s+1} = c_{s+2} = \ldots = c_n = 0$ and let $c_i = a_i c_1$ for $i = 2, 3, \ldots, s$. Then

$$r = \frac{n\,[c_1^2 + (a_2c_1)^2 + (a_3c_1)^2 + \cdots + (a_sc_1)^2]}{[c_1 + (a_2c_1) + (a_3c_1) + \cdots + (a_sc_1)]^2}$$

$$= \frac{n\,(1 + a_2^2 + a_3^2 + \cdots + a_s^2)}{(1 + a_2 + a_3 + \cdots + a_s)^2} \ .$$

It is asserted that this expression for $r$ assumes a minimum when $a_2 = a_3 = \ldots = a_s = 1$. To show this let $a_i = 1 + \epsilon_i$ $(i = 2, 3, \ldots, s)$. Then

$$r = \frac{n\left[1 + (1 + \epsilon_2)^2 + (1 + \epsilon_3)^2 + \cdots + (1 + \epsilon_s)^2\right]}{[1 + (1 + \epsilon_2) + (1 + \epsilon_3) + \cdots + (1 + \epsilon_s)]^2}$$

$$= \frac{n\left(s + 2\sum_{i=2}^{s} \epsilon_i + \sum_{i=2}^{s} \epsilon_i^2\right)}{\left(s + \sum_{i=2}^{s} \epsilon_i\right)^2}$$

$$= \frac{\frac{n}{s}\left\{s^2 + 2s\sum_{i=2}^{s} \epsilon_i + \left[\left(\sum_{i=2}^{s} \epsilon_i\right)^2 - \left(\sum_{i=2}^{s} \epsilon_i\right)^2\right] + s\sum_{i=2}^{s} \epsilon_i^2\right\}}{\left(s + \sum_{i=2}^{s} \epsilon_i\right)^2}$$

$$= \frac{n}{s}\left(1 + \frac{s\sum_{i=2}^{s} \epsilon_i^2 - \left(\sum_{i=2}^{s} \epsilon_i\right)^2}{\left(s + \sum_{i=2}^{s} \epsilon_i\right)^2}\right) .$$

But from Schwartz' inequality (which doesn't depend on the $\epsilon_i$'s being positive) it follows that

$$\left(\sum_{i=2}^{s} \epsilon_i\right)^2 \leqq (s-1) \sum_{i=2}^{s} \epsilon_i^2 ,$$

and making the right hand side still larger,

$$\left(\sum_{i=2}^{s} \epsilon_i\right)^2 \leqq s \sum_{i=2}^{s} \epsilon_i^2 ,$$

and

$$s \sum_{i=2}^{s} \epsilon_i^2 - \left(\sum_{i=2}^{s} \epsilon_i\right)^2 \geqq 0 ,$$

whence

$$\frac{s \sum_{i=2}^{s} \epsilon_i^2 - \left(\sum_{i=2}^{s} \epsilon_i\right)^2}{\left(s + \sum_{i=2}^{s} \epsilon_i\right)^2} \geqq 0 .$$

Thus since $r$ takes its minimum value when this last expression is zero it follows that:

$$r \geqq \frac{n}{s} .$$

*Lemma:* Let $u$ and $v$ be the values of the non-zero bars of an $n$-bar histogram with $i$ bars equal to $u$ and $j$ bars equal to $v$, where $i + j \leqq n$, the area of the histogram: $i\,u + j\,v = A$, and the sum of squares of the histogram: $i\,u^2 + j\,v^2 = rA^2/n$. Then

$$u = \frac{A}{i+j}\left(1 + \sqrt{\frac{j}{i}}\,R\right) , \tag{A3}$$

and

$$v = \frac{A}{i+j}\left(1 - \sqrt{\frac{i}{j}}\,R\right) , \tag{A4}$$

where

$$R^2 = \frac{(i + j)\, r}{n} - 1 \ .$$

Furthermore $u \geq v$ whenever $r \geq n/(i + j)$ and $v \geq 0$ whenever $n/i \geq r$.

