General Disclaimer

One or more of the Following Statements may affect this Document

- This document has been reproduced from the best copy furnished by the organizational source. It is being released in the interest of making available as much information as possible.

- This document may contain data, which exceeds the sheet parameters. It was furnished in this condition by the organizational source and is the best copy available.

- This document may contain tone-on-tone or color graphs, charts and/or pictures, which have been reproduced in black and white.

- This document is paginated as submitted by the original source.

- Portions of this document are not fully legible due to the historical nature of some of the material. However, it is the best reproduction available from the original submission.

(NASA-CR-160197) STATISTICAL CONSIDERATIONS IN DESIGN OF SPACELAB EXPERIMENTS (General Electric Co.) 41 p HC A03/MF A01 CSCL 22A

N79-25047

Unclas
G3/12 22186

STUDY REPORT:

STATISTICAL CONSIDERATIONS IN DESIGN OF SPACELAB EXPERIMENTS

Prepared by:

John Robinson
General Electric Company
Information Systems Programs
Space Division
Arlington, Virginia 22202

October 16, 1978

1.0 PRINCIPAL CONSIDERATIONS IN EXPERIMENTAL DESIGN AND DATA ANALYSIS

The purpose of this section is to present the main steps taken in setting up an experiment to furnish data on a hypothesis and then analyzing these data in order to obtain information leading to acceptance or rejection of the hypothesis.

Figure 1 shows the main quantitative factors which affect the result of a statistical hypothesis test on the data furnished by an experiment. These are:

1. Measurement error
2. Subject-to-subject variation
3. Day-to-day variation
4. Sample size (number of subjects)
5. Number of measurements on a subject
6. Number of measurements taken over a period of days.

Figure 1 shows schematically the effect of these factors on the outcome of the experiment and of the post-experiment analysis. That is, after an experiment is performed, a statistical analysis is generally carried out to test one or more hypotheses. The results of this analysis are, for each such hypothesis, (1) a decision to accept or reject the hypothesis, and (2) a numeric "confidence" in the correctness of this acceptance or rejection. This confidence is expressed by two sets of parameters: the significance level of the hypothesis, and confidence intervals about the parameters used in the statement of the hypothesis. The significance level is the probability of rejecting the hypothesis when it is true.

The measurement error is usually normally distributed about a mean, which is ideally equal to zero. This means that individual errors of measurement vary randomly, sometimes above the mean and sometimes below. The average of a sequence of measurements will tend to be closer to the

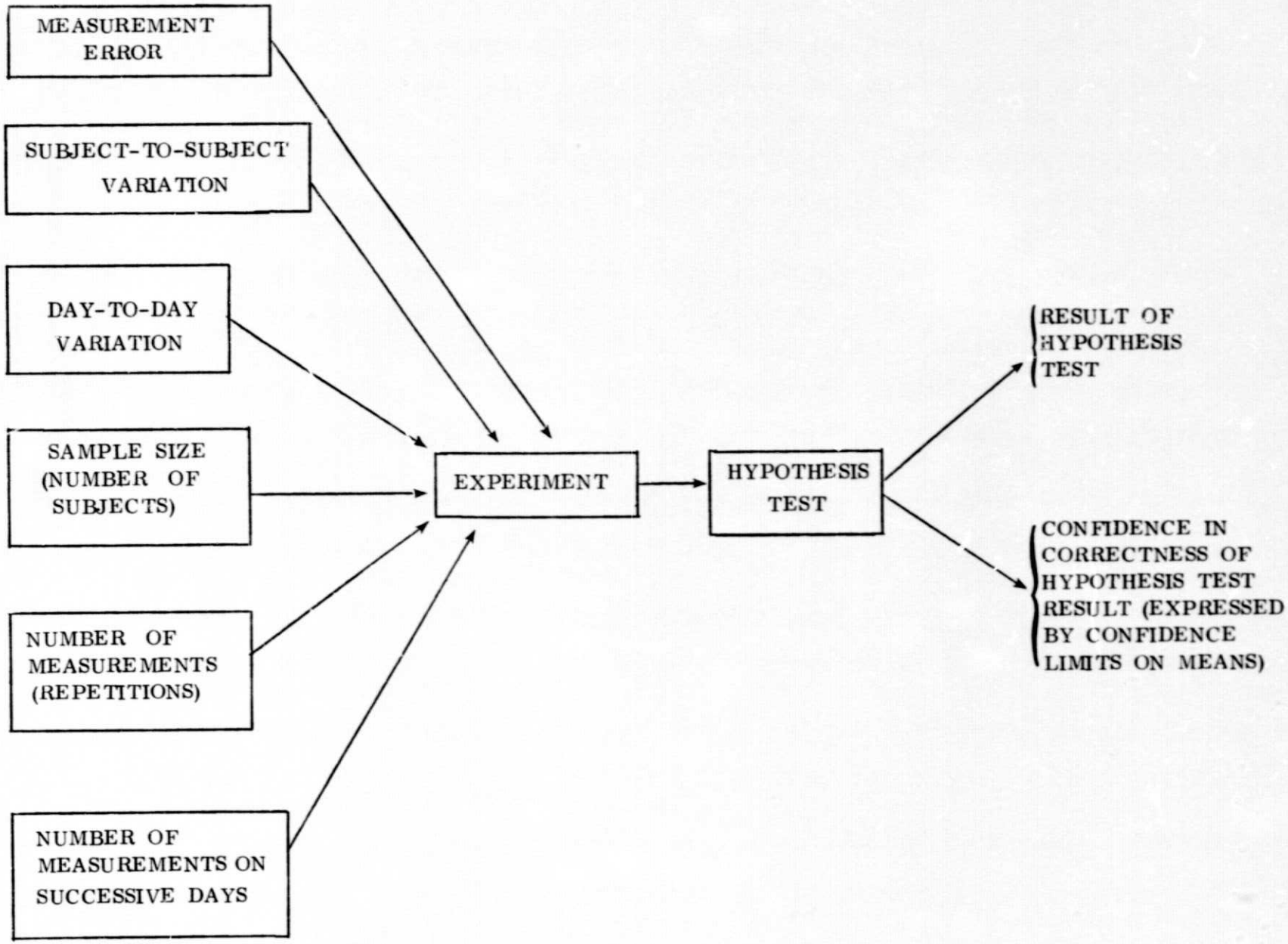

MEASUREMENT ERROR
SUBJECT-TO-SUBJECT VARIATION
DAY-TO-DAY VARIATION
SAMPLE SIZE (NUMBER OF SUBJECTS)
NUMBER OF MEASUREMENTS (REPETITIONS)
NUMBER OF MEASUREMENTS ON SUCCESSIVE DAYS
EXPERIMENT
HYPOTHESIS TEST
RESULT OF HYPOTHESIS TEST
CONFIDENCE IN CORRECTNESS OF HYPOTHESIS TEST RESULT (EXPRESSED BY CONFIDENCE LIMITS ON MEANS)

FIGURE 1

mean, as the number of such measurements increases. Thus, if successive measurements on a single subject are statistically independent of each other and normally distributed with mean of zero, and if it is possible to make a series of such measurements, the probability that the average of all such measurements has error quite close to zero is higher than for the case of only one such measurement.

Following is a table of such probabilities for the case of a measuring device whose errors are normally distributed with mean of zero and standard deviation (denoted by the symbol σ) equal to 1:

No. of Measurements	$P(\lvert \text{error}_n \rvert < 1)$	$P(\lvert \text{error}_n \rvert < 0.1)$
1	0.6826	0.0796
2	0.8414	0.1114
3	0.9164	0.1350
4	0.9544	0.1586
5	0.9742	0.1742
10	0.9984	0.2510
30	>0.9999	0.4176
100	>0.9999	0.6826
200	>0.9999	0.8414
400	>0.9999	0.9544
1000	>0.9999	>0.9999

Here the term error_n is defined as follows: If $e_1, e_2, \ldots, e_k$ are the errors of the first, second,..., kth measurements, respectively, then $\text{error}_1 = e_1$; $\text{error}_2 = (e_1 + e_2)/2$, ..., $\text{error}_n = (e_1 + e_2 + \ldots + e_n)/n$.

The second quantitative factor mentioned above, viz. subject-to-subject variation, is the natural variation between subjects of any quantity.

For instance, some subjects may lose more bone than others, or more body water than others, under the influence of bed rest or space flight. It seems reasonable to assume that this variation is an expression of the variance of a normally distributed random variable. For example, the loss of trabecular bone sustained by male subjects of a certain age and physical condition after three weeks of space flight may average 16% with a standard deviation of 3%, with the losses normally distributed about the mean. The same type of loss for female subjects is likely also normally distributed, quite possibly with a different mean, but with much the same standard deviation.

More generally, any statistical parameter, such as average total body water (TBW) loss or bone loss, may be <u>estimated</u> from any sample of one or more subjects. If the numbers $X_1, \ldots, X_n$ represent TBW losses or bone losses for each of n subjects, the estimate of average TBW or bone loss for the population from which the subjects were taken is calculated as

$$\overline{X} = (X_1 + X_2 + \ldots + X_n)/n.$$

The experimenter may naturally wish to know how close to the true value of mean TBW loss, or bone loss he or she has come by making this estimate. This question is answered statistically by means of <u>confidence intervals</u>. That is, for a given number n of observations:

$$X_1, X_2, \ldots, X_n,$$

the experimenter can calculate intervals about the quantity $\overline{X}$ which contain the true mean value (of the observed quantity) with any known probability. Specifically, given n independent observations $X_1, X_2, \ldots, X_n$ of a normally distributed random variable X, a $100\,(1-\alpha)\,\%$ confidence interval about $\overline{X} = (X_1 + X_2 + \ldots + X_n)/n$ is

$$\left(\bar{X} - \frac{S}{\sqrt{n}}\, t_{n-1}\left(1 - \frac{\alpha}{2}\right),\ \bar{X} + \frac{S}{\sqrt{n}}\, t_{n-1}\left(1 - \frac{\alpha}{2}\right)\ .\right)$$

That is, the probability is $1 - \alpha$ that the true mean value (denoted by μ) for X is contained in this interval. Here the t_{n-1} values may be found in any table of the t distribution. The quantity S is defined by

$$S = \left(\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n - 1}\right)^{1/2}.$$

An example may make this clearer: If n = 10 and the ten observed values are 4.8, 5.2, 5.0, 5.5, 4.7, 4.9, 5.4, 5.1, 4.8, 4.6, then

$$\bar{X} = 5.0 \text{ and } S = 0.298.$$

Then if we want the 95% confidence interval for μ, we set $\alpha = 0.05$ and look up $t_9(0.975)$; this value is 2.262. Thus, we have

$$\frac{S}{\sqrt{n}}\, t_{n-1}\left(1 - \frac{\alpha}{2}\right) = \frac{0.298}{3.162}\,(2.262) = 0.2132.$$

Therefore, we have 95% confidence that the true value of the mean μ of the random variable X is in the interval

$$(5 - 0.2132,\ 5 + 0.2132), \text{ or } (4.7868,\ 5.2132).$$

In other words, the probability that the mean of X is in this interval is 0.95.

