Journal of Andrology, Vol. 12, No. 2, March/April 1991
Copyright © American Society of Andrology

Issues in the Statistical Analysis of Sperm Motion Data Derived from Computer-assisted Systems

BETH C. GLADEN,* JACQUELINE WILLIAMS,† AND ROBERT E. CHAPIN†

*From the *Statistics and Biomathematics Branch, and the †Developmental and Reproductive Toxicology Group, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina.*

ABSTRACT: Computer-assisted sperm motion measurements present certain features that must be accounted for in statistical analyses. Some are specific to this type of data, whereas others are standard considerations. For example, the measurement of multiple sperm from individuals creates correlations that must be accounted for if each sperm's measurement is used, and unequal variances may arise that need to be addressed if an average measurement from the individual is used. Also, the limitations on the ranges of some measurements create discrepancies between observed and actual means and may make treatment-related effects more difficult to detect — a circumstance that has an impact on study design. When variables that are truly continuous are measured in a discrete fashion, odd effects may arise and care is needed. Other considerations, such as the shapes of distributions and correlations among various measurements, should also be examined. Attention to these details of statistical analysis are vital to proper interpretation of data.

Key words: Sperm motility, statistics, computer-assisted semen analysis (CASA).

J Androl 1991;12:89–97.

Computer-assisted sperm motion measurement technology has provided researchers with new kinds of data in large quantities. This wealth of information brings both new opportunities and new problems, some of which were discussed by Amann (1989) and Tash and Wolf (1989). As investigators, we need to be aware of the problems while taking advantage of the opportunities. In this article, we explore some of the potential problems that arise in the statistical analysis of these data. Most of the concepts are not new and may be found in standard reference texts. Similarly, many of these problems are not unique to andrology, but most are not familiar to andrologists. Thus, a discussion of some statistical problems and some potential solutions seems appropriate and timely. Where space does not permit full discussion, we refer the reader to appropriate statistical literature.

To illustrate our points, we will refer to data collected in a study of the effects of ethylene dibromide on rabbits (Williams et al, 1990, and Williams et al, 1991). We took weekly semen samples from 42 rabbits before, during, and after exposure; only data from the pre-exposure period will be used. The pre-exposure period lasted 6 weeks, and we were able to collect 223 usable samples during that time; 29 samples were unusable, primarily because of the presence of urine. The number of sperm analyzed varied among samples. For curvilinear velocity, for example, the number of sperm measured per semen sample ranged from 2 to 185, and the total number of sperm measured from all samples from all rabbits was 13,674. These data were generated using the CellSoft® system (Cryo Resources Ltd., New York, NY). The principles, however, apply to any species and to any of the current automated systems in use, as well as to any semi-automated systems using the same general methods of data generation.

Correlated Measurements

When the treatment or exposure under study is given to the whole animal but observations are made on individual sperm from that animal, the result is correlated measurements. This study method has led to questions about what is the "experimental unit" or the "unit of analysis." Use of such terms can be somewhat misleading since they imply a simplicity that does not exist; no choice of unit allows us to ignore the hierarchical nature of these data. Because of this hierarchical structure, simple analyses may not apply. Regardless of the unit chosen, analyses must accommodate the added complexity. Simple analyses assume that observations are both independent and identically distributed (at least within a group). These conditions do not necessarily apply to measurements from multiple sperm per animal. We discuss the implications of using the sperm or the animal as the unit and the considerations that arise in each case.

Correspondence to: Beth C. Gladen, Statistics and Biomathematics Branch, Mail Drop B3-02, National Institute of Environmental Health Sciences, P.O. Box 12233, Research Triangle Park, North Carolina 27709.

Received for publication July 13, 1990; accepted for publication September 19, 1990.

This question of the appropriate unit is not unique to sperm motion analysis. It arises in other fields with hierarchical observations, such as teratology, with multiple fetuses per female (Haseman and Kupper, 1979); ophthalmology, with two eyes per person (Rosner, 1989); and education, with multiple pupils per classroom or teacher (Hopkins, 1982). The main statistical principles are the same in all of these cases, although the question of primary interest may vary. For example, in ophthalmology, it is possible that the two eyes of a diseased individual will undergo different treatments; thus, methods that accommodate this difference are needed.