*Proof:* Let us verify the four assertions about u and v. First let us consider the difference u-v:

$$u - v = \frac{A}{i + j} \left( 1 + \sqrt{\frac{j}{i}}\, R - 1 + \sqrt{\frac{i}{j}}\, R \right)$$

$$= \frac{A}{i + j} \left( \sqrt{\frac{j}{i}} + \sqrt{\frac{i}{j}} \right) R \geq 0.$$

This result will be true whenever R is a real number which is the case whenever $R^2 \geq 0$ or equivalently whenever $r \geq n/(i + j)$.

Next let us consider $v \geq 0$. This will be true whenever $1 - \sqrt{i/j}\, R \geq 0$ or equivalently whenever $1 \geq (i/j)\, R^2$. Substituting for $R^2$ we have $1 \geq (i/j)\, [(i + j)r/n - 1]$. This result will be true whenever $j/i + 1 \geq (i + j)r/n$ or equivalently $n/i \geq r$.

Next let us consider the sum

$$iu + jv = \frac{iA}{i + j} \left( 1 + \sqrt{\frac{j}{i}}\, R \right) + \frac{jA}{i + j} \left( 1 - \sqrt{\frac{i}{j}}\, R \right)$$

$$= \frac{A}{i + j} \left( i + \sqrt{ij}\, R + j - \sqrt{ij}\, R \right) = A \ .$$

Finally let us consider the sum of squares:

$$iu^2 + jv^2 = \frac{iA^2}{(i + j)^2} \left( 1 + 2\sqrt{\frac{j}{i}}\, R + \frac{jR^2}{i} \right) + \frac{jA^2}{(i + j)^2} \left( 1 - 2\sqrt{\frac{i}{j}}\, R + \frac{iR^2}{j} \right)$$

$$= \frac{A^2}{(i + j)^2} \left( i + j + (j + i)\, R^2 \right) = \frac{A^2}{i + j} \left( 1 + \frac{(i + j)\, r}{n} - 1 \right) = \frac{rA^2}{n} \ .$$

*Corollary:* Let $\xi = u/v$, where u and v are as defined in the lemma. Then $1 \geq \xi \geq 0$ if $n/(i + j) \leq r \leq n/i$, $\xi = 1$ only for $r = n/(i + j)$ and $\xi = 0$ only for $r = n/i$.

*Proof:* If $n/(i + j) \leqq r \leqq n/i$ then $u \geqq v \geqq 0$. Dividing by $u$ gives $1 \geqq \xi \geqq 0$. From Equation A5 and the definition of $r$ we note that $u = v$ only if $r = n/(i + j)$ whence $\xi = 1$ only if $r = n/(i + j)$. Also we note that by replacing the symbol $\geqq$ by equality in the proof that $v \geqq 0$ when $n/i \geqq r$ we obtain $v = 0$ when $n/i = r$.

## Theorem 4

Let the real numbers $c_1, c_2, \ldots, c_n$ represent a histogram with area $A$ and sum of squares $rA^2/n$. Order these numbers so that $c_1 \geqq c_2 \geqq \ldots \geqq c_n \geqq 0$. Then if $n/s \leqq r \leqq n/(s - 1)$ where $s$ is an integer between 2 and $n$, a histogram that has the smallest possible value for $c_1$ is of the form of the lemma with $i = s - 1$ and $j = 1$. The value of $c_1$ under these conditions is

$$\frac{A}{s}\left(1 + \sqrt{\frac{rs - n}{n(s - 1)}}\right).$$

*Proof:* Pick an integer, $s$, between 2 and $n$. Then let $i = s - 1$, $j = 1$ in Equations A3 and A4 of the lemma to obtain the histogram:

$$\left.\begin{aligned} q_1 = q_2 = \cdots = q_{s-1} &= \frac{A}{s}\left(1 + \sqrt{\frac{rs - n}{n(s - 1)}}\right) \\ q_s &= \frac{A}{s}\left(1 - \sqrt{\frac{(s - 1)(rs - n)}{n}}\right) \\ q_{s+1} = q_{s+2} = \cdots = q_n &= 0 \end{aligned}\right\} \qquad \text{(A5)}$$

which must satisfy the hypothesis of the theorem whenever $n/s \leqq r \leqq n/(s - 1)$. To prove that this histogram assumes the greatest lower bound for the largest histogram bar let the numbers $c_i$ $(i = 1, 2, \ldots, n)$ represent any other histogram which satisfies the hypothesis of the theorem. We wish to show that $c_1 \geqq q_1$.