If, however, we had had only <u>five</u> observed values, say 4.8, 5.2, 5.0, 5.3, 4.7, we would have $\bar{X} = 5.0$ as before, and S = 0.255. This time, since $t_4(0.975) = 2.776$, we would get

$$\frac{S}{\sqrt{n}}\, t_4\left(1-\frac{\alpha}{2}\right)=\frac{0.255}{2.236}\left(2.776\right)=0.3166,$$

and our 95% confidence interval is now (4.6834, 5.3166). So the 95% confidence interval has widened with decreased number of observations. This is true in general of confidence intervals; as the sample size gets smaller, the interval gets wider. Or, if the interval is held constant, the confidence level decreases. It is, of course, intuitively clear that this should be so; the greater the number of observations we make, the higher should be our confidence that the true mean will fall into a given interval about our estimate, and the narrower should be the interval about our estimate for a given confidence level.

The next quantitative factor is day-to-day variation. Most physiological quantities are subject to some variation from day to day (examples are blood pressure, TBW, etc.) These variations appear to be random, and thus such physiological quantities may be treated as random variables in the same way as above; confidence intervals may again be calculated for the true mean of a quantity over a period of days, if we can make the assumption that the variation is only statistical and does not indicate a time change in the mean itself.

Each of the three quantitative factors just described will have an adverse effect on the confidence the experimenter may have in the conclusions he or she may draw from analyzing the data obtained by experiment. The greater the factors (i.e., the larger the σ of the corresponding distributions) the more adverse this effect will be.

Conversely, the next three factors to be discussed, viz. sample size (number of subjects), number of measurements (repetitions), and the number of measurements taken each on successive days have a <u>favorable</u> effect on the confidence the experimenter may have in the conclusions drawn. This effect counteracts the adverse effect of the first three factors, and if the sample size and the number of repetitions and daily measurements can be raised high enough, the experimenter can achieve <u>any</u> desired level of confidence in these conclusions.

Here the concept of "confidence in conclusions drawn" is denoted on the right of Figure 1; it is usually expressed by confidence limits on means, since means are usually used in expressing statistical hypotheses.

To make this clearer, an example will be presented, using TBW loss as the subject of a statistical hypothesis. Here there are only two treatments; zero-g and 1-g, and hence a t-test is appropriate. For the present case the t statistic has the form

$$\bar{X} \Big/ \left(S \Big/ \sqrt{n} \right)$$

where n is the number of subjects, $\bar{X}$ is the sample mean of the TBW losses for n different subjects:

$$\bar{X} = \left(X_1 + X_2 + \ldots + X_n \right) / n,$$

and S is the sample standard deviation:

$$S = \left[\sum_{i=1}^{n} (X_i - \bar{X})^2 / (n - 1) \right]^{1/2} .$$

The question the experimenter desires to answer is: Is there a real loss in TBW, and if so, how much, and what confidence can I have in these conclusions, given my set of data, sample size, etc. ? (Either there <u>is</u> a real loss, or any apparent loss is really only due to random variation in the data.) This question is translated into statistical language as a hypothesis, namely:

$$\mu = (\text{ mean TBW loss due to zero-g environment}) \leq 0.$$

The hypothesis is clearly equivalent to stating that there is no loss. It is statistically tested by calculating the t statistic $\overline{X}/(S/\sqrt{n})$ and

a. rejecting the hypothesis if $\overline{X}/(S/\sqrt{n}) > t_{n-1}(1-\alpha)$

or

b. accepting the hypothesis if $\overline{X}/(S/\sqrt{n}) \leq t_{n-1}(1-\alpha)$.

As an example, suppose we have TBW loss values for three subjects of 0.6 liters, 1.1 liters and 0.5 liters. These data give

$$\overline{X} = 0.7333, \quad S = 0.3215, \quad \overline{X}/(S/\sqrt{n}) = 3.9511.$$

For $\alpha = 0.05$, we have $t_{n-1}(1-\alpha) = t_2(0.95) = 2.920$. Thus we would, in this case, reject the hypothesis at the 0.05 significance level; i.e., we would reject the assertion that average loss of TBW in the zero-g environment is zero or less. This is equivalent to concluding that there is a real loss in TBW, induced by zero-g conditions, for the general population from which we drew the subjects for the experiment.

For such a case we also have $100(1-\alpha)$ % confidence that the true mean TBW loss satisfies the following inequality:

$$\mu \geq \overline{X} - \frac{S}{\sqrt{n}} t_{n-1}(1-\alpha).$$

For $\alpha = 0.05$, this is translated into saying that

$$P\left(\mu \geq 0.7333 - (0.1856)(2.920)\right) =$$

Probability that $\mu \geq 0.1913$ is equal to 0.95. Thus, given the three measurements 0.6, 1.1, 0.5, without even knowing the true mean or standard deviation of the distribution, we can say with 95% confidence that the true mean is at least 0.1913.

More generally, given a particular true mean for a change in a parameter (e.g. bone loss or TBW loss) induced by spaceflight, and given particular values for the first three values on the left-hand side of the block diagram, the experimenter may wish to know how many measurements he must have (1) to ensure that the results of a hypothesis test will call the change statistically significant (i.e. reject the statistical hypothesis of no change, mentioned above), and (2) assure the experimenter of a particular level of confidence that the true mean is greater than a given value. The answer to the question posed by (1) is given by the set of curves in Figure 2, and the answer to the question posed by (2) is given by the curves in Figure 3.

The abscissa of the curves in Figure 2 is the ratio μ / σ, where μ is the mean of the quantity being measured, and σ is the composite standard deviation of this quantity. That is, this value of σ is the standard deviation for measurements of a quantity pertaining to <u>one subject</u>, measured possibly several times each day over a number of days. We define the following quantities:

σ_1^2 = Subject-to-subject variance of the quantity being measured.

σ_2^2 = Variance introduced by the measuring device, or "reproducibility."

σ_3^2 = Variance introduced by day-to-day variation of a measured quantity in the same subject.

n_1 = Number of subjects.

n_2 = Number of times a measurement is taken on one subject in one day.

n_3 = Number of days on which measurements are taken on one subject. It is assumed that n_2 is the same for all these days.

PROBABILITY THAT $H_0: \mu \leq 0$ WILL BE REJECTED, AS A FUNCTION OF μ/σ AND SUBJECT SAMPLE SIZE, n_1.

