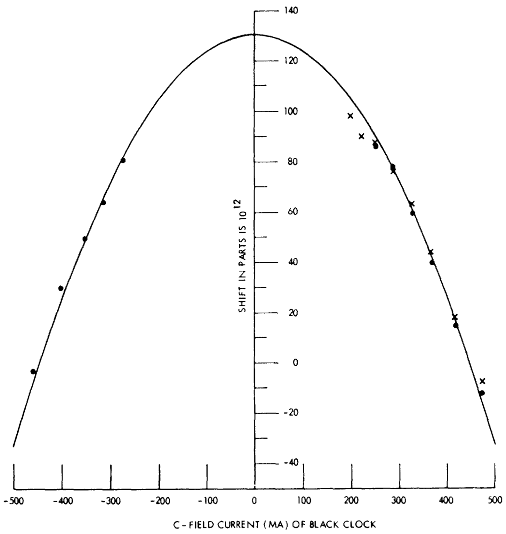

Fig. 9. *C*-field effect on operating frequency.

the zero current point can be attributed to a permanent field in the drift region of the tube and can be evaluated from (8). By operating at a known point on the parabola and varying externally applied fields along the two orthogonal axis, the total field present in the drift region could be determined by a similar technique.

Conclusion

Square-wave phase modulation has been shown to be a superior technique for the precision control of cesium atomic clocks since it causes the operating frequency to be that of the true magnetically-offset cesium resonance. Even when used for conventional single-loop operation its simplicity and the ease with which it meets specified tolerances makes it the logical modulation choice. The method of determining precisely the magnetic offset by observing the effect of magnetic field on frequency may resolve the remaining barrier to true accuracy.

The Statistical Construction of a Single Standard from Several Available Standards

EDWIN L. CROW

***Abstract*—In the case of physical quantities for which no international standards exist, such as microwave power and microwave noise, it may be desirable to establish an international standard by combination of the data from several laboratory standards. Even if there is an international standard, it conceivably might be replaced by a new composite standard, for example the mean of several laboratory atomic-time standards. A weighted mean may be preferable.**

It is noted that the "best" combination of observations for estimating a theoretical mean or median value depends on the theoretical frequency distribution from which the observations are drawn. Since the theoretical distribution for observations from different laboratories would in general be unknown, two weighted means are suggested which are reasonably good for many distributions, allowing for wild observations particularly, and in which the weight of each observation depends only on its order when the observations are ordered in size. The efficiencies of these and other estimates are evaluated for each of four distributions.

I. Introduction

A PHYSICAL STANDARD may be defined in terms of a prototype object, such as the meter bar in the vault of the International Bureau of Weights and Measures, or in terms of a specification for determining a conceptual physical quantity, such as the present definition of the meter as a certain number of wavelengths of the orange-red radiation of krypton 86.

In the former case the standard, although uniquely defined, may vary in time due to environmental influences that are not quite constant despite all efforts to make them so. In the latter case the standard will in effect depend on the particular experimental instrumentation through which the standard is made operational. Thus, in either case a standard might be established more accurately by averaging in some way over several independent prototype objects or over several independent instrumentations, with positive and negative errors tending to cancel each other.

This conjecture can be put on a more rigorous basis by the use of results from probability theory. In particular, it is well known that if the objects or instrumentations give rise to independent observations from a common probability distribution possessing a standard deviation σ, then the mean of n such observations has a standard deviation $\sigma/\sqrt{n}$. Thus, if the goodness of a standard is judged by the standard deviation of realizations of the standard, the combination standard using the mean of n independent instrumentations is better than any single instrumentation.

But is the mean the best combination of the observations for estimating the true value of the quantity under

Manuscript received July 13, 1964.
The author is with the National Bureau of Standards, Boulder, Colo.

observation? No, in general, although it is if the common probability distribution of the observations is normal, i.e., Gaussian. It is the purpose of this paper to advocate the use of composite standards, to show quantitatively how far short of the optimum combination the mean falls for several distributions, and to recommend a weighted mean as a satisfactory alternative. This is not to say that a simple mean (with equal weights) will not be satisfactory in many cases, but a properly weighted mean has the advantage of being less affected by outlying observations.