Lack of Independence: Using the Individual Sperm as the Unit

If data from individual sperm are analyzed, measurements of any particular endpoint for different sperm from the same animal are correlated. This correlation arises because values for sperm from some animals are consistently high while others are consistently low. Thus, two sperm from the same animal will tend to be more alike than two sperm from different animals. To show the magnitude of the problem, consider the following example of curvilinear velocity: the means for 42 individual rabbits ranged from 86 to 145 μm/second. Figure 1 shows the variation both within and between animals. Three animals are shown for illustration. Sperm from the animal shown in the top panel averaged 96 μm/second; those in the middle panel averaged 119 μm/second; and those in the lower panel averaged 141 μm/second. In general, if the variance within animals is W^2 and the variance between animals is B^2, then the correlation between two sperm from the same animal is as follows: $B^2/(B^2 + W^2)$. We calculated that W^2 for curvilinear velocity was 166 (μm/second)^2 and B^2 was 1446 (μm/second)^2. From the formula, we see that the correlation is 0.10. If such a correlation is ignored, estimates of variability will be too small and statistical tests will produce too many false-positive results (Haseman and Kupper, 1979; Miller, 1986, section 1.3).

These problems can be illustrated by considering what happens when we estimate a group mean. Suppose we have R animals and n_i sperm from the i^{th} animal, for a total of: $N = \sum_{i=1}^{R} n_i$. Let the j^{th} measurement from the i^{th} animal be denoted X_{ij}. If we use the sperm as the unit without considering the correlations among sperm from the same animal, we would simply use the overall mean of $\bar{X} = \sum_{i=1}^{R}\sum_{j=1}^{n_i} X_{ij}/N$ with the usual estimated squared standard error (ie, estimated variance of the mean): $\sum_{i=1}^{R}\sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2/(N[N - 1])$. It can easily be shown that the mean of this estimated variance is the true value minus a quantity that is greater than zero (unless the n_i's are all equal to 1 or B equals 0, neither of which is likely). Thus, the estimated variance of the group mean will be too small. This underestimation can be quite severe. The ratio of the incorrect variance to the true value is given by:

$$1 - \frac{(\Sigma n_i[n_i - 1])B^2/(N - 1)}{W^2 + \Sigma n_i^2 B^2/N}$$

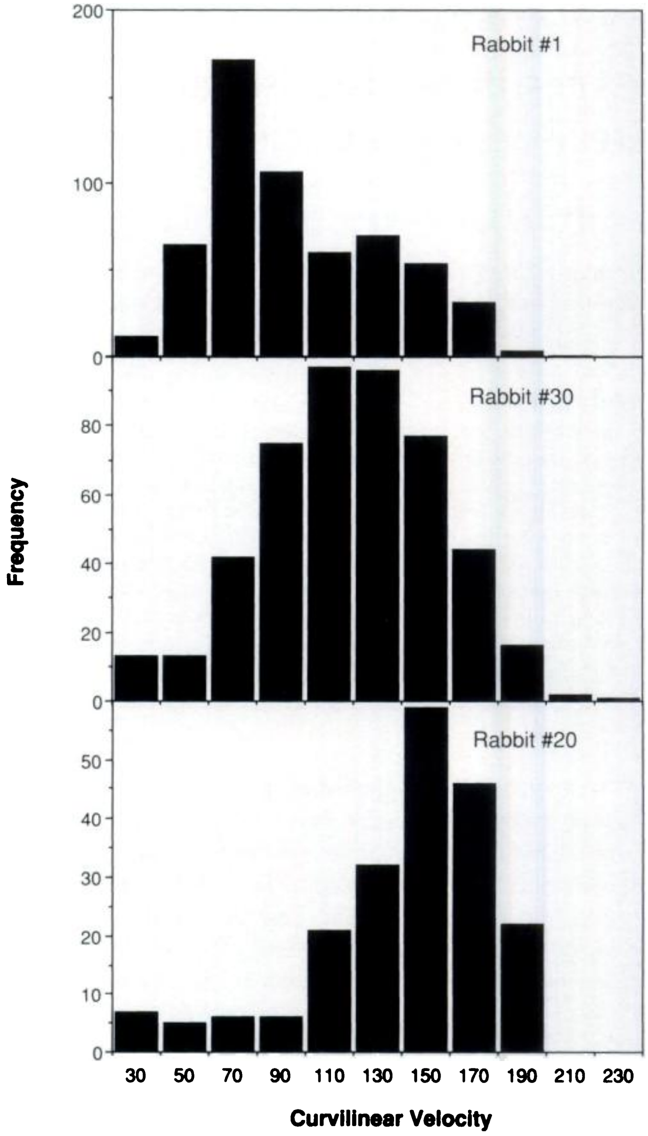

FIG. 1. Histograms of curvilinear velocity measurements of sperm from three untreated rabbits. Sperm from rabbit 1 tend to be relatively slow; those from rabbit 30 are about average; and those from rabbit 20 tend to be relatively fast.