Let the $c_i$'s be related to the $q_i$'s so that

$$c_i = q_i + q_1 \epsilon_i \quad (i = 1, 2, \cdots, n), \qquad \text{(A6)}$$

where the $\epsilon_i$'s may be either positive, negative, or zero. Then we have

$$A = \sum_{i=1}^{n} c_i = \sum_{i=1}^{n} (q_i + q_1 \epsilon_i) = \sum_{i=1}^{n} q_i + q_1 \sum_{i=1}^{n} \epsilon_i = A + q_1 \sum_{i=1}^{n} \epsilon_i .$$

Thus we have

$$\sum_{i=1}^{n} \epsilon_i = 0 \cdot \tag{A7}$$

In this and the succeeding theorems we shall consider a function, $f$, of $n$ variables defined as

$$f = f(\epsilon_1, \epsilon_2, \ldots, \epsilon_n) = \left(\frac{1}{q_1}\right)^2 \left(\sum_{i=1}^{n} c_i^2 - \sum_{i=1}^{n} q_i^2\right)$$

$$= \sum_{i=1}^{n} \left[\left(\frac{c_i}{q_1}\right)^2 - \left(\frac{q_i}{q_1}\right)^2\right] \cdot$$

Substituting for $c_1$ from Equation A6

$$f = \sum_{i=1}^{n} \left[\left(\frac{q_i}{q_1} + \epsilon_i\right)^2 - \left(\frac{q_i}{q_1}\right)^2\right] \tag{A8}$$

Letting $\xi = q_s/q_1$ (which is consistent with the definition of $\xi$ in the corollary to the lemma) and substituting in Equation A8 we have

$$f = \sum_{i=1}^{s-1} [(1 + \epsilon_i)^2 - 1] + [(\xi + \epsilon_s)^2 - \xi^2] + \sum_{i=s+1}^{n} \epsilon_i^2$$

$$= \sum_{i=1}^{s-1} (1 + 2\epsilon_i + \epsilon_i^2 - 1) + (\xi^2 + 2\xi\epsilon_s + \epsilon_s^2 - \xi^2) + \sum_{i=s+1}^{n} \epsilon_i^2$$

$$= 2 \sum_{i=1}^{s-1} \epsilon_i + 2\xi\epsilon_s + \sum_{i=1}^{n} \epsilon_i^2 \cdot$$

Since $\sum_{i=1}^{n} \epsilon_i = 0$ we have $\epsilon_s = - \sum_{i=1}^{s-1} \epsilon_i - \sum_{i=s+1}^{n} \epsilon_i$ ,

$$f = 2 \sum_{i=1}^{s-1} \epsilon_i - 2\xi \sum_{i=1}^{s-1} \epsilon_i - 2\xi \sum_{i=s+1}^{n} \epsilon_i + \sum_{i=1}^{s-1} \epsilon_i^2$$

$$+ \epsilon_s \left( - \sum_{i=1}^{s-1} \epsilon_i - \sum_{i=s+1}^{n} \epsilon_i \right) + \sum_{i=s+1}^{n} \epsilon_i^2$$

$$= \left( \sum_{i=1}^{s-1} \epsilon_i \right) (2 - 2\xi - \epsilon_s) + \sum_{i=1}^{s-1} \epsilon_i^2$$

$$+ \left( \sum_{i=s+1}^{n} \epsilon_i \right) (- 2\xi - \epsilon_s) + \sum_{i=s+1}^{n} \epsilon_i^2$$

$$= \sum_{i=1}^{s-1} \epsilon_i [\epsilon_i - \epsilon_s + 2(1 - \xi)]$$

$$+ \sum_{i=s+1}^{n} \epsilon_i [\epsilon_i - \epsilon_s - 2\xi] \cdot \qquad \text{(A9)}$$