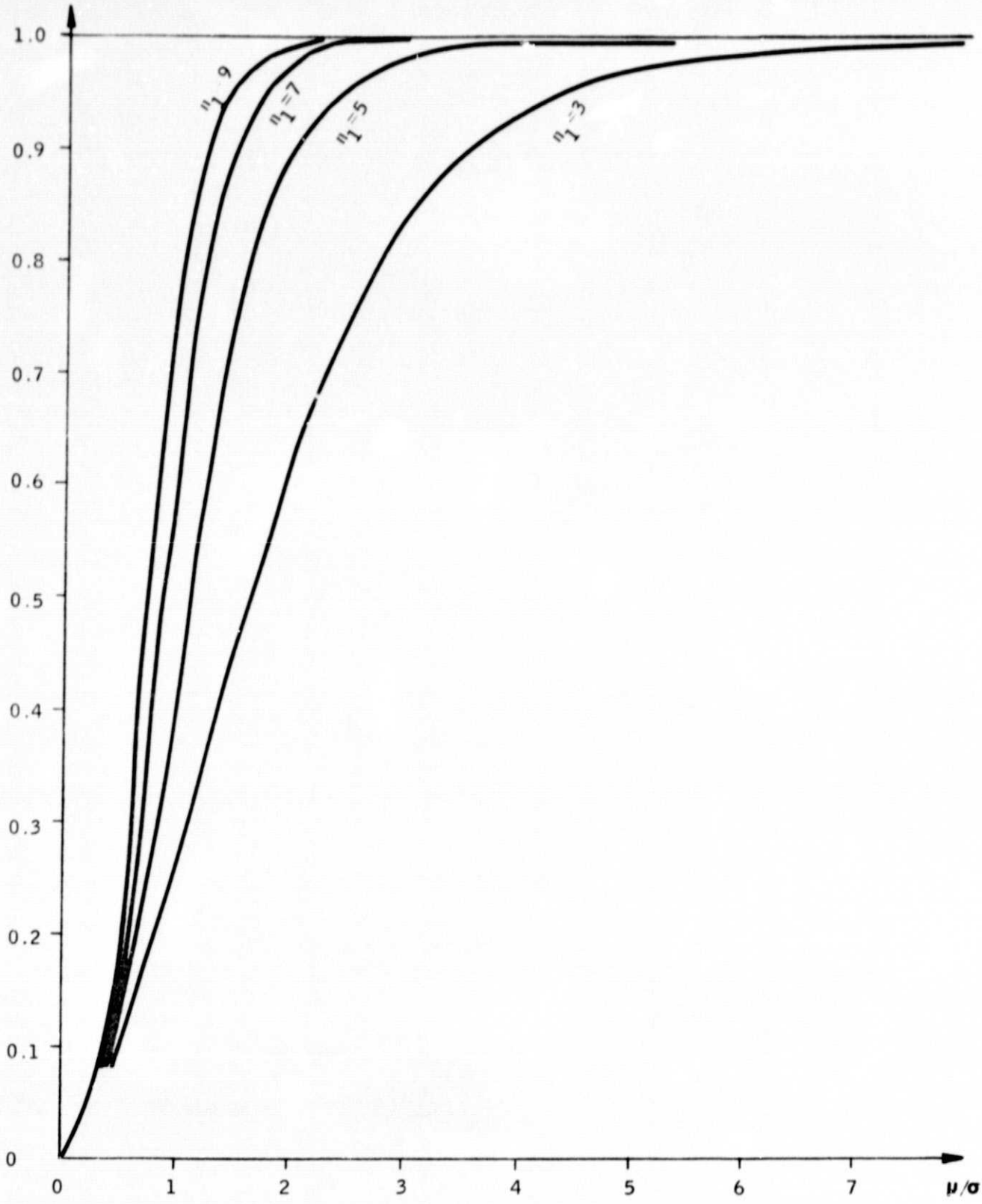


FIGURE 2

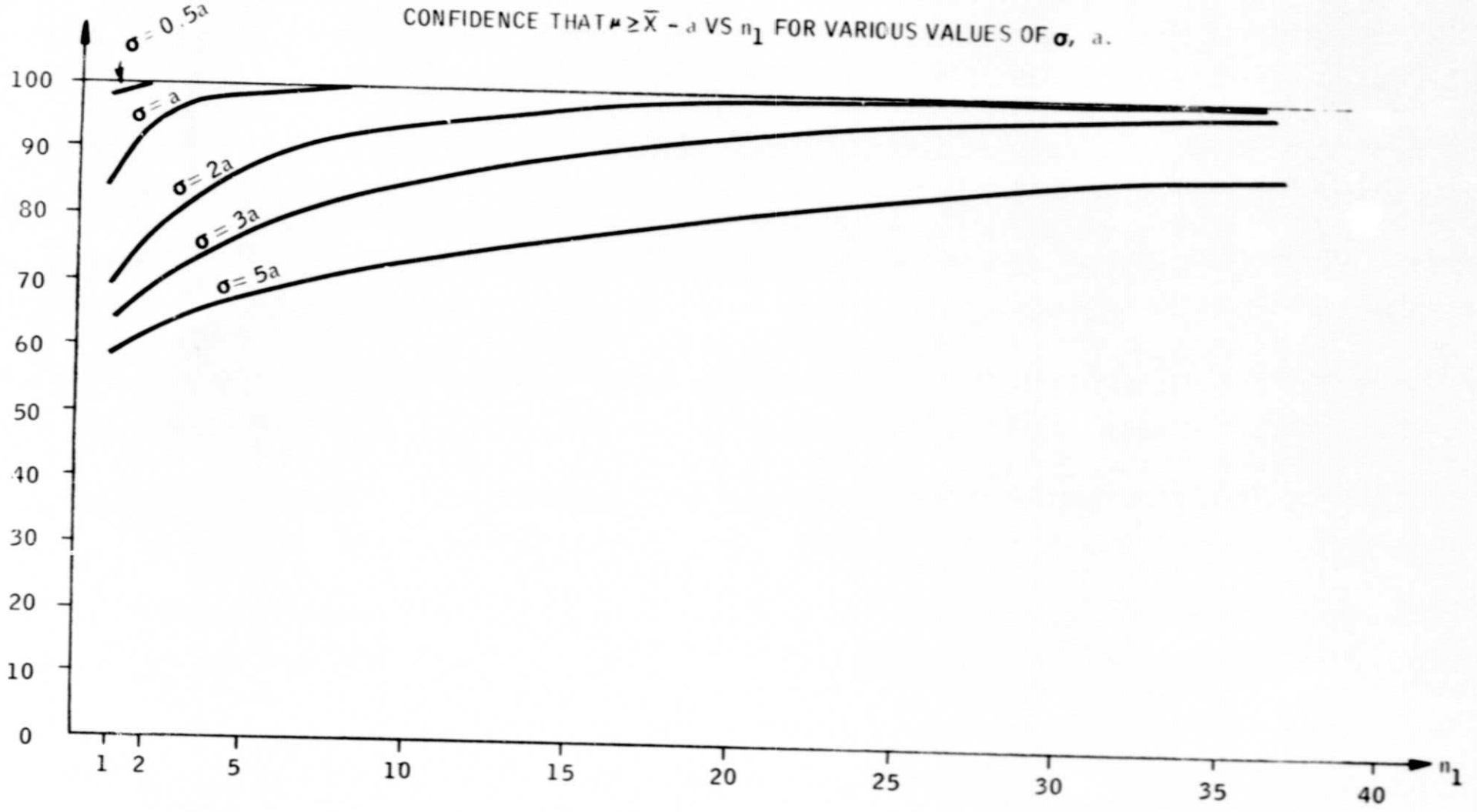

CONFIDENCE THAT $\mu \geq \bar{X} - a$ VS n_1 FOR VARIOUS VALUES OF σ, a.
$\sigma = 0.5a$
$\sigma = a$
$\sigma = 2a$
$\sigma = 3a$
$\sigma = 5a$
100
90
80
70
60
50
40
30
20
10
0
1
2
5
10
15
20
25
30
35
40
n_1

FIGURE 3

If the quantities are averaged over all measurements on each day, the variance due to measurement error is reduced to σ_2^2 / n_2. If these quantities are averaged over the total number of days on which measurements are taken, the combined variance due to both measurement and day-to-day variation is

$$\frac{\sigma_2^2}{n_2 n_3} + \frac{\sigma_3^2}{n_3} .$$

Thus, the quantity measured for one subject and averaged as described above is normally distributed with mean μ and standard deviation equal to

$$\sigma = \left[\sigma_1^2 + \frac{\sigma_2^2}{n_2 n_3} + \frac{\sigma_3^2}{n_3} \right]^{1/2} .$$

This is the σ used in the abscissa value μ / σ in Figure 2. The values labeled n_1 in the figure represent the number of subjects, n_1.

The ordinate of the curves in Figure 2 represents the probability that the hypothesis test will call a change significant; i.e., reject the hypothesis that $\mu \leqq 0$. For example, if average TBW loss has the same mean equal to 1 (i.e., $\mu = 1$) over three days and overall n_1 subjects (note that this does not say that actual TBW loss is the same for all subjects; it simply says that the subjects may be considered as belonging to the same statistical population for the three days, and that the mean for this population is equal to μ), but $\sigma_1 = \sigma_2 = \sigma_3 = 0.5$, then from the formula for σ above, we have

$$\sigma = \left[(0.5)^2 + \frac{(0.5)^2}{9} + \frac{(0.5)^2}{3} \right]^{1/2} = (0.5)\ (1.20) = 0.6.$$

Thus, $\mu / \sigma = 1.6667$, and the curves in Figure 2 show that the $\mu \leqq 0$ hypothesis will be rejected with a probability of about 0.55 if $n_1 = 3$, 0.85 if $n_1 = 5$, 0.94 if $n_1 = 7$, and 0.98 if $n_1 = 9$.

On the other hand, if σ is still 0.5, but we take so many measurements over so many days that the effect from σ_2, σ_3 may be neglected, we shall have σ now approximately equal to 0.5, so that $\mu / \sigma = 2$. For this case the probabilities rise to about 0.62, 0.90, 0.97, and 0.995 for $n_1 = 3, 5, 7, 9$, respectively.

Or, if the σ's all decreased to 0.4167 for the first case of three measurements and three days these latter numbers would again result.

The following conclusion is evident from the second example: the probability that can be obtained by increasing n_2 and n_3 is bounded by the value of n_1. The second example is tantamount to assuming that $n_2, n_3 = \infty$. The only way to raise the probabilities higher than these values is to raise n_1. On the other hand, it is clear that we can achieve as high a probability as we like by increasing n_1.

The curves in Figure 3 represent the confidence that $\mu \geqq \overline{X} - a$, where $\overline{X}$ is the sample mean of measurements on n_1 subjects and a is some positive number. Here the confidence is entirely independent of the actual value of $\overline{X}$; the only dependence is on the values of a, σ and n_1, where σ is defined as above.

If we suppose that $\mu = 1$ and $\sigma = 0.6$ as in the first example above, and that a = 1, then we have that $\sigma = 0.6a$ and thus, the confidence lies between the curves $\sigma = 0.5a$ and $\sigma = a$. Therefore, it is about 95%, even for a sample of only one subject. That is, the confidence that $\mu \geqq 0$ is about 95%.

On the other hand, if $\sigma = 0.5$ and a = 1/2, we get $\sigma = a$, and thus the confidence that $\mu \geqq 1/2$ is 84%, 92%, 96%, and 97.5% for $n_1 = 1, 2, 3, 4$, respectively. For this same case if σ increases to 1.0, then $\sigma = 2a$, and the confidence that $\mu \geqq 1/2$ will be 69%, 76%, 80%, 84%, 87%, and 89% for $n_1 = 1, 2, 3, 4, 5$, and 6, respectively.

In this way, the confidence values may be determined by using the curves for any given values of n_1, n_1, n_1, σ_1, σ_2, σ_3.

2.0 APPLICATION OF MEASUREMENT ERROR ANALYSIS TO SPACE-FLIGHT STUDIES

2.1 COMPUTER TOMOGRAPHY MEASUREMENT ERRORS

Background

In the work by Elsasser*, the author presents results from two types of experiments designed to give an indication of the accuracy that can be expected from measurements by computed tomography. The first type of experiment consists of measurements on objects designed to simulate actual bone both as to shape and as to absorptive properties relative to the radiation used in computed tomography. The materials used are aluminum and plexiglas, which are also used in models for the photon absorption method. The author writes that these materials provide a satisfactory approximation of physiological conditions. For modeling of trabecular bone and marrow, a mixture of aluminum powder and PMMA cement ("beracryl®") is used. The author also presents considerable detail on the actual structure of these objects for modeling bones, but these will not be given here.