If we make the simplifying assumption that the number of sperm per animal (n_i) are all equal, this formula becomes: $1 - B^2Rn(n - 1)/[(Rn - 1)(W^2 + nB^2)]$. If we also assume that the number of sperm per animal (n) is large, the approximate formula is as follows: $W^2/(W^2 + nB^2)$. The larger n becomes, the larger the relative underestimation of the variance of the group mean becomes. For the curvilinear velocity example, recall that W^2 was 166 (μm/second)^2 and B^2 was 1446 (μm/second)^2. If 50 sperm per animal were measured, the incorrect variance would be about 15% of the true variance. If $n = 200$, it would be only 4% of the true value. Even for the small $n = 10$, the incorrect variance would still be only about 47% of the true value.

Whenever variance estimates of means are too small, any test of a hypothesis concerning those means, such as testing for equality of means, will produce too many false-positive results. This occurs because we will be more certain of the means than we should and thus more ready to declare that two values are unequal. Accordingly, ordinary t tests, ordinary (ie, fixed-effect) analyses of variance (ANOVA), and ordinary regression techniques, all of which are based on estimates like those in the preceding paragraphs, will not be appropriate to these data. Nonparametric analogs such as the Wilcoxon or Kruskal-Wallis tests (Hollander and Wolfe, 1973) suffer from the same problem in these circumstances.

When analyzing data from individual sperm, the correlated structure of these data can and should be accommodated. Models and tests explicitly incorporating the two sources of variability (within animals and between animals within a group) are the most appropriate techniques. Mixed-model ANOVA is an example of such a procedure; it differs from ordinary ANOVA by explicitly incorporating random animal-to-animal variability in addition to treatment effects. Mixed-model analyses are dealt with in standard texts (Snedecor and Cochran, 1980; Sokal and Rohlf, 1981) and are available in standard statistical packages such as BMDP® (Dixon et al, 1988) and SAS® (SAS Institute Inc., 1985). Unfortunately, for discrete (categorical or yes/no) outcomes, fewer software packages are currently available.

Unequal Variances: Using the Whole Animal as the Unit

If statistics from whole animals (such as the mean velocity or the percentage of motile sperm) are analyzed, they will typically be independent of each other; thus, the problems described above will not arise. However, the number of sperm analyzed often differs from animal to animal, a difference that will lead to differences in variance. For example, in the rabbit study, eight fields were examined for each semen sample, and the total number of sperm with measurable velocity in those fields varied from 2 to 185. Clearly, a result based on 185 sperm is more accurate than one based on 2. Even if an attempt had been made to examine the same number of sperm per sample by examining more fields, it is unlikely that the samples with only a few sperm in each field would have had sufficient sperm. Since it may not be feasible to examine equal numbers of sperm per sample in other studies as well, awareness of the implications is needed.

One of the assumptions underlying most statistical procedures is equality of variances; this is true of both parametric and nonparametric procedures. The t test, for example, is based on the assumption that each measurement in both groups has the same variance. If the variance is the same within groups but differs between groups, we have what is known as the Behrens-Fisher problem. In this situation, the t test has been shown to be relatively robust to moderate inequality of variance in any design and to be relatively robust to any size inequality when the groups have the same numbers of animals (Miller, 1986, section 2.3.1). It is reasonable to expect that the same robustness will occur when the distribution of variances is not too different from group to group. If an exposure has changed sperm concentration dramatically, so that the number of sperm available in each group is quite different, classic procedures like t tests may not be appropriate. There are modifications available for the t test, such as Welch's approximation (Miller, 1986, section 2.3.3), and similar extensions to ANOVA also have been proposed (Miller, 1986, section 3.3.3). There is also some research indicating that nonparametric procedures are less affected by inequality of variance (Miller, 1986, section 2.3.3).

An alternate approach for dealing with unequal variances is to perform weighted analyses. The advantages of weighting are smaller variances for the estimates of the means and the availability of appropriate standard errors and tests. Variance estimation and testing in weighted analyses are discussed in standard texts (Draper and Smith, 1981, section 2.11). Let the measurement for the i^{th} animal be X_i and the weight be w_i. Both the usual unweighted average of $\Sigma X_i/R$ and the weighted average of $\Sigma w_iX_i/\Sigma w_i$ correctly estimate the group mean, ie, the mean of both averages is the true group mean. However, the weighted average will have a smaller variance if the proper weights are chosen. It can easily be shown that the best weight is the inverse of the variance of X_i, which is denoted $Var(X_i)$. If X_i is the average measurement over n_i sperm, then it follows that: $Var(X_i) = B^2 + W^2/n_i$. Thus, estimates of the variability within animals and the variability between animals within groups must be available to weight properly. Note that most computer programs that do weighted analyses assume that optimal weights are being used (so that variances and inverse weights are interchangeable) and that the variances are of the form $a_i\sigma^2$, where the a's are known constants and the common σ^2 is a variance parameter to be estimated. Since $Var(X_i)$ can be written as either $[1 + (W/B)^2/n_i]B^2$ or

$[(B/W)^2 + 1/n_i]W^2$, we need only specify the ratio W/B (and the n's). In summary, if the number of sperm per sample varies, summary measurements for the sample will have unequal variances. While small inequalities in variance will have little impact, especially when the number of animals per group is the same, larger inequalities should be addressed by modifications to the standard tests or by weighting.