A composite standard defined as a weighted mean could be quite generally used but may be particularly useful where no international standards now exist, for example for microwave power. Even if there is an existing single international standard, it conceivably might be replaced by a composite standard based on a new principle; for example, the ephemeris second might be replaced by the weighted mean of several laboratory atomic-time standards. Markowitz has been forming such a standard experimentally at the U. S. Naval Observatory [1].

How should weights for combining the data of several laboratories be determined? Often they have been determined from the number and internal variability of the observations within each laboratory which are averaged to provide the contributed values. However, it is well known that systematic external errors often are far more important in physical experiments than the observed internal fluctuations [2]. Hence, internally determined weights are not usually satisfactory by themselves. Another approach is the use of the external consistency of the several laboratory values, but the number of such values is often too small to contain sufficient information for weighting.

An extreme weight, zero, is attained if a value is rejected completely as being too outlying. Statistical methods for rejection of outlying values may be invoked and lead to some surprising conclusions. For example, even if one is sampling from a normal distribution, if he has only three observations, the longer of the two nonoverlapping intervals formed will be 16 or more times as long as the shorter interval in ten per cent of such samples in the long run and 32 or more times as long in five per cent of such samples [3]. This suggests that a person proceeding intuitively may tend to reject an observation when in fact it is a valid observation from the intended distribution. However, Dixon has analyzed processing of data for outlying values further and given rules by which one may either end up with the median of the original data or the mean of data with outlying observations removed as the better estimate, depending on the expected proportion of "contaminating" values, i.e., values from a distribution other than the intended one [4]. In contrast, the present paper seeks to avoid complete rejection of data by assigning smaller weights to the more extreme observations and the largest weight to the median.

In order to compare the behavior of various estimates of the true value of a physical quantity, we must consider the observations as random variables with a common "parent" probability distribution, of which the true value is then the mean or median. The estimates, such as the observed (or sample) mean or median, then have probability distributions also, generated by hypothetical repetition of taking samples of the specified size. By the behavior of an estimate we then mean simply its probability distribution, in particular its spread as measured by its standard deviation. It is fairly evident that the distribution of an estimate of the mean or median of the parent distribution depends on what that parent distribution is, on its shape in particular. Hence, in Section II the mathematical expressions are given for the four parent probability distributions to be considered in detail.

The particular estimates to be considered, all weighted means, are listed in Section III. Section IV and the accompanying Figs. 2–4 give the quantitative results of this paper, the *percentage efficiencies* of the various estimates of the parent mean or median. The results are applicable not only to the construction of composite standards but to data analysis in general.

II. Distributions of Observations

It is assumed that each observation is a real number drawn independently from the same probability distribution, called the parent distribution. It is assumed that the parent distribution is symmetrical in any case. Thus, although we may not in practice know the shape of the distribution, we are assuming that negative errors of any given size are just as likely as positive errors of the same size. It follows that a parent mean is identical with the parent median.

A parent distribution will be defined in terms of its probability density function (pdf) $f(X)$. Since we shall always be comparing estimates of the parent mean for the same parent distribution, it turns out that the parent distributions may be expressed in the simplest possible form, in particular centered at $X=0$.

Normal Distribution:

$$f(X) = \frac{1}{\sqrt{2\pi}} e^{-X^2/2}, \qquad -\infty < x < \infty$$

Rectangular Distribution:

$$f(X) = \begin{cases} 1, & |X| \le \frac{1}{2} \\ 0, & |X| > \frac{1}{2} \end{cases}$$

Double-Exponential Distribution:

$$f(X) = \tfrac{1}{2} e^{-|X|}, \qquad -\infty < X < \infty$$

Cauchy Distribution:

$$f(X) = \frac{a}{\pi} \frac{1}{a^2 + X^2}, \qquad -\infty < X < \infty$$

For all essential purposes we may set the constant a in the Cauchy distribution equal to 1, but in Fig. 1 the Cauchy pdf is shown with $a=0.6745$ along with the above normal pdf. This value of a produces the same central 50 per cent interval as the normal pdf shown.

