

Methodological Issues in the Analysis of Human Sperm Concentration Data

NANCY G. BERMAN,* CHRISTINA WANG,* AND C. ALVIN PAULSEN†

From the *Departments of Pediatrics and Medicine, Harbor-UCLA Medical Center, Torrance, California; and †Department of Medicine, University of Washington, Seattle, Washington.

ABSTRACT: We examined two methodological issues in the analysis of sperm concentration data using a large database of sperm concentrations in healthy men that were collected at the University of Washington. We showed that the raw data were skewed and that log transformation should be used to assure that the data meet the assumptions underlying most statistical estimation and testing procedures. We also addressed the issue of the great variability in sperm concentrations within a single individual and the necessity and utility of multiple sampling to reduce variance. We conclude that log-trans-

formed data should be used for statistical analysis of sperm concentration and recommend that such analyses be based on the geometric mean of several samples from each subject to reduce variability, increase accuracy of estimation, and improve statistical power. This is particularly important when the objective is to detect small but important differences or subtle effects.

Key words: Sperm concentration, log transformation, multiple sampling, variation, statistical power.

J Androl 1996;17:68-73

When spontaneous or induced changes in human spermatogenesis are analyzed, it is important that the particular characteristics of semen parameters be addressed in the study design and statistical methods. It is recognized that healthy men may show large changes in sperm concentration in a short time period in the absence of any intervention, so that data based on a single sample may not be a reliable or precise estimate of sperm production. Moreover, the distribution of sperm concentration is highly skewed to lower concentrations, so that standard techniques of analysis cannot be used without transformation of the data. Our laboratory has addressed these issues and suggested improved methods to achieve more accurate estimates (Paulsen et al, 1984). In this report, we will clarify the methods of sperm concentration data analysis utilizing a large data base from healthy men to illustrate their effectiveness.

Supported in part by National Institutes of Child Health and Human Development grant P-50 HD 12629; Contraceptive Research and Development Program grant CSA-88-024, Eastern Virginia Medical School under a cooperative agreement with the USAID (CCP-3044-A-00-2015-00); and National Institutes of Health grant RR-00425 Clinical Research Center, Harbor-UCLA Medical Center. The views expressed by the authors do not necessarily reflect the views of USAID and CONRAD.

Correspondence to: Dr. Nancy Berman, Department of Pediatrics, Harbor-UCLA Medical Center, Box 446, 1000 West Carson Street, Torrance, California 90509.

Received for publication May 5, 1995; accepted for publication September 19, 1995.

Materials and Methods

The Database

The database consists of sperm concentration obtained from 510 normal healthy males participating in 14 different studies over a period of 20 years, from December, 1972 to July, 1993. Subjects included either control subjects who received no drugs or other interventions or men who submitted serial semen samples prior to drug exposure. Subjects were recruited from the urban population in the Seattle, Washington area and were mostly Caucasian. They ranged in age from 18 to 52 years (mean age 28 years) and in weight from 100 to 275 pounds (mean weight 171 pounds). The men gave no history of infertility problems or chronic systemic medical illness. They had normal annual physical examinations, including examination of external genitalia and testes size. They reported no alcoholism, heavy tobacco use (>20 cigarettes per day), or recreational drug abuse and were not taking any regular medication. Each revealed normal blood chemistries. Subjects were asked to refrain from sexual activity for 2-7 days before each sample collection. Sperm concentration was measured using the Coulter counter, as described previously (Gordon et al, 1967).

Each man submitted multiple samples, at intervals of at least 2 weeks. The median number of samples for a subject over his entire time in the study was 6; the range was 4-30. Only 7% of the subjects had only 4 samples, 18% had 5 samples, and 11% had >10 samples. The median difference in time between the earliest and latest sample for a subject was 81 days, with a range from 42 to 630 days. In the analysis described below we used the earliest sample for each subject as the "single" sample, to test distributions and to calculate the sample variances that would result from a study based on just one sample from each subject. To show the effect of multiple sampling, we used all samples

collected during the first 3 months to calculate the geometric mean and variance and compared the reduction in variance due to using this multiple sampling protocol to those from the "single" sample. Although the sampling period for many subjects was longer than 3 months, this limit provided a more consistent number of samples per subject. The maximum number of samples any subject had in the 3-month period was seven, the median was six, and only 4.7% of the subjects had fewer than four samples.