Measurements Within Limited Ranges

The algorithms used to measure motion endpoints have their limitations. An example is velocity measurements. Since discrimination between sperm and other particles is based partly on a minimum threshold value for velocity, a slower-moving sperm will not have its velocity measured. In addition, since velocity measurements are based on comparing positions at multiple time points, a fast-moving sperm may already have disappeared from view at the later points and also will not be measured. Thus, sperm from both ends of the velocity distribution may be missing from the data. Various authors (eg, Knuth et al, 1987; Vantman et al, 1988; Toth et al, 1989) have shown how changes in the number of points examined or other machine settings determining the measurable velocities and other parameters can affect the mean values.

This truncation, however, also has implications for comparing groups measured at the same settings. The observed mean for a group will differ from the actual mean, with the amount of difference depending on where the limits of measurement are in relation to the distribution of velocities. Thus, even with fixed machine settings, the effect of truncation will potentially differ among groups. If an exposure affects mean velocities, the change between groups will seem less severe than it actually is. If an exposure affects variances, it may appear to affect the means. To illustrate these points, Table 1 shows some calculations based on the assumptions that the velocity distribution is normal (similar results would occur with other distributions) and that measurements can only be made between 20 and 250 μm/sec (the actual limits used for curvilinear velocity in the rabbit study). For simplicity, B^2 was assumed to be zero, but the point to be made will hold even if B^2 is not zero. For two values of the standard deviation and for a range of values for the mean, the apparent mean (ie, the mean of those sperm that have measurable velocities) is shown. The standard deviation of curvilinear velocity measurements (from randomly chosen sperm from randomly chosen animals) was about 40 μm/sec in the rabbit study; results for a standard deviation of 60 are also shown for comparison. The mean in the rabbit study was approximately 120 μm/sec. Note that for a mean of 135, the boundaries are symmetrically placed, and the true and apparent means are equal. In all other cases, the apparent mean is closer to 135 than is the true mean.

Table 1. *Comparison of true and apparent means when the range of measurement is restricted*

	Apparent mean	
True Mean	SD = 40	SD = 60
160	158.8	153.3
150	149.4	146.1
140	139.8	138.7
135	135.0	135.0
130	130.2	131.3
120	120.6	123.9
110	111.3	116.7
100	102.2	109.7
90	93.6	103.1
80	85.6	96.8
70	78.2	90.9
60	71.5	85.5
50	65.6	80.4

Entries for apparent mean are calculated under the assumptions (i) that the data are normally distributed with true mean and standard deviation as shown, and (ii) that measurements can be made only in the range of 20 to 250.

Table 1 shows that differences between groups are diminished. For example, if an exposure changed the true mean from 120 to 100, the apparent mean would only change by 18 or 14 (depending on the standard deviation) instead of 20. This shrinkage affects our ability to detect a change. Often, the effect will be minor, either because the shrinkage is small or because the true change is so large that even the shrunken apparent change is detectable. However, the possibility exists that effects will be (for practical purposes) undetectable because of this phenomenon; studies involving more animals would then be needed to detect the same actual change. Study designs for velocity or other endpoints that can only be measured within a limited range should be chosen with the awareness that shrinkage will occur.

Another type of problem would arise if we compared two groups of animals with the same mean but different variances. As shown by the difference between the two columns of Table 1, the apparent means for the groups would differ even though the only true difference is in the variance. For example, if two groups had the same true mean of 160 but different standard deviations (40 and 60), their means would appear to be 158 and 153. Thus, it is possible to be misled about the nature of the difference between groups and the magnitude of the difference.