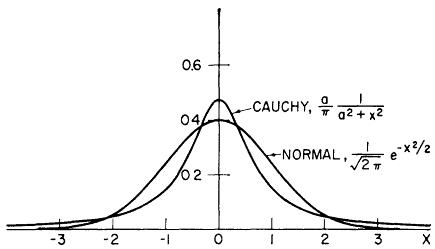

Fig. 1. Probability density functions of normal and Cauchy distributions, the latter drawn for $a=0.6745$.

The mean of each of the above distributions is zero except for the Cauchy, for which the mean, according to the definition as an integral, does not exist. However, the median exists for any distribution and, for symmetrical pdf's, equals the mean when the mean does exist; for this reason we consider in general the estimation of the median. The standard deviation (root-mean-square deviation from the mean) exists for the first three distributions also but not for the Cauchy. This nonexistence results from the very slowly decreasing tails of the Cauchy pdf. The double-exponential pdf has tails decreasing more slowly than the normal tails but much more rapidly than the Cauchy tails.

The four distributions above can be considered simply as arbitrary mathematical assumptions. However, there are good reasons for their selection. The normal distribution is often a good approximation to distributions of physical measurements. The lack of tails in the rectangular distribution provides an effect on estimates of the mean in great contrast to that of thick-tailed distributions, and its simplicity enables complete analytical results to be obtained. The double-exponential and the Cauchy may be considered as rough models for taking observations from a normal distribution subject to contamination by outlying observations from some other distribution when there is no information to suggest whether the outliers are too large or too small, i.e., when the composite distribution including contamination is symmetrical and has thicker tails than the normal distribution does [5].[1] More realistic models of contamination have been considered by Dixon [4] and Tukey [6] providing explicitly for contamination from another distribution, but such models appear to be susceptible only to Monte Carlo or asymptotic investigation.

[1] Tukey and McLaughlin [5] state on p. 332:

"The typical distribution of errors and fluctuations has a shape whose tails are longer than that of a Gaussian distribution . . . . It is the Gaussian distribution that has to be regarded as somewhat pathological from the standpoint of practice. And distributions with shorter tails, while they do occur, are rather more pathological. Thus, frequency of occurrence directs our attention to longer-tailed distributions."

III. Some Estimates of the Mean or Median

We assume that N independent observations from a distribution having a symmetric pdf are available. We consider the observations in order of size and so label them:

$$X_1 \leq X_2 \leq \cdots \leq X_i \leq X_{i+1} \leq \cdots \leq X_N$$

Since optimum estimates of the mean are often linear in the observations and since nonlinear estimates would be difficult to treat, we restrict attention to estimates that are linear combinations of the ordered observations with symmetric coefficients, that is, to the weighted means

$$\sum_{i=1}^{N} a_i X_i \Big/ \sum_{i=1}^{N} a_i, \qquad a_{N-i+1} = a_i,\ i = 1, 2, \cdots, N \tag{1}$$

We shall consider the following weighted means in which the weights are prescribed:

Mean:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{2}$$

Median:

$$\begin{aligned}\tilde{X} &= X_n && \text{if } N = 2n-1\\ &= \tfrac{1}{2}(X_n + X_{n+1}) && \text{if } N = 2n\end{aligned} \tag{3}$$

Midrange:

$$X_{MR} = \tfrac{1}{2}(X_1 + X_N) \tag{4}$$

Linearly weighted mean:

$$\begin{aligned}\tilde{\mu} &= \frac{1}{n^2}\left[\sum_{i=1}^{n-1} i(X_i + X_{2n-i}) + nX_n\right] \quad \text{if } N = 2n-1\\ &= \frac{1}{n(n+1)}\sum_{i=1}^{n} i(X_i + X_{2n+1-i}) \quad \text{if } N = 2n\end{aligned} \tag{5}$$

Square-weighted mean:

$$\begin{aligned}\tilde{\mu}' &= \frac{3}{n(2n^2+1)}\left[\sum_{i=1}^{n-1} i^2(X_i + X_{2n-i}) + n^2X_n\right] \quad \text{if } N = 2n-1\\ &= \frac{3}{n(n+1)(2n+1)}\sum_{i=1}^{n} i^2(X_i + X_{2n+1-i}) \quad \text{if } N = 2n\end{aligned} \tag{6}$$

These weighted means have been listed in the order of their frequency of use, the last two apparently being new. In the order of decreasing weights for the more extreme observations they are:

$$X_{MR},\ \overline{X},\ \tilde{\mu},\ \tilde{\mu}',\ \tilde{X}$$

Thus the midrange would be most affected by contaminating outlying observations, and the median least.