The advantages of carrying out measurements on such a model are, of course, that the accuracy of the method may be tested by comparing the results with the known density of the object being measured.

The results of the tests for these models are given on p. 86 of Elsasser's dissertation, in Tables 5.16, 5.17, 5.18, 5.19, and 5.20. In Table 5.16 are given results for the following models:

a) Model to simulate the total bone: Plexiglas, aluminum tubing and PMMA/aluminum cylinder, an illustration of which is given in Figure 5.7.

b) Model without "compact bone": The aluminum tubes in a) are replaced by plexiglas tubes of the same dimensions.

c) Model without any "soft tissues": The outer plexiglas cylinder is omitted, and only the aluminum tubing, with the PMMA/aluminum cylinders inserted, is measured.

d) "Compact bone" alone: Only the aluminum tubes are measured.

*Quantifizierung der Spongiosadichte an Rohrenknechen mittels computer tomographic (Quantification of Trabecular Bone Density in Tubular Bones by Computed Tomography).

e) Models without "spongy bone": Instead of the PMMA/aluminum mixture, the interior of the aluminum tubing contains plexiglas cylinders. The thickness of the aluminum tube wall, i.e., the "compact bone" thickness, varies as follows:

e1) Wall thickness of tubing = 1.5 mm.

e2) " = 3.0 mm.

e3) " = 4.0 mm.

e4) " = 5.0 mm.

The numbers given in the table are not densities, but rather "mean linear absorption coefficients" of the materials being measured. These coefficients have units of cm^{-1}. The model configuration a) (line a) in the table) gives trabecular bone density in these units for the model of an actual bone. The measured result is the value 0.677 ± 0.007 for a "true value" of 0.675 ± 0.005. The author notes that the ± 0.005 is included because the "true" value of the linear absorption coefficient for the PMMA/aluminum powder mixture cannot be precisely determined.

Comparing this measured value with the "true value" shows that for simulation of the actual bone by the aluminum/plexiglas/PMMA/aluminum powder model, the method of computed tomography appears to measure trabecular bone density to a very high degree of accuracy; the relative error is only +0.3%.

Models b) and c) correspond to trabecular bone without any compact bone and without any tissue outside the compact bone, respectively. These are, of course, deviations from physiological conditions so severe that they will never arise in applications with astronauts as subjects. Nevertheless, even with such severe deviations, the relative error is bounded in absolute val ue by the level of 2.1%. Models e1) through e4) simulate the case where there is no trabecular bone; nevertheless, the given results are still labeled "trabecular bone density" (Spongiosa-Dichte) in the dissertation. It is not explicitly stated what trabecular bone density means for this case.

Table 5.18 shows the results of measurements to determine "reproducibility." This is simply a compilation of results of repeated measurements on the same object 10 successive times without changing the position of the plane of measurement. Thus, an idea of the variability of the measurement is provided, with the result that here the estimate, S, of the standard deviation of the measured density is 0.9% of the density. These measurements were all carried out on the same day.

Table 5.19 shows results of 10 measurements taken at intervals of 10 mm along the longitudinal axis of the model. Here the author cites a standard deviation of 1.2% (on p. 88), which he attributes to inhomogeneities of the density of the PMMA/aluminum powder mixture.

Long range reproducibility (over 12 months or more) is given by Table 5.20. The author comments that there is no systematic error detectable due to age of the radiation source used in the measurements.

The increased standard deviation of these measured values is attributed to density variations in the model and a decreasing statistical accuracy as a function of age of the radiation source (p. 88, , section 5.2.5).

Before going on to general measurements on human subjects, we mention a remark of the author on p. 97 to the effect that the observed difference between the digital tomographically determined trabecular bone densities of normal and osteoporotic femurs are greater by a factor of 10 than the total observed mineral content.

The second type of experiments which were carried out were those on human subjects. Here the location of measurement was on the radius of the right arm at a distance of 10% of the ulna length from the ulna styloid process. The density of trabecular bone is defined as the mineral value of all matrix elements of the area located in the interior of the radius, equidistant from the outer edge of the bone, and comprises 50% of the total bone cross-sectional area.

The first experiment discussed was carried out on 96 subjects, of which the majority fall in the two categories 5-16 years and 20-40 years of age. The author says that the presence of most of the subjects in these two age groups is random.

The results of this experiment are presented in Figure 6.4 and Table 6.5. The latter gives numerical estimates for mean and standard deviation of the subgroups: 14 girls, 23 boys ; 13 women, 46 men ; 37 boys and girls, 59 men and women, 27 girls and women, 69 boys and men, and, finally, the total of 96 subjects. The mean trabecular bone density is found to be about 0.765, with an estimated standard deviation of 0.120. Perhaps the most significant result is that there is no apparent difference in the measured results as a function of age.

The reproducibility experiments in human subjects yield the results of most concern to planners of the Space Shuttle experiments. These results are presented and discussed in Section 6.3.3 on p. 113 of Elsasser's dissertation. The author mentions in a general way that the reproducibility error is, as in the case of the nonhuman models, a function of the positioning of the plane of measurement along the arm. He cites some examples of measurements made on humans where the reproducibility was of the same order of magnitude for both humans and the models; i.e., about 1.5%. He then mentions that for all subjects one may not necessarily hope for such a good reproducibility. He cites two prerequisites for good reproducibility: (1) a high degree of cooperation on the part of the subject, and (2) the trabecular bone density in the area being investigated must not change by more than two percent for each one percent change in measuring position along the longitudinal axis of the ulna.

In the experience of the author, the proper measurement location can seldom be found with an accuracy of less than 2.5 mm. This means that the reproducibility depends not only on the length of the arm, but also on the density gradient along the longitudinal axis of the bone. As an example of a sharply changing trabecular bone density, he presents Figure 6.1.4. Here a shift of 3 mm in one

direction or another leads to a change in trabecular bone density from +6% to -7% of the value at 10% of ulna length.

The author sums up by stating that under optimal conditions the trabecular bone density has a reproducibility of about $\pm$ 1.5%, but under less favorable conditions two measurements on the same subject may differ by more than 10%. A more precise estimate of reproducibility for a single subject may only be made with knowledge of the axial density gradient; this requires several measurements. One possible way to improve reproducibility is to take a plaster cast of the arm and to locate the measurement plane by use of this cast on each subsequent measurement. The only drawback the author sees with this method is in measuring children, primarily because their arms tend to grow in size between measurements over a period of months. Thus, there appear to be no foreseeable problems in taking measurements in Space Shuttle astronauts.

In Section 6.3.4 (p. 117) the author mentions another problem which tends to decrease the accuracy of the measurements: movements of the subject during the time in which measurements are being taken. He states that it is impossible to eliminate entirely this source of error, and says that the two sources of error: positioning and movement result in about 10% of all measurements being termed worthless.

Application

It seems clear that the primary source of error in the computed tomography measurement method is due to reproducibility, primarily because of uncertainty in positioning the plane of measurement along the arm. However, as the author notes, the likely magnitude of the reproducibility may be estimated by making several measurements 1 mm apart along the arm of the subject. Thus, if the accuracy cannot be improved, at least the experimenter may obtain an idea of the magnitude of the error, for each subject.

However, measurement errors are commonly assumed to be normally distributed about some mean. If this assumption is made, it implies that any number of successive measurements on the same subject may be used to approximate the true value of trabecular bone density more accurately than a single measurement. Under this assumption, the mean of a number n of such measured values, i.e., the quantity

$$\frac{X_1 + X_2 + \ldots + X_n}{n}$$

is distributed with the same mean as the X_i, but with standard deviation $\sigma / \sqrt{n}$, where σ is the standard deviation of a single measurement.

Thus, theoretically the standard deviation of an estimate of trabecular bone density may be brought arbitrarily close to zero, simply by taking a sufficiently large number of successive, mutually independent measurements.

One problem with this approach is that the estimate thus obtained for the mean error may have some bias; i.e. may not be zero, but may lie on one side or the other of zero. In the case of a plaster cast (to help in determining the location of the measurement plane), a small bias could easily ensue for the measurement of density, but since it would be essentially the same for preflight and postflight, it would vanish for the measurement of the spaceflight-induced change in density. If the position for measurement is always selected by only one person,

it is conceivable that some bias could be present, but here too, it would tend to cancel out if the same person makes the selection for both preflight and postflight measurements.

However, if one person selects the position for preflight measurements and another for postflight measurements, and their location biases (if any) reinforce each other (e.g., the preflight person tends to locate the measurement plane one mm too close to the wrist and the postflight person tends to locate it too close to the elbow), then serious systematic errors in measured density change may well result, especially for arms with high density gradients. Hence, it appears advisable either to have location done by the same person on both occasions, or to make several measurements with the location done by different persons for each measurement, both before and after.

Because of this and other considerations (e.g. Elsasser states that the statistical properties of the radiation source change slightly with time; by this he means apparently a progressive increase of the standard deviation of measurements; i.e., reproducibility error; his statement of no systematic error means that the mean is zero), it seems reasonable to assume that while we can bring the standard deviation of successive measurements quite close to zero, we probably cannot, in real life, bring it arbitrarily close to zero, as was mentioned above for the theoretical case. Another practical consideration supporting this reservation is that we only have a relatively short time in which to perform postflight measurements, since the bone density is expected almost immediately to start increasing back toward the 1-g level.

So to be conservative, we might assume that by taking 9 or 10 preflight measurements and 9 or 10 postflight measurements (both under such conditions that we may be sure that density change during measurement is negligible; probably all measurements on a subject should be taken on the same day), we could reduce the variance of the mean of the measurements by a factor of at least 8 (instead of 9 or 10, as in the theoretical case), which means a reduction in the

standard deviation of each (preflight/postflight) density estimate by a factor of $\sqrt{8} = 2\sqrt{2}$. If σ represents reproducibility, or standard deviation of the actual densities, then the standard deviation of the difference between the densities is given by σ times $\sqrt{2}$. Thus, this assumption implies a reduction in the standard deviation of the actual difference between zero-g and 1-g densities by a factor of two.