Data Transformations

Data transformations of sperm concentration are required to stabilize the variance of the data sample and to assure a statistically valid analysis. When looking at the distribution of untransformed biological data in a plot such as a frequency histogram there is often an obvious lack of symmetry. Frequently the data appear to "bunch up" near the lower end of values and space themselves out at the higher end. The data are not normally distributed, and if one were to take samples at different levels, the variance would increase as the mean increases. The arithmetic average does not estimate the central tendency of the data. Data transformation is required to create a new variable that has a normal distribution and a variance that is independent of the mean. For sperm concentration, which has these characteristics, the simple logarithmic transformation $y = \log(x)$, where x is the raw data value, has been shown to meet these criteria (Paulsen et al, 1984; Wang et al, 1985; Nagao et al, 1986; Douglas et al, 1988; Bromwich et al, 1994; Farrow, 1994; Auger et al, 1995).

Either the natural or base 10 log may be used, because the difference is simply multiplication by a constant. The choice of one or the other is usually dictated by the context of the results. In the examples we use the natural log. The true mean, i.e., central value of x , is simply the geometric mean of x , computed as the antilog of the mean of $\log(x)$. The true variance of x is more complex to calculate because variation is a mixture of addition and multiplication (Johnson and Kotz, 1970). (The formula is given in Appendix I, equation (3).) Confidence intervals for the mean are calculated by computing the confidence interval for the log mean and standard deviation (SD), then taking the antilogs of the endpoints. This confidence interval will not usually be symmetric about the mean.

Multiple Sampling

Biological data, particularly sperm concentration values, often show a high degree of variability, even in healthy, normal individuals despite the absence of any particular intervention. It is assumed that the values for an individual fluctuate around a constant mean with some variation, referred to as the within-subject or intra-subject variation. A single sample may not be an accurate estimate of this mean for many applications. Therefore it is preferable to take several samples from each subject and use the average of these samples as the representative value for the subject. Intuitively, one can see that this would give a better estimate of the true value for the subject and, by extension, better estimates of group means for testing hypotheses.

Statistically, taking the average of several samples, say n of them, reduces the within-subject variation by a factor of $1/n$;

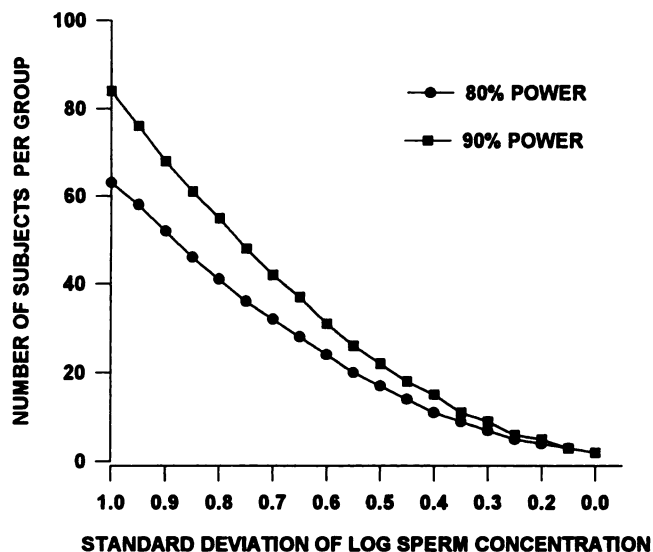


FIG. 1. Sample size requirements for an independent group's t -test as a function of the standard deviation of log sperm concentration. Curves are shown for 80% and 90% power to detect a difference in sperm concentration between two groups of $0.5 \log(1 \times 10^6/\text{ml})$.

thus the expected error of this individual average is much less than that of a single sample. The concept carries over to the mean of a group. The total variance in a group of subjects is the weighted sum of the variation between subjects plus the within-subjects variation (Dunn and Clark, 1987; Cooper et al, 1991). Reducing the within-subjects variation therefore reduces the total variation for the group. The increased precision can be seen as a reduction in the length of confidence intervals, which are generally proportional to the SD, which is the square root of the variance. The statistical equations describing this concept are given in the Appendix II, equations (4)–(8).