Now assume $\epsilon_1 < 0$. By the ordering hypothesis on the $c_i$'s we have, upon dividing by $q_1$:

$$(1 + \epsilon_1) \geqq (1 + \epsilon_2) \geqq \cdots \geqq (1 + \epsilon_{s-1}) \geqq (\xi + \epsilon_s) \geqq \epsilon_{s+1} \geqq \cdots \geqq \epsilon_n \geqq 0 \ ,$$

whence

$$\epsilon_1 \geqq \epsilon_2 \geqq \cdots \geqq \epsilon_{s-1} \ ,$$

so that for $i = 1, 2, \ldots, s-1$ we have $\epsilon_i < 0$ since $\epsilon_1 < 0$. Also for $i = 1, 2, \ldots, s-1$ we have $1 + \epsilon_i \geq \xi + \epsilon_s$. Whence $\epsilon_i - \epsilon_s + 1 - \xi \geq 0$. From the corollary to the lemma we have $1 - \xi \geq 0$ so that by adding inequalities we have $[\epsilon_i - \epsilon_s + 2(1 - \xi)] \geq 0$ for $i = 1, 2, \ldots, s-1$. Thus it follows that the first $s - 1$ terms in Equation A9 are $\leq 0$.

It also follows from the ordering that for $i = s+1, s+2, \ldots, n$ we have $\epsilon_i \geq 0$. Also for $i = s+1, s+2, \ldots, n$ we have $\xi + \epsilon_s \geq \epsilon_i$. From the corollary we have $\xi \geq 0$ so that $[\epsilon_i - \epsilon_s - 2\xi] \leq 0$. Combining these results it follows that $f \leq 0$.

Also from the corollary we have $1 - \xi < 0$ if $r > n/s$ from which it follows that $f < 0$ since we already know that $f \leq 0$ and the first $s - 1$ terms of Equation A9 are all $< 0$.

If $r = n/s$, the corollary states that $\xi = 1$. In this case the ordering hypothesis gives us $1 + \epsilon_s \geq \epsilon_i$ for $i = s+1, s+2, \ldots, n$ so that $\epsilon_i - \epsilon_s - 1 \leq 0$ whence $\epsilon_i - \epsilon_s - 2 < 0$. Now if none of the first $s - 1$ terms of Equation A9 is $< 0$ we can conclude that $\epsilon_i(\epsilon_i - \epsilon_s) = 0$ for $i = 1, 2, \ldots, s-1$ and therefore that $\epsilon_1 = \epsilon_2 = \ldots = \epsilon_s$. Since

$$\sum_{i=1}^{n} \epsilon_i = 0$$

at least one of $\epsilon_{s+1}, \epsilon_{s+2}, \ldots, \epsilon_n$ must be $> 0$ and the corresponding term of Equation A9 will be $< 0$. Hence in all cases $f < 0$. But this implies that

$$\sum_{i=1}^{n} c_i^2 < \sum_{i=1}^{n} q_i^2 = \frac{rA^2}{n},$$

which contradicts the assumption that the $c_i$'s satisfy the hypothesis of the theorem. Thus the assumption that $\epsilon_1 < 0$ must be false, that is $\epsilon_1 \geq 0$ so that

$$c_1 \geq q_1 = \frac{A}{s}\left(1 + \sqrt{\frac{rs - n}{n(s-1)}}\right).$$

## Theorem 5

Let the real numbers $c_1, c_2, \ldots, c_n$ represent a histogram with area $A$ and sum of squares $rA^2/n$. Order these numbers so that $c_1 \geq c_2 \geq \ldots \geq c_n \geq 0$. Then a histogram that has the largest possible value for $c_p$ ($p$ between 2 and $n$) is of the form of the lemma with $i = 1$ and $j = p - 1$ if $n/p < r \leq n$ for any $p$'s and with $i = p$ and $j = n - p$ if $1 \leq r \leq n/p$. The corresponding values of

$c_p$ under these conditions are

$$\frac{A}{p}\left(1 - \sqrt{\frac{pr - n}{n(p-1)}}\right) \text{ if } \frac{n}{p} \leqq r \leqq n \quad \text{for} \quad p \geqq 2$$

and

$$\frac{A}{n}\left(1 + \sqrt{\frac{(n-p)(r-1)}{p}}\right) \text{ if } 1 \leqq r \leqq \frac{n}{p} \quad \text{for all } p$$