While this factor may seem to be low, it should be kept in mind that experiments on actual subjects may show it to be somewhat higher. Also, since good estimates of reproducibility error may be obtained by taking computed tomographic estimates at several adjacent points on the arm spaced equidistant from the desired measuring point, subjects who will have very large reproducibility σ's can be identified. It seems likely that measurements can be carried out on a sufficiently large randomly selected sample of the population from which subjects are chosen to determine the distribution of density gradients in the radius over the total population with a high level of confidence. If this distribution then indicated that, say, only 2% of the population have gradients implying σ of more than 9%, such people could be excluded from the experiment without strongly impinging on the representativeness of the sample finally chosen; i.e., it would still represent at least 98% of the population.

Examples

An example of trabecular bone density differences between an immobilized (for three weeks, due to a fracture) arm and the opposite arm of 14 children is given in the reference: Dynamics of Trabecular and Compact Bone Mineral of the Radius after Immobilization of the Upper Arm in Children, by Elsasser, Exner, Prader and Anliker. Here there is a rather large sample standard deviation (13% of trabecular bone density in the healthy arm) for trabecular bone density loss; this leads to a 95% confidence interval of $\pm$ 6.21 about the sample mean of 17%; i.e., the probability is 0.95 that the true mean of the population lies somewhere

in the interval (10.79, 23.21). Further, the fact that comparisons were made with the healthy arm of the same subject tends to confuse deviations in bone loss of the type which would occur in a zero-g environment with deviations which occur as a consequence of the variation in physical activity of the healthy arms of the subjects in this experiment. Also, unless care was taken in selection of the subjects, some of the variation may be due to the possibility that some of the arms fractured were dominant, and others nondominant.

Nonetheless, it seems instructive to see what the results of a t-test would be for a population having a true mean of 17% and 13% standard deviation. Here, if we assume that the reproducibility standard deviation is (after we have reduced it as much as we can) 2%, then the σ for the curves in Figure 2 is

$$\sigma = \left[13^2 + 2^2\right]^{1/2} = \left[173\right]^{1/2} \approx 13.15,$$

and hence, μ / σ, the abscissa of these curves, is about $17/13.15 \approx 1.29$. The curves then show that for such a high standard deviation for the population, the probabilities that a t-test will deliver a verdict of $\mu > 0$ for the general population are 0.39, 0.67, 0.84 and 0.94 for number of subjects equal to 3, 5, 7, 9, respectively, where the significance level of the test is 0.05 (probability of rejecting the hypothesis $\mu \leq 0$ when it is true). If the reproducibility of the computed tomography method should be so bad that its standard deviation is 10%, then the σ becomes

$$\sigma = \left[13^2 + 10^2\right]^{1/2} = \left[269\right]^{1/2} \approx 16.4,$$

so that $\mu / \sigma \approx 1.07$, and the probabilities for the t-test are now 0.33, 0.56, 0.77 and 0.89 for number of subjects equal to 3, 5, 7, and 9, respectively.

These results indicate that for such a large σ due to subject-to-subject variation, even errors introduced by what is close to the worst possible reproducibility (10% + only one measurement) make relatively little difference in the

probability of rejection/acceptance of the hypothesis that the average loss in trabecular bone over the entire population is greater than zero.

To get an idea how the t-test will behave if the σ due to subject-to-subject variation is somewhat less than in the example above, we present another example.

In the introduction to his dissertation, Elsasser says that total trabecular bone loss seems to vary from 10% to 41% for subjects who are inactive for 3 to 4 weeks. If these are extremes, i.e., if 99% of all subjects suffer losses over this period between these limits, and if, further, the losses are normally distributed, then typical values for σ, and mean loss are 6% and 25%, respectively, of total trabecular bone density. Here if measurement standard deviation is again rather high (=10%) and we can reduce it by a factor of two, then the σ value for Figure 2 becomes

$$\sigma^2 = (0.06)^2 + (0.05)^2 = 0.0778.$$

Since μ is 25% we get $\mu / \sigma = 0.25/0.0778 \approx 3.22$, and Figure 2 shows then that the t-test will say $\mu > 0$ with probabilities 0.89, 0.98, 0.999, 0.999 ~ for numbers of subjects = 3, 5, 7, 9, respectively. Again the computed tomographic reproducibility error plays a relatively minor role in determining what the test will do.

Next, consider a single subject with rather high reproducibility σ_2, say $\sigma_2 = 0.09$, or 9% of total density, and suppose we can only reduce this by a factor of 2 by repeated measurements, etc. It may be of interest to know, under these circumstances, the probability that the measured value is at least equal to 90% of μ (the actual value which would be returned by computed tomography if there were no reproducibility error) or 75%, 50%, 25% of μ, or simply the probability that the measured value is greater than zero. Curves for these probabilities are given in Figure 4. The abscissa here is the value of μ (μ =0.1 means μ is 10% of trabecular bone density), the mean bone loss.

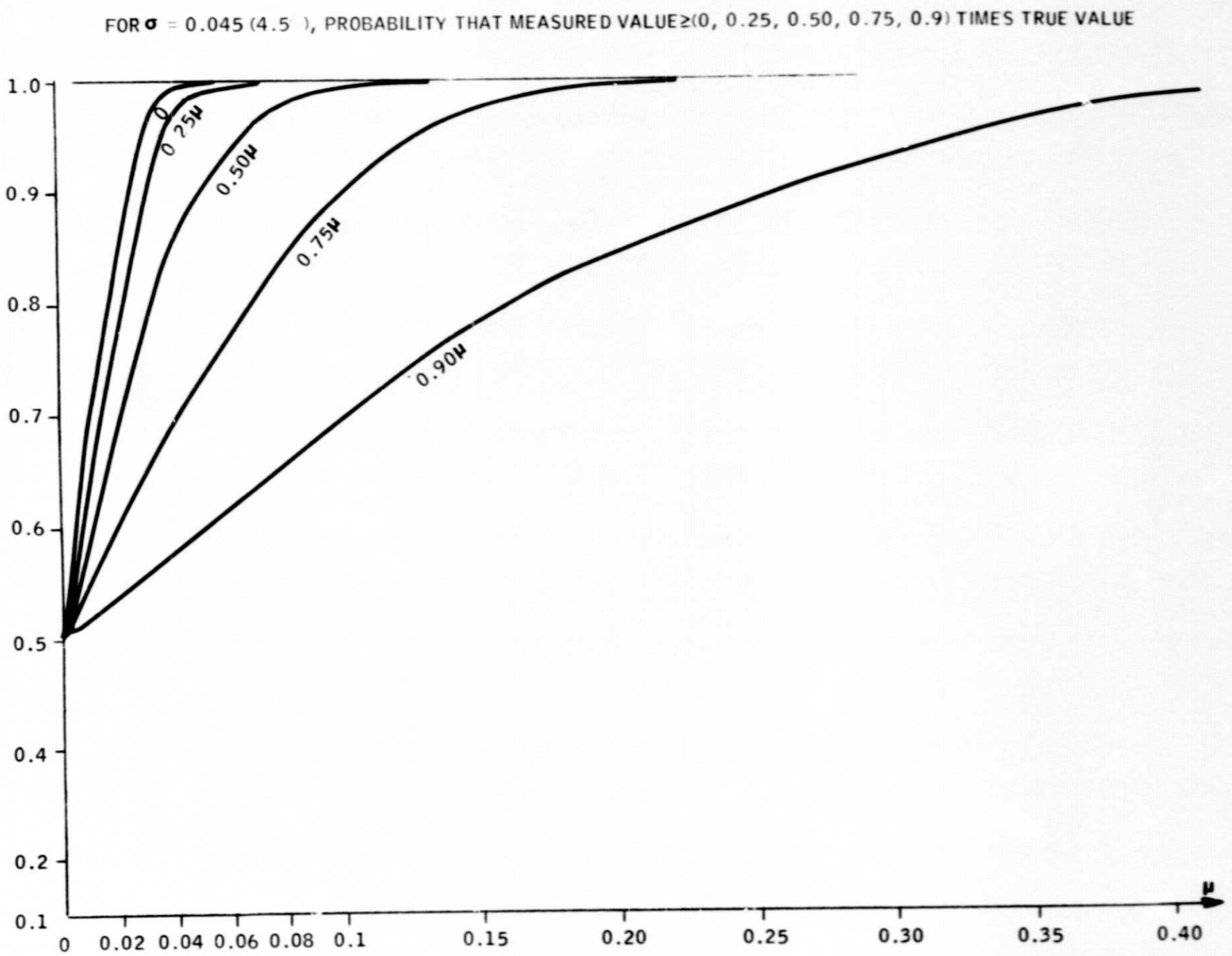

FOR σ = 0.045 (4.5), PROBABILITY THAT MEASURED VALUE≥(0, 0.25, 0.50, 0.75, 0.9) TIMES TRUE VALUE
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.2
0.1
0
0.25μ
0.50μ
0.75μ
0.90μ
μ
0 0.02 0.04 0.06 0.08 0.1 0.15 0.20 0.25 0.30 0.35 0.40

FIGURE 4

We see from Figure 4 that if the true loss is 5%, the probability that the measured loss will be at least 5%, 4.5%, 3.75%, 2.5%, 1.25%, or 0 is 0.500, 0.520, 0.609, 0.709, 0.800, and 0.870, respectively. For a 10% true loss, the corresponding numbers are 0.5, 0.587, 0.712, 0.867, 0.953, and 0.98. Thus, for a single subject with $\sigma_2 = 9\%$, we could feel quite confident (95.3% confidence) that the measured value would be at least 2.25%.