On the other hand, there has been much investigation of determining weights from the sample itself. The rejection of observations is equivalent to assigning zero weight [4]. Tukey has advocated replacing outlying observations by their nearest neighbors ("Winsorizing") [7]. Earlier Jeffreys had given an iterative method for determining continuously varying weights from the sample [8]. Recently Huber published a rigorous foundation for an iterative method which is asymptotically (for large sample size) equivalent to Winsorizing [9]. However, determining weights from the sample itself usually requires a substantial number of observations, some assumptions, and a substantial amount of work.

One particular method of determining weights from the data may be noted: First calculate $\overline{X}$; weight each observation (or individual standard) inversely as its absolute deviation from $\overline{X}$, and calculate the weighted mean; repeat, using the deviation from the weighted mean, and so on until there is no change in the weighted mean. Joan Rosenblatt has shown that this sequence converges to the median unless the mean or some iterate coincides exactly with an individual observation [10]. Hence one should simply use the median to begin with. (In the exceptional case of coincidence the sequence stops at the coincident observation.)

IV. Efficiencies of Estimates

Since the parent distributions and estimates of the median considered are both symmetrical, the estimates are unbiased, and the goodness of estimates may be judged by their standard deviations. The *efficiency* of an estimate is defined as the square of the ratio of the minimum standard deviation possible with any weighted mean to the standard deviation for the estimate. An efficiency of 0.64, or 64 per cent (as in the case of the sample median and a normal distribution) means that the best possible estimate based on 64 observations is as good as the estimate considered based on 100 observations. The percentage efficiencies of the estimates (2)–(6) were either given by Sarhan and Greenberg [11] or can be calculated from their results.

Normal Distribution

The percentage efficiencies of the median $\tilde{X}$, linearly weighted mean $\tilde{\mu}$, and square-weighted mean $\tilde{\mu}'$ (relative to the mean $\overline{X}$) for observations from a normal distribution are shown as a function of sample size in Fig. 2. The midrange is less efficient than the median for $N \geq 6$ and is not shown. The rapid approach of the median's efficiency to its asymptotic value of 63.7 per cent is shown. The most interesting feature of Fig. 2 is the fairly high efficiencies of $\tilde{\mu}$ and $\tilde{\mu}'$ for all sample sizes. Only the indicated points were calculated, but the excellent fit by rectangular hyperbolas suggests that the fitted asymptotic efficiencies of 85 per cent for $\tilde{\mu}'$ and 92 per cent for $\tilde{\mu}$ are essentially correct.[2]

[2] Added in proof: These values have been confirmed by asymptotic theory developed by M. M. Siddiqui which is part of a joint paper with the present author. The paper has been submitted to a statistical journal.

Rectangular Distribution

For observations from a rectangular distribution the best linear unbiased estimate is the midrange [11]. Its variance is of order $1/N^2$, whereas all the other estimates (2), (3), (5), and (6) have variances of order $1/N$ and hence their efficiencies approach zero as N becomes infinite. The numerical results are given in Fig. 3. The sample mean is more efficient than the median, $\tilde{\mu}'$, and $\tilde{\mu}$, but none of these four is satisfactory for $N > 4$.

Double-Exponential Distribution

Efficiencies for the double-exponential distribution for $N \leq 5$, based on the work in Sarhan and Greenberg [11] and Sarhan [12], are graphed in Fig. 4. Fig. 4 shows that $\tilde{\mu}'$ has efficiency at least 98.6 per cent for $N \leq 5$, that the median and $\tilde{\mu}$ are quite efficient also, and that $\overline{X}$ rapidly decreases in efficiency, down to 79 per cent for $N=5$. The midrange is less efficient than $\overline{X}$ and is not considered in Fig. 4.