*Proof:* Case I: Let $n/p \leqq r \leqq n$ and let the $i = 1$ and $j = p - 1$ in Equations A3 and A4 of the lemma to obtain the histogram:

$$\left.\begin{aligned} q_1 &= \frac{A}{p}\left(1 + \sqrt{\frac{(p-1)(pr-n)}{n}}\right), \\ q_2 = q_3 = \cdots = q_p &= \frac{A}{p}\left(1 - \sqrt{\frac{pr-n}{n(p-1)}}\right), \\ q_{p+1} = q_{p+2} = \cdots = q_n &= 0, \end{aligned}\right\} \tag{A10}$$

which must satisfy the hypothesis of the theorem whenever $n/p \leqq r \leqq n$. Let the numbers $c_i (i = 1, 2, \ldots, n$ represent any other histogram which satisfies the hypothesis of the theorem. We wish to show that $c_p \leqq q_p$. If we define the numbers $\epsilon_i$ $(i = 1, 2, \ldots, n)$ as in Equation A6 using the $q_i$'s as defined in Equation A10 we conclude that Equation A7 holds by the same argument as before. We shall also define the function f as before (Equation A8) and noting that $q_p/q_1 = \xi$ we obtain

$$f = [(1 + \epsilon_1)^2 - 1] + \sum_{i=2}^{p} [(\xi + \epsilon_i)^2 - \xi^2] + \sum_{i=p+1}^{n} \epsilon_i^2$$

$$= \epsilon_1(2 + \epsilon_1) + \sum_{i=2}^{p} \epsilon_i (2\xi + \epsilon_i) + \sum_{i=p+1}^{n} \epsilon_i^2 \cdot$$

From $\sum_{i=1}^{n} \epsilon_i = 0$ we have

$$\epsilon_1 = - \sum_{i=2}^{n} \epsilon_i \cdot$$

Substituting this we have

$$f = - 2 \sum_{i=2}^{n} \epsilon_i + \left( \sum_{i=2}^{n} \epsilon_i \right)^2 + 2\xi \sum_{i=2}^{p} \epsilon_i + \sum_{i=2}^{p} \epsilon_i^2 + \sum_{i=p+1}^{n} \epsilon_i^2$$

$$= \left( \sum_{i=2}^{n} \epsilon_i \right)^2 + \sum_{i=2}^{n} \epsilon_i^2 - 2 (1 - \xi) \sum_{i=2}^{p} \epsilon_i - 2 \sum_{i=p+1}^{n} \epsilon_i$$

and

$$f = \sum_{i=2}^{p} \epsilon_i \left( \sum_{j=2}^{n} \epsilon_j + \epsilon_i - 2 (1 - \xi) \right) + \sum_{i=p+1}^{n} \epsilon_i \left( \sum_{j=2}^{n} \epsilon_j + \epsilon_i - 2 \right)$$

$$= \sum_{i=2}^{p} \epsilon_i \left( \sum_{j=2}^{n} \epsilon_j - (1 - \xi) \right) + \sum_{i=p+1}^{n} \epsilon_i \left( \sum_{i=2}^{n} \epsilon_j - 1 \right)$$

$$+ \sum_{i=2}^{p} \epsilon_i [\epsilon_i - (1 - \xi)] + \sum_{i=p+1}^{n} \epsilon_i (\epsilon_i - 1)$$

$$= \left( \sum_{i=2}^{p} \epsilon_i \right) \left[ \sum_{j=2}^{n} \epsilon_j - (1 - \xi) \right] + \left( \sum_{i=p+1}^{n} \epsilon_i \right) \left( \sum_{j=2}^{n} \epsilon_j - 1 \right)$$

$$+ \sum_{i=2}^{p} \epsilon_i [\epsilon_i - (1 - \xi)] + \sum_{i=p+1}^{n} \epsilon_i (\epsilon_i - 1).$$

Since the $c_i$'s must satisfy the ordering hypothesis of the theorem, we have upon dividing by $q_j$

$$(1 + \epsilon_1) \geqq (\xi + \epsilon_2) \geqq \cdots \geqq (\xi + \epsilon_p) \geqq \epsilon_{p+1} \geqq \cdots \geqq \epsilon_n \geqq 0.$$