If on the other hand we have a subject with $\sigma_2 = 2\%$, or one for whom we can bring σ_2 down to 2% by repeated measurements, etc. the curves will look as in Figure 5. Here an actual loss of 5% gives probabilities of measured values at least 5%, 4.5%, 3.75%, 2.5%, 1.25% and 0 of 0.500, 0.579, 0.734, 0.894, 0.969, and 0.994, respectively. Thus, the probability that measured value $\geq 1.25\%$ is here 0.969, so that a 5% change seems quite reliably detectable.

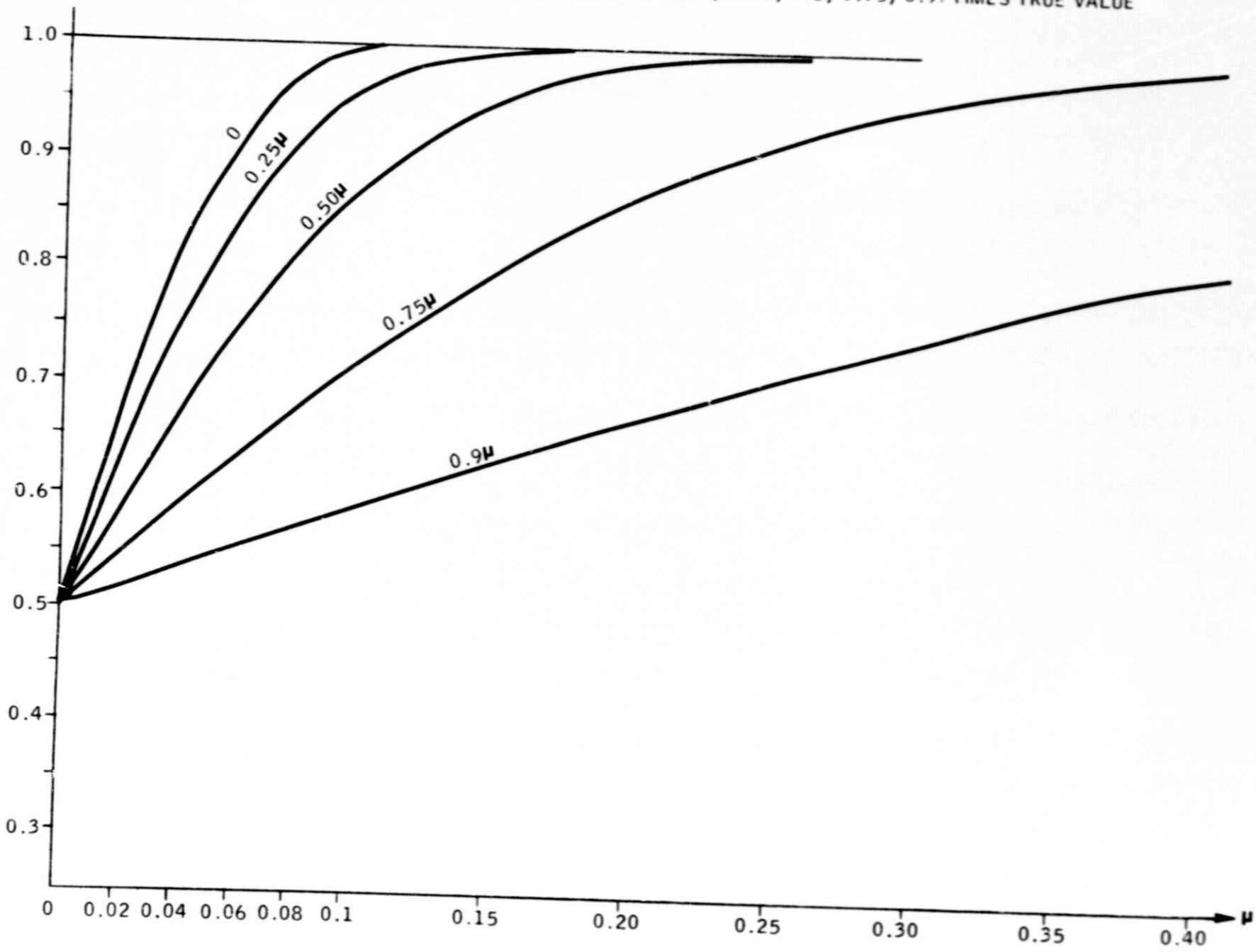

FOR σ = 0.02 (2%), PROBABILITY THAT MEASURED VALUE ≥ (0, 0.25, 0.5, 0.75, 0.9) TIMES TRUE VALUE
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0
0.25μ
0.50μ
0.75μ
0.9μ
0 0.02 0.04 0.06 0.08 0.1 0.15 0.20 0.25 0.30 0.35 0.40
μ

FIGURE 5

2.2 BASAL METABOLISM MEASUREMENT ERRORS

The following questions have been posed on this subject:

In order to detect, at the 95% confidence level, a 5%, 10%, 20%, or 30% difference in basal metabolic rate, how many subjects do I need, how many repetitive measurements should I make on a single subject, how many times over a period of days should I repeat the measurements, and what statistical analysis techniques should I use?

The answer to these questions is given in terms of the parameters n_1, n_2, n_3; σ_1, σ_2, σ_3, all of which were previously defined. First, the question of detecting the above mentioned changes with 95% confidence will be discussed, using some curves presented below, and then the question of analysis will be discussed using curves in Figure 2, presented previously.

For basic metabolic rate, the quantities σ_1, σ_2, σ_3 may, of course, have different values for each of the parameters which define this rate. We shall discuss the problem of detecting a change in one of these parameters and mention that the statistical features of this problem are identical for any one of the parameters.

The quantities σ_1, σ_2, σ_3 represent standard deviations due to (1) subject-to-subject variation, (2) measurement-to-measurement variation, and (3) day-to-day variation, respectively. Since we are here only interested in changes in a parameter, we may expect to be measuring values applying to post-flight conditions (i.e. conditions which obtain at the end of flight, not those for a day or two after completion of flight), and those applying to preflight conditions. The measurements will presumably be made over a sufficiently short period as to preclude any significant variation with time, because time variations will affect the measured value of a difference. Hence, the value of σ_3 here is zero, and $n_3 = 1$.

If such parameters have never previously been measured during spaceflight, the experimenter has no way of knowing the value of σ_1, although he may be able to obtain a fairly close estimate for this through bed rest studies. If there are data available from previous flights, the value of σ (the composite σ given by contributions from σ_1, σ_2, σ_3 and particular values of n_2, n_3; i.e. $\sigma = \left[\sigma_1^2 + \sigma_2^2/n_2 n_3 + \sigma_3^2/n_3\right]^{1/2}$) for the change in a parameter from 1-g to zero-g may be estimated by the formula for S given previously. This formula is also used, of course, in estimating σ_1 from bed rest data.

The user should usually have a fairly good estimate for σ_2 from the manufacturer of the measuring instrument, or some other source.

The procedure for measurements is to take the measurements of all subjects under conditions which the experimenter considers to represent satisfactorily the 1-g environment, and then take identical measurements of the same parameters immediately after return from spaceflight, while the parameters still are as close as possible to the zero-g values. Or, if it is feasible to take measurements daily aboard the spacecraft, this may be done. It affords the advantages of permitting detection of any appreciable time trends in the parameters under the influence of spaceflight.

Before the actual experiments, if measurements are to be taken as described above, for any given values of μ (to be defined below) and σ (already mentioned above), certain things may be said about the likely outcome of the experiment, and the confidence the experimenter may have in his or her results. For this case, $\sigma = \left[\sigma_1^2 + \sigma_2^2/n_2\right]^{1/2}$, and if μ is the true mean change in a parameter for the population from which the subjects were selected, then a particular number n of subjects will be required for 95% confidence that the sample mean $\bar{X}_n$ of the measured values of a parameter:

$$\bar{X}_n = (X_1 + X_2 + \ldots + X_n)/n$$

(where the X_1, X_2, ... are the measured values and the total number of such values is n) is at least 90%, 75%, 50%, or 25% of the true mean μ. It can be shown

mathematically that this number n depends only on the ratio μ/σ. The functional dependence of n on μ/σ is depicted in Figure 6. Here it is assumed the measured value is a normal random variable with mean zero and standard deviation σ_2, so that the standard deviation of the sample mean of n_2 measurements is $\sigma_2/\sqrt{n_2}$.

The tables from which the curves in Figure 6 were plotted are the following:

0.9μ:

μ/σ	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	8.0	10
n	1089	273	121	68	44	31	22	17	4	3

0.75μ:

μ/σ	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
n	175	44	20	11	7	5	4	3

μ/σ :	0.5	1.0	1.5	2.0	2.5	3.5
n	44	11	5	3	2	1

0.25μ:

μ/σ	0.5	1.0	1.55	2.2
n	20	5	2	1

0:

μ/σ	0.499	0.522	0.55	0.58	0.623	0.67	0.74	0.83	0.96	1.17	1.65
n	11	10	9	8	7	6	5	4	3	2	1

It is thus seen that for a case where $\mu = \sigma$ (i.e. a rather large standard deviation σ) we have 95% confidence that if five subjects are chosen (n = 5), then $\bar{X}_5$ will be at least equal to 0.25μ, where μ is the true population change for the parameter we are estimating, and that even with only three subjects we still may have about 95% confidence that $\bar{X}_3$ will at least not be negative.