Cauchy Distribution

Corresponding results for the Cauchy distribution cannot be shown because the standard deviations of the estimates (2)–(6) may be infinite except for that of the median. In fact the distribution of $\overline{X}$ for samples from a Cauchy distribution is exactly the same Cauchy distribution and thus has neither mean nor standard deviation. Even the standard deviation of the sample median is infinite for $N=1$, 2, 3, and perhaps $N=4$ [13]. The best linear unbiased estimate of the parent median is not known, but an estimate about nine per cent more efficient than the sample median is obtained for large samples by averaging the middle quarter of the ordered observations, and no substantial improvement is possible [14].

An Even Thicker-Tailed Distribution

It might be thought that the Cauchy distribution is a rather pathological example, not experienced in practice. However, from a finite number of observations on an unknown distribution it is impossible to estimate the extreme tails of the pdf and thus to be sure how thick the tails are. Furthermore, the Cauchy distribution is but a middling case of a family of distributions, the stable distributions, depending on a constant α between 0 and 2 (as well as another constant β not relevant here). The normal distribution corresponds to $\alpha=2$, the Cauchy to $\alpha=1$. The pdf of another stable distribution, for $\alpha=1/2$, is given by [15]

$$f(X) = (2\pi X^3)^{-1/2} e^{-1/(2X)}, \quad x \geq 0 \tag{7}$$

The pdf of the means of samples of two observations from the distribution is

$$g(\overline{X}) = (\pi \overline{X}^3)^{-1/2} e^{-1/\bar{X}}, \quad x \geq 0 \tag{8}$$

Thus $g(\overline{X})$ is the same curve as $f(X)$ plotted to a different scale; in fact each percentage point of $\overline{X}$, such as

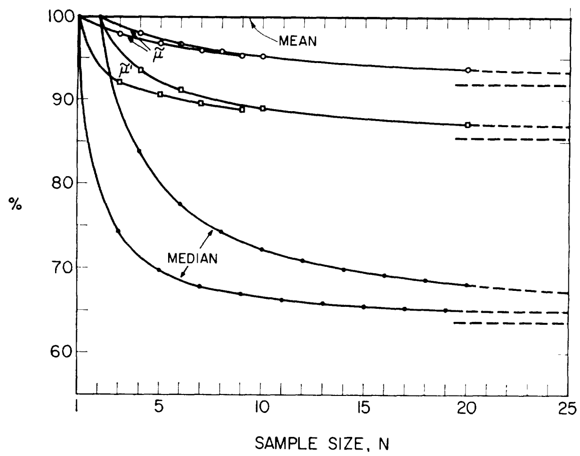

Fig. 2. Percentage efficiencies of estimates of μ for samples from a normal distribution.

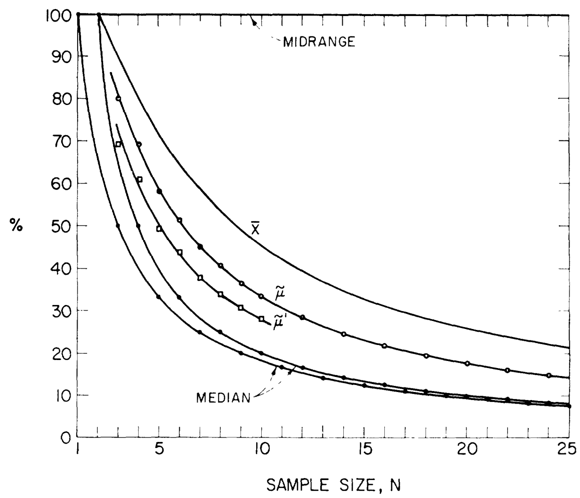

Fig. 3. Percentage efficiencies of estimates of μ for samples from a rectangular distribution.