Thus $\epsilon_i \geqq 0$ for $i = p + 1, p + 2, \ldots, n$. Furthermore $\epsilon_2 \geqq \epsilon_3 \geqq \ldots \geqq \epsilon_p$. Assume $\epsilon_p > 0$. Then $\epsilon_i > 0$ for $i = 2, 3, \ldots, p$. Thus in Equation A11 the left factor in each term is at least non-negative. Also

$$1 - \xi \geqq \epsilon_2 - \epsilon_1 > - \epsilon_1 = \sum_{j=2}^{n} \epsilon_j ,$$

whence

$$\left( \sum_{j=2}^{n} \epsilon_j - (1 - \xi) \right) < 0.$$

From the corollary to the lemma it follows that $\xi \geqq 0$ so that

$$\left( \sum_{i=2}^{n} \epsilon_j - 1 \right) < 0 \quad \text{also.}$$

Similarly $[\epsilon_i - (1 - \xi)] < 0$ for $i = 2, \ldots, p$ and $(\epsilon_i - 1) < 0$ for $i = p + 1, \ldots, n$. From these results it follows that $f < 0$, since the right hand factor in each term of Equation A11 is negative and the left hand factor in the first term of Equation A11 is positive.

But then,

$$\sum_{i=1}^{n} c_i^2 < \sum_{i=1}^{n} q_i^2 = \frac{rA^2}{n} ,$$

which is a contradiction. Thus $\epsilon_p \leqq 0$ so that $c_p \leqq q_p$.

*Case II:* Let $1 \leqq r \leqq n/p$ and let $i = p$ and $j = n - p$ in Equations A3 and A4. We then obtain the histogram:

$$\left.\begin{aligned} q_1 = q_2 = \cdots = q_p &= \frac{A}{n}\left(1 + \sqrt{\frac{(n-p)(r-1)}{p}}\right), \\ q_{p+1} = q_{p+2} = \cdots = q_n &= \frac{A}{n}\left(1 - \sqrt{\frac{p(r-1)}{n-p}}\right), \end{aligned}\right\} \tag{A12}$$

which satisfies the hypothesis of the theorem whenever $1 \leqq r \leqq n/p$. As in *Case I* we let the numbers $c_i$ $(i = 1, 2, \ldots, n)$ represent any other histogram which satisifes the hypothesis of the theorem, and define quantities $\epsilon_i$ $(i = 1, 3, \ldots, n)$ as in Equation A6. We again wish to show $c_p \leqq q_p$ and so proceed as before to show that Equation A7 holds and to define the function, f, using Equation A8. By substituting $q_n/q_1 = \xi$

$$f = \sum_{i=1}^{p} [(1 + \epsilon_i)^2 - 1] + \sum_{i=p+1}^{n} [(\xi + \epsilon_i)^2 - \xi^2]$$

$$= \sum_{i=1}^{p} (2\epsilon_i + \epsilon_i^2) + \sum_{i=p+1}^{n} (2\xi\epsilon_i + \epsilon_i^2)$$

$$= \sum_{i=1}^{n} \epsilon_i^2 + 2 \sum_{i=1}^{p} \epsilon_i + 2\xi \sum_{i=p+1}^{n} \epsilon_i \cdot$$

If we now substitute

$$\sum_{i=p+1}^{n} \epsilon_i = - \sum_{i=1}^{p} \epsilon_i$$

it follows that

$$f = \sum_{i=1}^{n} \epsilon_i^2 + 2(1 - \xi) \sum_{i=1}^{p} \epsilon_i . \tag{A13}$$

Assume now that $\epsilon_p > 0$. Then from the ordering hypothesis it follows that for $i = 1, 2, \ldots, p$ we have $\epsilon_i > 0$ whence

$$\sum_{i=1}^{p} \epsilon_i > 0 .$$

But from the corollary it follows that $1 - \xi \geq 0$ so that the second term of Equation A13 is $\geq 0$. Clearly the first term of Equation A13 is positive since each term in the summation is $\geq 0$ and the first p terms are actually $> 0$. Thus we have $f > 0$.

But this implies that

$$\sum_{i=1}^{n} c_i^2 > \sum_{i=1}^{n} q_i^2 = \frac{rA^2}{n} ,$$

which is a contradiction. Therefore the assumption that $\epsilon_p > 0$ must be false, that is $\epsilon_p \leq 0$ which implies:

$$c_p \geq q_p .$$

## Theorem 6

Let the real numbers $c_1, c_2, \ldots, c_n$ represent a histogram with area A and sum of squares $rA^2/n$. Order these numbers so that $c_1 \geq c_2 \geq \ldots \geq c_n \geq 0$. Then if $1 \leq r \leq n/(p-1)$ a histogram that has the smallest possible value for $c_p$ (p between 2 and n) is of the form of the lemma with $i = p - 1$ and $j = n - p + 1$. The corresponding value for $c_p$ under these conditions is