While there is no way to know what is happening outside of the observed range, the shape of histograms can be suggestive. For illustration, two histograms are presented in Figure 2. In the upper panel, data for curvilinear velocity from the rabbit study are shown. There is little reason to suspect a problem; since there are few sperm near the thresholds, there are probably few sperm outside of the

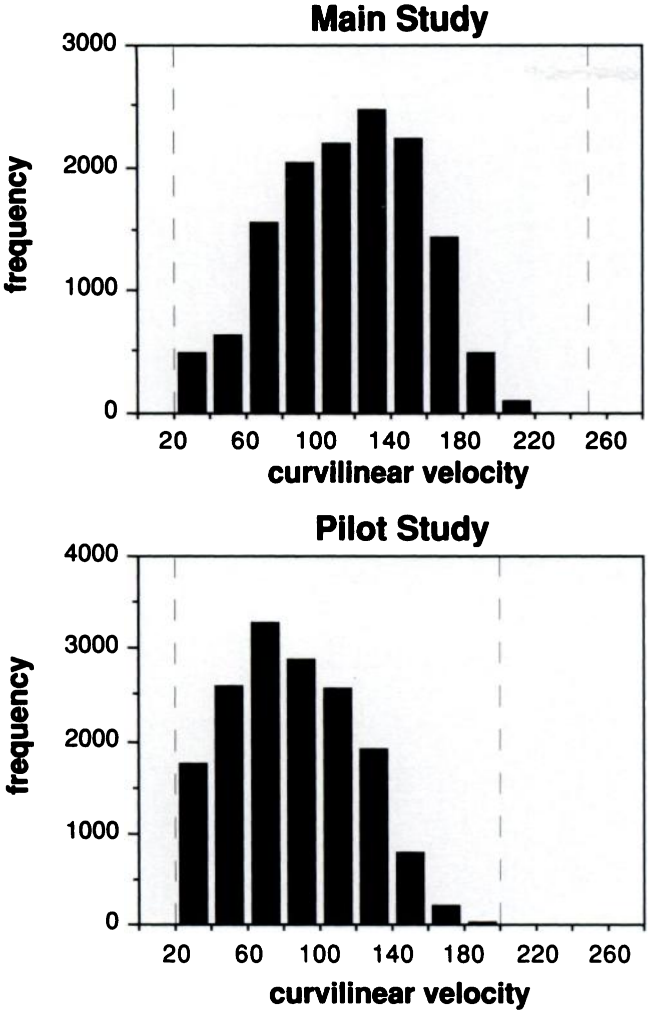

FIG. 2. Two distribution histograms of sperm curvilinear velocity generated by CASA from untreated rabbits in a pilot study (10 rabbits, semen samples collected weekly from each rabbit for 6 to 21 weeks, 15,976 sperm measured in all, lower panel) and the main study (42 rabbits, 6 weeks, 13,674 sperm, upper panel). Ordinate represents number of sperm. The larger number of sperm near the lower limit of the machine settings in the pilot study suggest that there might have been sperm moving more slowly than the lower velocity limit that were not counted for this study.

thresholds. In the lower panel, data from an unpublished pilot study using different rabbits, experimental conditions, and machine settings are shown. The pilot study had the same lower limit of measurement and the upper limit was 200. It seems probable that most of the slower sperm in the pilot study were not measured since there are a large number of sperm near the lower threshold. Thus, it is plausible that there would be more difficulty in detecting any decrease under these conditions. This would mean that larger studies with more animals would be needed unless the conditions (eg, media, threshold) could be changed.

In summary, the existence of thresholds (beyond which measurements cannot be made) distorts the information we can obtain. The usual effect will be to shrink differences between groups, necessitating larger studies than would be needed if the thresholds did not exist. If possible, pilot data should be examined when planning studies to see whether such problems are likely, so that appropriate changes can be made in machine settings or study designs.

Artificial Discreteness of Measurements

Some quantities that are truly continuous have discrete measurements because of the way they are obtained. Beat-cross frequency can, in theory, be any value; however, it is calculated by counting the number of times the head centroid crosses the average path and dividing by a factor related to the number of frames tracked to convert it into hertz (Amann, 1989). Since the number of crossings and the number of frames must be integers, the measurements are discrete, ie, only a limited number of values can occur. To illustrate, for the 8,463 measurements obtained from the rabbits, only 52 distinct values occurred.

The distribution of all the beat-cross frequency measurements produces the unusual multi-modal histogram shown in Figure 3. To examine this more closely, the counts of all the values obtained are shown in Table 2. This histogram is actually a composite of nine separate histograms. Sperm were tracked for a maximum of 15 frames; at least seven frames were required for beat-cross frequency to be calculated. Separating the original histogram by the number of frames tracked gave the grouping in Table 3. Each of these nine separate histograms is smooth and unimodal.

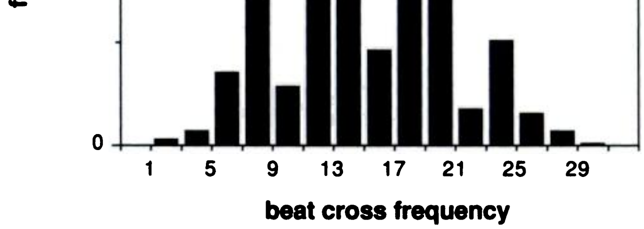

FIG. 3. A multi-modal distribution histogram of sperm beat-cross frequency generated by CASA from untreated rabbits (42 rabbits, semen samples collected weekly for 6 weeks, 8,463 sperm measured in all). Ordinate represents number of sperm. This distribution is a composite of several distributions that are broken down in Table 3.