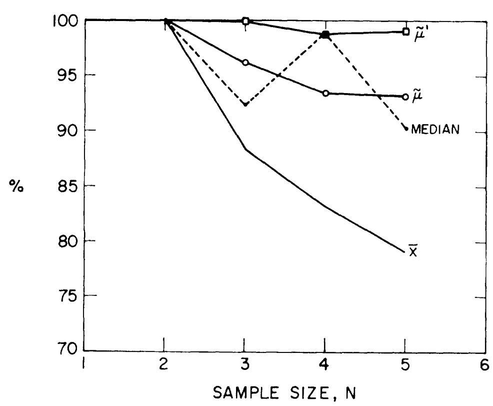

Fig. 4. Percentage efficiencies of estimates of μ for samples from a double-exponential distribution.

its median, is twice the corresponding percentage point of X. Thus $\overline{X}$ is twice as dispersed as X, in complete contrast to the usual situation where $\overline{X}$ for $N=2$ is $1/\sqrt{2}$ as dispersed as X. (The usual situation is that in which the parent distribution possesses a finite standard deviation and hence the Central Limit Theorem holds.) This remarkable fact can be easily checked because (7) is the distribution of the reciprocal of the square of a standard normal variable; hence one need only take, say, 200 random normal numbers [16], square and invert, calculate 100 values of $\overline{X}$ from successive pairs, and form two histograms. For this distribution (which is not symmetrical) we conclude simply that $\overline{X}$ is not as good an estimate as any single observation and that the median has good estimating properties based on general theory [17].

It should be noted that if all observations on a distribution beyond a *fixed* finite value of X are rejected, then the standard deviation of the resulting distribution is finite, the Central Limit Theorem applies, and the above type of result is excluded.

V. Conclusion

The results of Section IV and Figs. 2–4 are examples of a general law for the variation of efficiency of an estimate of the median of a symmetrical distribution with the weights of the coefficients and the thickness of the tails of the distribution [11]. For short-tailed distributions like the rectangular it is most efficient to weight the extreme observations heavily and the middle ones zero (or even negative for U-shaped distributions); for the normal distribution, to weight all observations equally; and for thick-tailed distributions like the double-exponential and Cauchy, to weight the more extreme observations zero and the middle ones heavily. While the sample mean $\overline{X}$ thus loses its popular position as the best of all estimates of parent mean or median, it is a simple estimate and reasonably good over a wide range of distributions.

In practice the experimenter or arbiter of standards may not know what form of distribution he is observing, or if he does, he cannot guard completely against wild observations or against the gradual drift of one or more standards from the norm set by the group as a whole. Distributions of errors in the physical sciences very often are close to normal but may have wild observations, so that thick-tailed distributions seem a reasonable model while short-tailed distributions are unrealistic (unless observations have been rejected, perhaps too readily). Hence it is reasonable to recommend tapered estimates like $\tilde{\mu}$ or $\tilde{\mu}'$, (5) or (6), if there is apt to be contamination of observations not immediately or obviously noticeable. It seems particularly appropriate in constructing a composite standard, in which one might be tempted to weight a highly-regarded individual standard heavily but find that in time this individual standard is in the outskirts of the group; the impartial weights based on order, as in $\tilde{\mu}$ and $\tilde{\mu}'$, prevent the outlying observation from affecting the composite standard greatly but yield good efficiency if the process is after all abiding by the normal distribution.

Acknowledgment

The author is indebted to J. M. Richardson for asking the question which led to this study and for several suggestions regarding applications.

References

[1] Markowitz, W., The atomic time scale, *IRE Trans. on Instrumentation*, vol I-11, Dec 1962, pp 239–242.
[2] McNish, A. G., The speed of light, *IRE Trans. on Instrumenta-*