$$\frac{A}{n}\left(1 - \sqrt{\frac{(p-1)(r-1)}{n-p+1}}\right) .$$

*Proof:* Let $1 \leq r \leq n/(p-1)$ and let $i = p - 1$ and $j = n - p + 1$ in Equations A3 and A4. We have thereby obtained the histogram:

$$\left.\begin{aligned} c_1 = c_2 = \ldots = c_{p-1} &= \frac{A}{n}\left(1 + \sqrt{\frac{(n-p+1)(r-1)}{p-1}}\right) , \\ c_p = c_{p+1} = \ldots = c_n &= \frac{A}{n}\left(1 - \sqrt{\frac{(p-1)(r-1)}{n-p+1}}\right) , \end{aligned}\right\} \qquad \text{(A14)}$$

which indeed satisfies the hypothesis of the theorem whenever $1 \leqq r \leqq n/(p-1)$. If we note here that this histogram is the same as that of *Case II* of Theorem 5 with $p - 1$ substituted for $p$ we may proceed similarly to obtain:

$$f = \sum_{i=1}^{n} \epsilon_i^2 + 2 \sum_{i=1}^{p-1} \epsilon_i + 2\xi \sum_{i=p}^{n} \epsilon_i ,$$

where here $\xi = q_p/q_1$. Since

$$\sum_{i=1}^{n} \epsilon_i = 0$$

we can substitute

$$\sum_{i=p}^{n} \epsilon_i = \sum_{i=1}^{p-1} \epsilon_i .$$

$$f = \sum_{i=1}^{n} \epsilon_i^2 - 2(1-\xi) \sum_{i=p}^{n} \epsilon_i . \tag{A15}$$

The first term is clearly non-negative. Assume $\epsilon_p < 0$. Then from the ordering hypothesis it follows that $0 > \epsilon_p \geqq \epsilon_{p+1} \geqq \ldots \geqq \epsilon_n$ whence

$$\sum_{i=p}^{n} \epsilon_i < 0.$$

From the corollary, it follows that $1 - \xi \geqq 0$ so that the second term of Equation A15 is $\geqq 0$. The first term of Equation A15 is clearly $> 0$ since each term of the summation is at least $\geqq 0$ and those terms corresponding to $i = p, p + 1, \ldots,$ and $n$ are $> 0$.

Thus $f > 0$. But this implies that

$$\sum_{i=1}^{n} c_i^2 > \sum_{i=1}^{n} q_i^2 = \frac{rA^2}{n} ,$$

which is a contradiction. Thus the assumption that $\epsilon_p < 0$ is false; that is, $\epsilon_p \geqq 0$. This implies that $c_p \geqq q_p$ .

*"The aeronautical and space activities of the United States shall be conducted so as to contribute . . . to the expansion of human knowledge of phenomena in the atmosphere and space. The Administration shall provide for the widest practicable and appropriate dissemination of information concerning its activities and the results thereof."*

—NATIONAL AERONAUTICS AND SPACE ACT OF 1958

# NASA SCIENTIFIC AND TECHNICAL PUBLICATIONS

TECHNICAL REPORTS: Scientific and technical information considered important, complete, and a lasting contribution to existing knowledge.

TECHNICAL NOTES: Information less broad in scope but nevertheless of importance as a contribution to existing knowledge.

TECHNICAL MEMORANDUMS: Information receiving limited distribution because of preliminary data, security classification, or other reasons.

CONTRACTOR REPORTS: Technical information generated in connection with a NASA contract or grant and released under NASA auspices.

TECHNICAL TRANSLATIONS: Information published in a foreign language considered to merit NASA distribution in English.

TECHNICAL REPRINTS: Information derived from NASA activities and initially published in the form of journal articles.

SPECIAL PUBLICATIONS: Information derived from or of value to NASA activities but not necessarily reporting the results of individual NASA-programmed scientific efforts. Publications include conference proceedings, monographs, data compilations, handbooks, sourcebooks, and special bibliographies.

***Details on the availability of these publications may be obtained from:***

SCIENTIFIC AND TECHNICAL INFORMATION DIVISION

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

**Washington, D.C. 20546**