Table 2. *Data used to generate a histogram of beat-cross frequency*

Value	Number	Value	Number
2.92	25	16.37	149
3.23	5	16.67	221
3.61	8	17.14	383
4.09	17	17.53	843
4.71	34	18.06	130
5.56	86	18.86	152
5.84	99	19.39	85
6.46	19	20.25	222
6.75	148	20.46	522
7.22	40	21.67	43
8.18	57	22.22	102
8.57	342	22.62	27
8.77	288	23.34	273
9.43	126	23.38	155
9.69	51	23.57	48
10.83	107	24.55	28
11.11	219	25.28	5
11.67	586	25.72	104
11.69	635	25.85	10
12.27	142	26.30	36
12.92	103	27.00	39
13.50	343	27.78	14
14.14	196	28.29	9
14.45	116	28.64	5
14.61	972	29.08	1
16.16	89	29.22	4

Number of sperm ("number" column) having each of the 52 distinct beat-cross frequency values ("value" column) that occurred in the data. Data are from 42 untreated rabbits sampled weekly for 6 weeks. Note lack of smoothness and existence of multiple modes.

There is no standard theory that is strictly appropriate for such a mixture of distributions. For most purposes, treating this mixture as normal should produce no serious problems; the approximate symmetry and the lack of outliers should ensure that standard analytic procedures do not behave much differently than theory predicts. However, while not likely, it is possible that groups could appear different simply because they tended to be tracked for different numbers of frames. As an extreme example, suppose that all the sperm in one group of animals were tracked for seven frames, while all the sperm in an otherwise identical group were tracked for nine frames. All the values for the first group would be either 11.67 or 23.34 and all the values for the second group would be 6.75, 13.50, 20.25, or 27.00. The distributions would thus look different even though the groups did not actually differ in their beat-cross frequencies. Comparing the histograms of the groups would indicate whether this occurred. Investigators should be aware of this effect when interpreting results.

Other Considerations

In addition to the issues mentioned above, there are, of course, others to be considered. Many of these are common

Table 3. *Rearranged data used to generate a histogram of beat-cross frequency*

Tracked frames	Beat-cross frequency value	Number of sperm
15	2.92	25
	5.84	99
	8.77	288
	11.69	635
	14.61	972
	17.53	843
	20.46	472*
	23.38	155
	26.30	36
	29.22	4
14	3.23	5
	6.46	19
	9.69	51
	12.92	103
	16.16	89
	19.39	85
	22.62	27
	25.85	10
	29.08	1
13	3.61	8
	7.22	40
	10.83	107
	14.45	116
	18.06	130
	21.67	43
	25.28	5
12	4.09	17
	8.18	57
	12.27	142
	16.37	149
	20.46	50*
	24.55	28
	28.64	5
11	4.71	34
	9.43	126
	14.14	196
	18.86	152
	23.57	48
	28.29	9
10	5.56	86
	11.11	219
	16.67	221
	22.22	102
	27.78	14
9	6.75	148
	13.50	343
	20.25	222
	27.00	39
8	8.57	342
	17.14	383
	25.72	104
7	11.67	586
	23.34	273

The same data as in Table 2 is presented, but rearranged by number of frames (*arbitrarily allocated to the two places this value appears). Unlike Table 2, each separate piece is smooth and unimodal; see text for full discussion.

to a much wider class of studies than those of sperm movement. We briefly discuss three of these in this section.

Non-normality of Measurements

Although many standard procedures are much more sensitive to the existence of correlation in the data than to lack of normality, the shape of the distribution is still of concern. Various authors (eg, Ratcliffe et al, 1987, Toth et al, 1989) have noted the non-normality of some motion measurements. When faced with non-normal distributions, one can use nonparametric procedures that do not rely on the assumption of normality, or one can transform the measurement.

The problem and the benefit of transformations can be illustrated with the rabbit data. For example, linearity was severely skewed in these data (Fig 4). By applying a logistic transformation (log[linearity/(10-linearity)]), a distribution closer to normal was obtained. It should be noted that if the linearity is at its maximum value of 10, the transformation will be infinite; to avoid this, values of 10 should be replaced by, for example, 9.99. The mean amplitude of lateral head displacement was also skewed, although not as severely. As shown in Figure 5, a logarithmic transformation worked reasonably well. There are, for each endpoint, an infinite number of transformations that could theoretically be considered. However, attention is usually confined to a group of relatively simple and interpretable ones such as those illustrated here. It is possible that two transformations will work equally well; in that case, either can be chosen. Transformations are discussed in standard texts (eg, Snedecor and Cochran, 1980; Sokal and Rohlf, 1981).