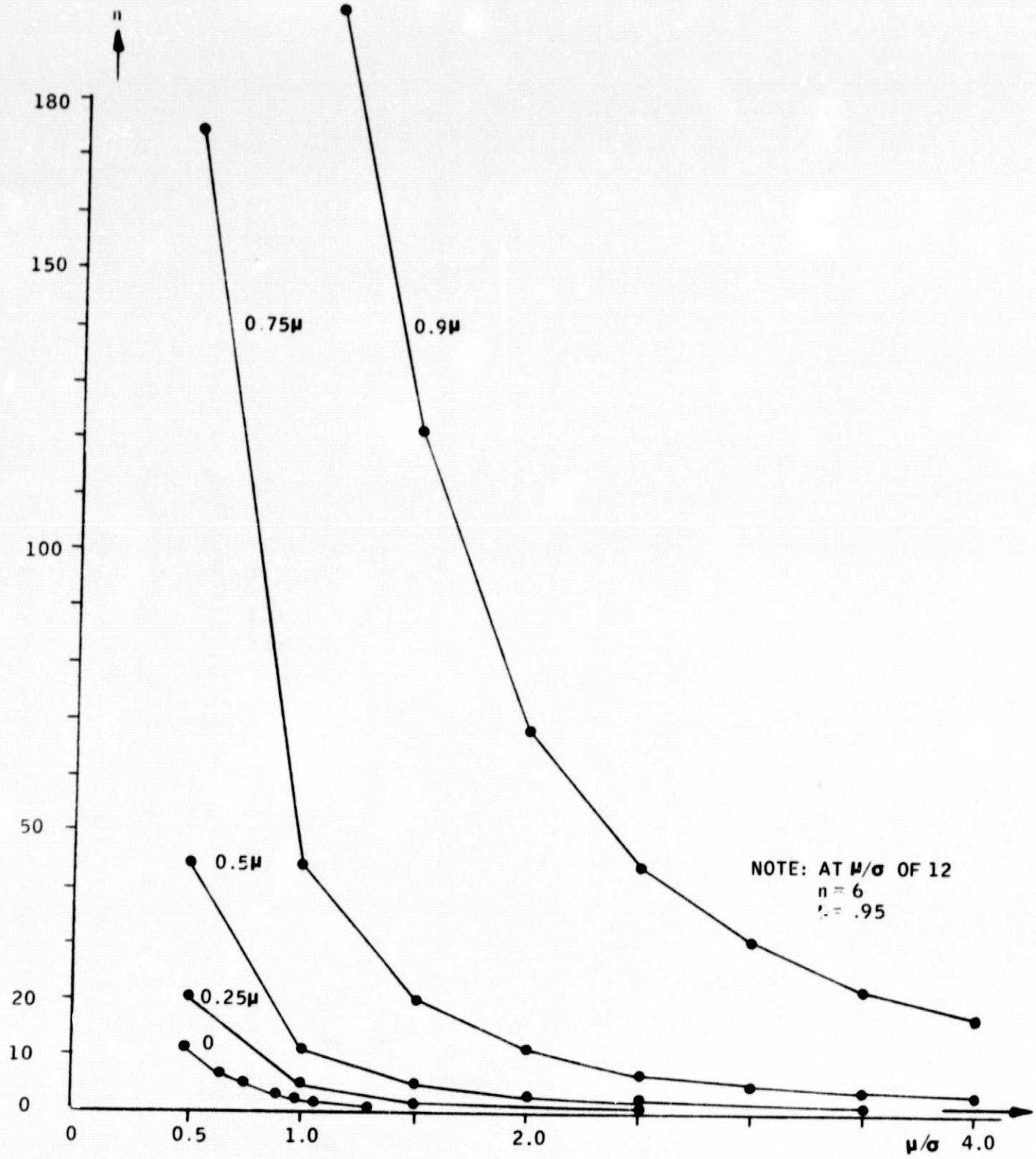

n
180
150
100
50
20
10
0
0.75μ
0.9μ
0.5μ
0.25μ
0
NOTE: AT μ/σ OF 12
n = 6
= .95
0
0.5
1.0
2.0
μ/σ
4.0

FIGURE 6

As for "trade-off" between repeated measurements and number of subjects, it is clear from the form of the tables and the curves that making the number of subjects large enough will enable us to detect <u>any</u> change, no matter how small μ/σ may be. For example, even if n_2 (number of measurements) is only 1, and if σ_2 is quite large, say $\sigma_2 = 3$, and σ_1 is 4, so that

$$\sigma = \left[3^2 + 4^2\right]^{1/2} = 5,$$

and if μ is only 2.5, so that $\mu/\sigma = 0.5$, then we can still achieve a 95% confidence level that the measured $\overline{X}_n \geq 0.9\mu$ by increasing n to 1089 subjects. However, if we hold σ_1, and n constant at 3 and increase the number n_2 of measurements to 1000, 2000, or 5000, we see by the formula

$$\sigma = \left[\sigma_1^2 + \sigma_2^2/n_2\right]^{1/2}$$

that our σ will still be at least 3. Thus, μ/σ is no more than 0.8333. , and thus the lowest curve applies; we may then only say that $\overline{X}_3$ will be greater than or at least equal to zero.

As for statistical analysis techniques, the data from an experiment are often used to test the hypothesis that the true mean change in a parameter for a population (e.g. the population of all healthy subjects between 28 and 38 years of age) is greater than zero or ≤ 0. This is done by a "t-test," which is explained in the text accompanying Figure 2. This figure presents the number of subjects necessary to ensure a particular probability that the hypothesis of no change, or of a change in opposite direction to that of the true change, will be rejected.

As an example of the use of these curves, suppose we have $\mu/\sigma = 0.5$, as mentioned above. Here it is clear from the curves in Figure 2 that many more than 9 subjects would be necessary to ensure a probability of 0.95 that a hypothesis contrary to the real change would be rejected. If, however, $\mu/\sigma = 2$, then only about

six subjects would be necessary for this, and with $\mu/\sigma = 4.3$, then only three subjects would be necessary. Again, we emphasize that such estimates for μ and σ might be obtained from bed rest studies or, if available, data for changes in the parameters of interest from past space flights.

2.3 BODY WATER MEASUREMENT ERRORS

Problem:

A new method* is proposed to measure total body water inflight based on ethanol dilution and non-invasive breath analysis. Assuming the changes in total body water during Shuttle Spacelab missions are similar to those observed in the nine Skylab crewmembers (see Figures 7 and 8), will this method provide the precision necessary to detect the expected water losses?

These data (Figure 7 and 8) serve to illuminate the problem of estimating what can be expected from future experiments to measure this parameter. We summarize these data below, and then apply them to the problem of estimating sample size and making other considerations for future experiments.

Data for TBW loss for a single crewman:

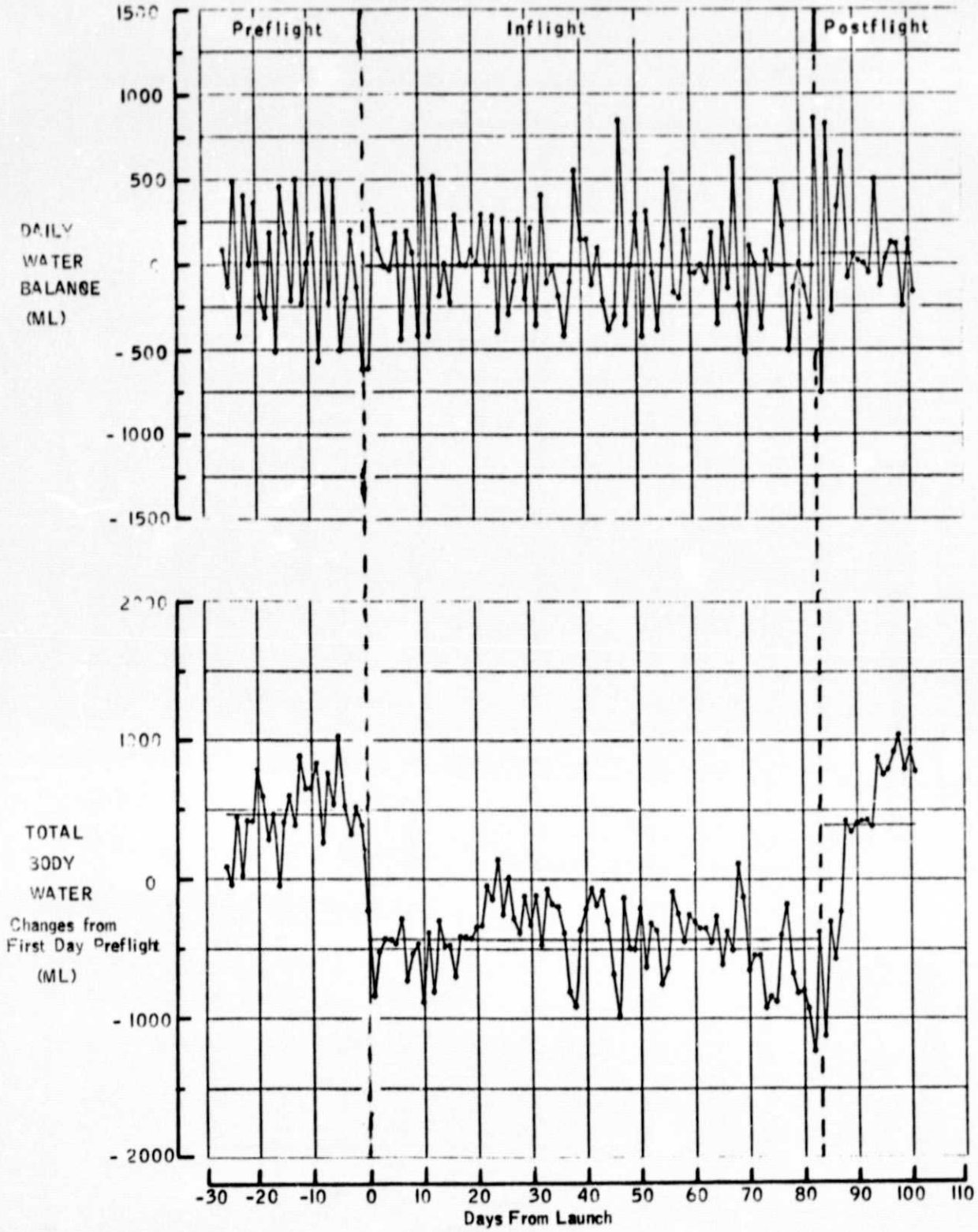
Here the mean change seems to be about 0.9 liter. Since 95% of the day-to-day measurements seem to fall within $\pm$ 0.5 liter of this postulated mean, it might be reasonable to assume that $2\sigma_3 \approx 0.5$ liter, or $\sigma_3 \approx 0.25$ liter, since this is the case if the day-to-day variation is normally distributed about the mean.