tion, vol I-11, Dec 1962, pp 138–148.
[3] Dixon, W. J., Ratios involving extreme values, *Ann. Math. Statist.*, vol 22, Mar 1951, pp 68–78.
[4] Dixon, W. J., Processing data for outliers, *Biometrics*, vol 9, Mar 1953, pp 74–89.
[5] Tukey, J. W., and D. H. McLaughlin, Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization 1, *Sankhyā*, Series A, vol 25, Sep 1963, pp 331–352.
[6] Tukey, J. W., A survey of sampling from contaminated distributions, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, Stanford, Calif.: Stanford University Press, 1960, pp 448–485.
[7] Tukey, J. W., The future of data analysis, *Ann. Math. Statist.*, vol 33, Mar 1962, pp 1–67.
[8] Jeffreys, H., *Theory of Probability*, 2nd ed., New York: Oxford, 1948, pp 187–192.
[9] Huber, P. J., Robust estimation of a location parameter, *Ann. Math. Statist.*, vol 35, Mar 1964, pp 73–101.
[10] Rosenblatt, J. R., National Bureau of Standards, Washington, D. C., Oct 1962. Private communication.
[11] Sarhan, A. E., and B. G. Greenberg, *Contributions to Order Statistics*, New York: Wiley, 1962, pp 383–397, and 200–205.
[12] Sarhan, A. E., Estimation of the mean and standard deviation by order statistics, *Ann. Math. Statist.*, vol 25, Jun 1954, pp 317–328.
[13] Rider, P. R., Variance of the median of samples from a Cauchy distribution, *J. Amer. Statist. Assn.*, vol 55, Jun 1960, pp 322–323.
[14] Rothenberg, T. J., F. M. Fisher, and C. B. Tilanus, A note on estimation from a Cauchy sample, *J. Amer. Statist. Assn.*, vol 59, Jun 1964, pp 460–463.
[15] Lukacs, E., *Characteristic Functions*, New York: Hafner, 1960, p 107.
[16] RAND Corporation, *A Million Random Digits with 100,000 Normal Deviates*, Glencoe, Ill.: Free Press, 1955.
[17] Cramér, H., *Mathematical Methods of Statistics*, Princeton, N. J.: Princeton University Press, 1946, p 483.

Exchange Collisions, Wall Interactions, and Resettability of the Hydrogen Maser

J. VANIER, H. E. PETERS, MEMBER, IEEE, AND R. F. C. VESSOT

***Abstract*—Experiments were performed in order to verify the resettability of the hydrogen maser. The method consisted of measuring the output frequency of one maser against the hydrogen pressure. It was found that at a given tuning of the cavity no shift larger than 2.1 parts in 10^{13} was observed for a change of 4 to 1 in pressure. This experiment also showed that the pressure shift due to exchange collisions, predicted by Bender, could not be observed for the field-independent transition in the hydrogen maser. Two masers, having the same storage bulb design and the same wall coating, were tuned by this technique and were found to have a frequency difference of 7.6 parts in 10^{13}. Experiments on the wall coating of the hydrogen maser storage bulb were made. Relaxation and decorrelation times of various materials were measured. The hyperfine splitting of the ground state of hydrogen measured against cesium is also given.**

I. Introduction

A FEATURE of the storage box technique [1], [2] used in the hydrogen maser is the presence of exchange collisions which depend on the hydrogen pressure. The resultant broadening of the emission line provides a means of tuning the cavity of the maser and making resettability measurements [10].

The storage box technique, however, relies on the use of a wall coating which does no appreciably perturb the energy state of the hydrogen atoms. The basic requirements are that the substance used as a coating must have low adsorption energy, must not react chemically with atomic hydrogen, and must not encourage H-H recombination. In the first type of interaction the atom may either make a transition to another level of the ground state (T_1 type of relaxation) or experience a phase shift which is reflected by an average frequency shift of the maser output frequency. Dispersion in adsorption times causes a loss of coherence which may be associated with a T_2 type of relaxation, where T_1 and T_2 relaxation times are used in analogy to nuclear magnetic relaxation. In the two other processes the atom as such is lost, which is essentially a T_1 type of relaxation. Various substances have been found in the past which meet the basic requirements to some degree; these are dichlorodimethylsilane, teflon, and long chain waxes [2], [4], [5].

The research reported here was conducted to investigate the properties of these materials and to search for new materials that could be used as wall coating. In this paper, resettability measurements made on two masers with the exchange collision tuning technique are reported. Results of the investigation of the properties of various wall coatings are given in Section IV. Finally, with the help of these results the free-space hyperfine splitting of the hydrogen ground state was determined, and the result is reported in Section V.

II. Apparatus

Four hydrogen masers were used in the experiments described here. Two of these masers incorporated a cavity thermal control; the design of these instruments has been described in another article [6]. Their relative stability measured for a period of several months is shown in Fig. 1. The other two masers used in the present research did not have a thermally controlled cavity and consequently did not have this degree of long-term stability.

Manuscript received June 23, 1964.
The authors are with Varian Associates, Beverly, Mass.