An alternative to transformation is the use of nonparametric procedures. Typical ones include the Wilcoxon-

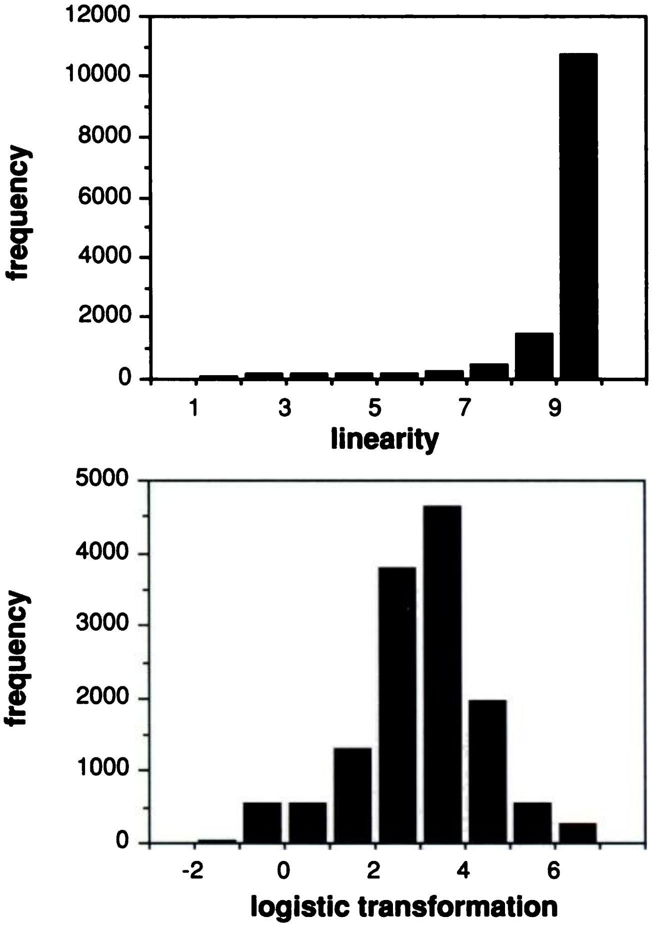

FIG. 4. Data for linearity before (upper panel) and after (lower panel) transformation to a more normal distribution. These data were obtained by CASA from untreated rabbits (42 rabbits, semen samples collected weekly for 6 weeks, 13,674 sperm measured in all). Ordinate represents number of sperm. A logistic transformation makes the skewed distribution more normal.

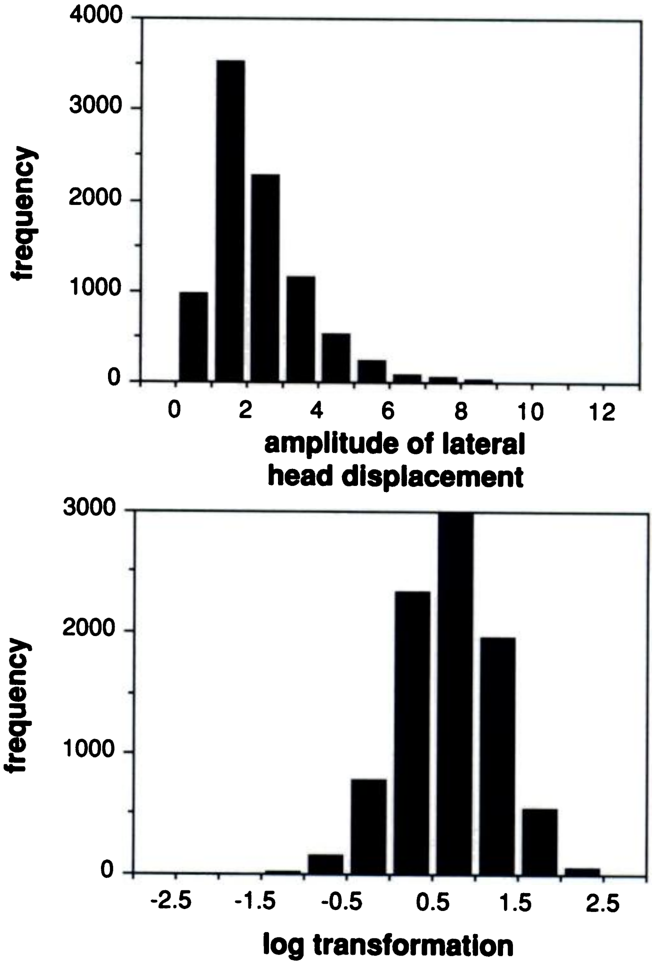

FIG. 5. Data for mean amplitude of lateral head displacement before (upper panel) and after (lower panel) transformation to a more normal distribution. These data were obtained by CASA from untreated rabbits (42 rabbits, semen samples collected weekly for 6 weeks, 8,842 sperm measured in all). Ordinate represents number of sperm. A logarithmic transformation makes the skewed distribution more normal.