Data for TBW losses for the entire Skylab crew:

It looks here as if the mean TBW loss is equal to 1.4 liters, and as if σ is around 0.3 (since the vertical line for two days after launch is about 0.6 liters in length).

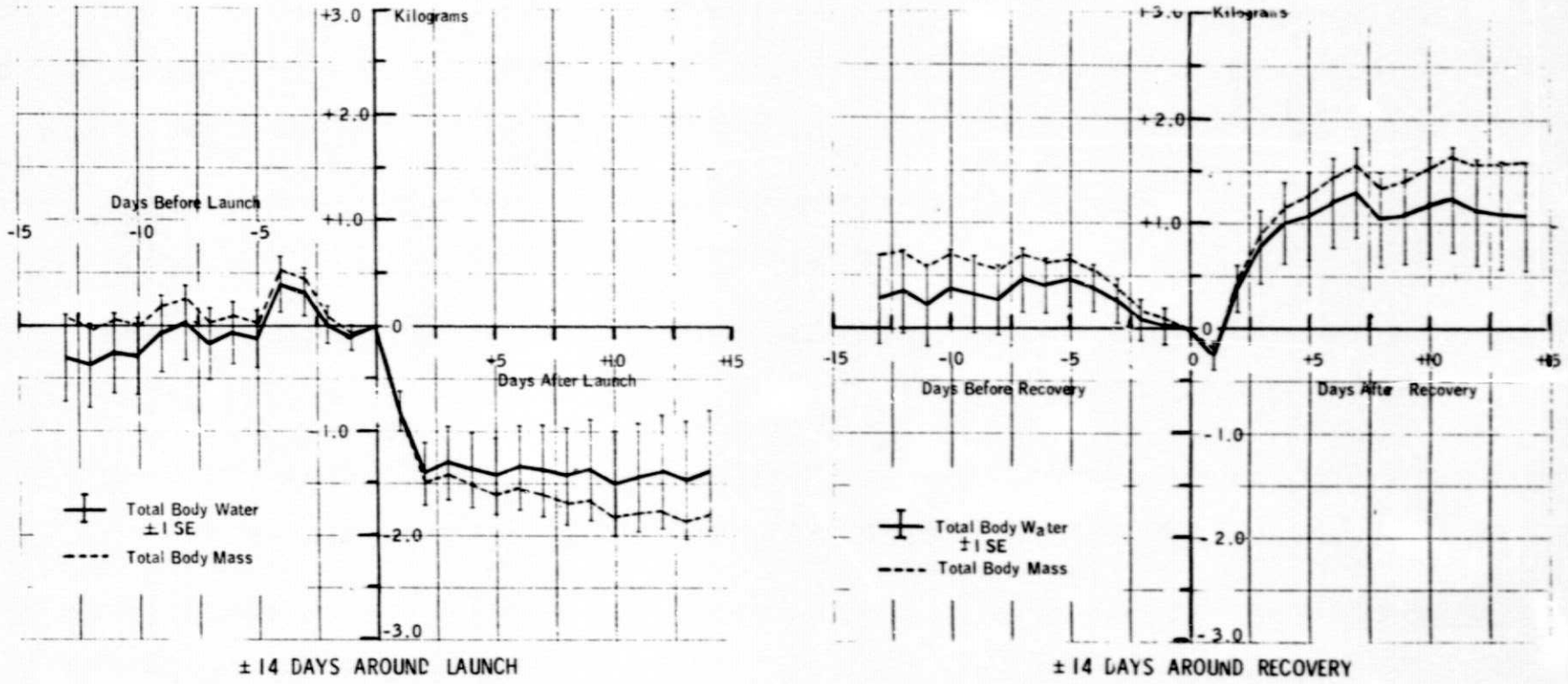
Since these values are radically different from the case of a single crewman, it appears possible that TBW loss might vary quite strongly with a person's normal (1-g) TBW level; i.e., change in TBW level might be quite strongly correlated with the 1-g TBW level. Thus, it might be advisable to use percentage of the 1-g TBW level as the parameter of interest for statistical analysis, rather than absolute

*Loeppky, et al (1977) Appl. Physiol. Respirat. Environ. Exercise Physiol. 42: 803-808.



DAILY WATER BALANCE AND TOTAL BODY WATER COMPUTED FROM WATER BALANCE EQUATION FOR A SINGLE SKYLAB CREWMAN (SL4/CDR)

FIGURE 7



DAILY CHANGES IN TOTAL BODY WATER AT LAUNCH AND RECOVERY OF ENTIRE SKYLAB CREW (N = 9)

(VALUES ARE SHOWN AS CHANGES FROM MORNING OF LAUNCH OR RECOVERY)

FIGURE 8

change in liters. If there is such a correlation, the measurement results might be appreciably changed, since the numbers in Table 1 of the article by Loeppky, et al (Total Body Water and Lean Body Mass Estimated by Ethanol Dilution) show a TBW range of 36.2 liters to 67.2 liters for a sample of 35 human subjects.

However, even the measurements in Figure 8 appear to show a striking uniformity in the individual TBW change, as indicated by the apparent standard deviation of 0.3 liter. If we suppose that these data yield sample standard deviation $S = 0.3$, then a 95% confidence interval for μ, the actual mean of the population, is

$$[1.1694,\ 1.6306\ ].$$

Furthermore, with these values a t-test at the 0.5 level would reject the hypothesis of change contrary to the true change, since

$$\frac{\bar{X}}{S/\sqrt{n}} = \frac{1.4}{0.3/3} = 14.0,$$

which is greater than $1.860 = t_8\ (0.95)$.

A study by Culebrat, et al (A Comparative Study of TBW as Measured by Isotope Dilution and Body Dessication in the Rat, Federation Proc. 35(3):450, 1976) reports TBW/wt = 0.702 by dessication and 0.714 by HTO. From Table 1 of the article by Loeppky, et al, the ETH measured value is 0.717 and that for HTO is 0.735, for a sample of 35 subjects. Therefore, there appears to be no significant difference in accuracy between the ETH and HTO methods, since the HTO methods deviates positively from dessication by about 1.68%, and the ETH method deviates negatively from the HTO method by about 2.45%, or from dessication negatively by about 0.77%, so that the ETH method may be slightly more accurate, if we regard dessication as a sort of absolute norm.

For purposes of predicting how future experiments might turn out in using the ETH method (assuming that the parameter values $\mu = 1.4$, $\sigma = 0.3$ mentioned above are valid), we may apply first the tables used in generating Figure 6. Here the graph of changes in TBW of the Skylab crew of 9 indicates that the σ value mentioned above incorporates both subject-to-subject and day-to-day variation, so that $\mu / \sigma = 1.4/0.3 = 4.67$. If to be conservative we drop this to 4.00, we see that we need 17 subjects to attain 95% confidence that the measured estimate will be at least 90% of the true value of μ. To attain this level of confidence that the estimate will be 0.75μ, we need only three subjects. For 0.5μ, 0.25μ, or 0 we need only one subject.

In other words, given a population TBW loss of 1.4 liters, and σ due to subject-to-subject variation and day-to-day variation of a little over 0.3 liters, the question, what loss could I detect, and how many subjects do I need to do it"? is answered by saying: "I would be able, with 95% confidence, to estimate a loss of at least 0.9 (1.4) = 1.26 liters with 17 subjects, at least 1.05 liters with three subjects, and 0.70 liters with only one subject."

Now let us postulate a less favorable scenario; namely, all parameters the same as before, but $\mu = 0.7$ liters, rather than 1.4 liters, as previously. With this assumption, μ / σ will be a bit higher than 2.0. If we conservatively assume $\mu / \sigma = 2.0$, the tables will now give 68, 11, 3, 2, 1 subjects needed in order to assert with 95% confidence that we shall obtain at least 0.63, 0.53, 0.35, 0.18, and 0 for the estimated population TBW losses, respectively.

Turning to the problem of estimating the outcome of the t-test, let us assume again the former case where the parameters are $\mu = 1.4$ and $\sigma = 0.3$. If we again suppose that $\mu / \sigma = 4.0$, the curves in Figure 2 give a probability of about 0.93 that the t-test will reject the hypothesis that $\mu \leq 0$. For 5, 7, or 9 subjects, this probability rises above 0.99.

With the second assumption made above, viz. that μ is only 0.7 liters, but the other parameters are unchanged, so that μ/σ may be taken to be 2.0, the curves give probability around 0.66 that the t-test will reject the hypothesis that $\mu \leq 0$. For 5, 7, or 9 subjects, this probability rises to about 0.92, 0.98, and 0.99, respectively.

3.0 CONCLUSION

In conclusion, the foregoing work represents a detailed discussion, first of the general aspects of experimental design as applicable to the Space Shuttle experiments, and then of statistical aspects particular to each of several of the proposed biomedical experiments for Space Shuttle. In retrospect, one impression seems to stand out somewhat more than any other. This is that, in the literature which was furnished to aid in statistical analysis, there was only one set of information which described results from previous actual flights in enough detail to form an idea of the statistical behavior which might be expected for spaceflight – induced changes in the parameters of interest. This is the set of information pertaining to TBW levels. As a result, the statistical commentary for all the other experiments had to assume a rather general character, largely in the form of curves and other information which should enable the experimenter to predict results for planned experiments only if he or she supplies statistical parameter values based either on previous spaceflight data or bed rest studies, or on the experimenter's subjective judgement.

For the TBW experiments, on the other hand, the information supplied permitted some fairly definite predictions to be made about what the results might be for future experiments. Here too, of course, the information presented for the other experiments can still be used to evaluate a variety of hypothetical scenarios which the experimenter may find useful in making predictions or decisions. It is hoped and suggested that past available data will be closely investigated (if this has not already been done) for any possible application in future experimental designs.