Mann-Whitney test as an analog to the t test and the Kruskal-Wallis test as an analog to one-way ANOVA. These are described in many standard texts (eg, Hollander and Wolfe, 1973, sections 4.1 and 6.1). These procedures are appropriate to relatively simple designs where there are two or more groups of animals with a single observation per animal. There are no standard nonparametric procedures appropriate for more complex designs, such as those with multiple samples and covariates from each animal.

In summary, nonparametric techniques have the advantage of avoiding the somewhat arbitrary choice of a transformation and are often preferred for that reason. However, for some designs, there may be no suitable nonparametric technique available, so that transforming and doing, for example, a complex ANOVA may allow a more appropriate analysis.

Multivariate Methods

The various measures describing motion are not independent of each other since they describe different facets of the same movement. For example, in the rabbit data, the correlation (Spearman rank) between curvilinear velocity and straight-line velocity was 0.94. The correlation of the mean amplitude of lateral head displacement (ALH) with curvilinear velocity was 0.57 and with straight-line velocity was 0.41. ALH was negatively correlated with linearity, with a value of −0.49. Linearity is the ratio of straight-line to curvilinear velocity (multiplied by 10); thus, since one is calculated from the other two, these three variables are certainly not independent, although the association is not readily expressed by correlation coefficients. In view of these dependencies, multivariate methods that look at several variables simultaneously may be a better approach in some cases. For example, Morales et al (1988) used multivariate ANOVA to compare fertile men to infertile patients on a variety of endpoints simultaneously. As another example, if the purpose of the study is to separate sperm into groups (normal vs. abnormal, hyperactivated vs. nonhyperactivated), various discriminant analysis or clustering techniques may prove useful (Snedecor and Cochran, 1980).

Choice of Statistic

It is most common and, ordinarily, most useful to describe an endpoint by its mean or median. This may not, however, always be the most appropriate choice. It is possible, for example, that a treatment or exposure may not change the mean but may increase the variance. (Methods for testing inequality of variance are discussed in Miller, 1986, chapter 7.) It is also possible that most sperm or most animals will be unaffected, but some will be affected substantially, leading to a change in the shape of the distribution. This could be characterized by looking at percentiles (eg, Toth et al, 1989) or the fraction of the distribution below a certain value. One should always compare the distributions in the groups to make sure that the chosen statistic captures the features of interest.

Summary

The new computer-assisted measurement technologies are powerful tools, but their limitations should be recognized as well. The inability to measure outside a restricted range and the inability to assign more than a limited number of values can have implications for the design and interpretation of studies. It should always be kept in mind that the things being measured are only approximations of what we really want to know.

Strengths and limitations of statistical techniques should also be kept in mind; any technique will be appropriate under some conditions and not under others. Most commonly used procedures require that observations be independent. This will not be the case when multiple sperm from the same sample are measured; in these cases, techniques that account for the correlation are needed. Many procedures require that all observations have the same variance. This may not occur when measurements are based on differing numbers of sperm. In this instance, the severity of the inequality should be assessed and possible adjustments considered. Some procedures require that observations be normally distributed. Many measurements are not, and either transformations or nonparametric procedures are needed. Univariate techniques may be more appropriate in some cases; in others, multivariate may be more appropriate. Similarly, measures of the center of the distribution may be better in one situation and measures of the tails may be more suitable in another. Before using any statistical tool, the data should be examined to see whether the tool is appropriate.

Acknowledgments

We thank Steve Schrader and Terry Turner of the National Institute for Occupational Safety and Health for their help in generating the data used for illustration throughout this article.

References

Amann RP. Can the fertility potential of a seminal sample be predicted accurately? *J Androl*. 1989;10:89–98.

Dixon WJ, Brown MB, Engleman L, Hill MA, Jennrich RI. *BMDP Statistical Software Manual*. Berkeley, CA: University of California Press; 1988.

Draper NR, Smith H. *Applied Regression Analysis*, 2nd ed. New York: John Wiley and Sons; 1981.

Haseman JK, Kupper LL. Analysis of dichotomous response data from certain toxicological experiments. *Biometrics*. 1979;35:281–293.