Reduction of the total variation increases the power of any given study. Power increases as the study size and/or size of the assumed effect increases; it decreases as the variation in the data increases. Investigators are encouraged to determine the power of a study as part of their planning, usually by choosing an appropriate sample size for given difference and required power. If the variation in the data is decreased by multiple sampling, the number of subjects required may be reduced or the power of the study to detect small but clinically important effects may be increased. Figure 1 is a plot of required sample size as a function of the SD to achieve power of 80% and 90% for a difference in mean log sperm concentration of $0.5 \log(1 \times 10^6/\text{ml})$.

Results

Data Transformations

We used the Shapiro and Wilk test (Shapiro and Wilk, 1965) to test for a normal distribution in each of the 14 studies in our database, using the a single sample, the earliest, for each subject. The hypothesis was rejected at the 0.01 level in all but four of the studies. When we ran

the same test on the (natural) log of the first sample, there were no studies that rejected the hypothesis. Before transformation the SD was highly dependent on the mean, as shown by a correlation of 0.71. The correlation between the mean and SD of the log-transformed data was markedly reduced, only -0.18 , showing that the log transformation was the appropriate choice to produce a distribution with a stable variation that is independent of the mean. Therefore we concluded that the distribution of sperm concentrations was lognormal, and the log of each sample was used for subsequent analysis.

Figure 2 shows the distribution of raw and log values for the first sample value in the 510 subjects. The arithmetic average value was $84.24 \times 10^6/\text{ml}$; the geometric mean was $64.85 \times 10^6/\text{ml}$. The 95% confidence interval for the mean of the data in Figure 2, computed using the log-transformed data, is $60\text{--}68 \times 10^6/\text{ml}$. One can see that this transformation has the effect of stretching out the data at the lower values and compacting them at the higher values, thus stabilizing the data and creating a more symmetric distribution.

Multiple Sampling

Figure 3 illustrates the reduction in total variation of sperm concentration due to having multiple samples per subject for any population. Figure 3a shows the percent reduction in total SD as a function of the within-subjects SD, assuming the total variation for a single sample is $1 \log(1 \times 10^6/\text{ml})$. Separate curves are shown for 2, 3, 4, 6, or 10 samples per subject. Figure 3b shows the percent reduction in total SD, due to the use of multiple samples per subject, as a function of the number of samples per subject for within-subject SDs of 0.6, 0.7, and 0.8 $\log(1 \times 10^6/\text{ml})$. For example, if the variation of sperm concentration were $1 \log(1 \times 10^6/\text{ml})$, and we assumed the within-subjects variation was 0.6 $\log(1 \times 10^6/\text{ml})$, using the geometric mean of four samples would give us a 25% reduction in total variation. Using 6 samples resulted in a 29% reduction in total variation and 10 samples a 32% reduction. The equations used to develop these curves are given in Appendix II, equations (4)–(8). In our database of 510 subjects, we computed the within-subjects mean and SD of sperm concentration using samples from the first 3 months. All computations were based on log-transformed data. The values of within-subjects SD ranged from 1 to 159% of the mean, with a median value of 25%. This showed that there could be considerable variation within an individual. We compared the total variation of sperm concentration in all subjects using a single sample (the first) to the total variation using the average value from the first 3 months. Use of the 3-month average resulted in a 19% reduction in the total SD. When we ran this analysis in each study, the amount of reduction ranged from 6 to 37%, with a mean of 19%.

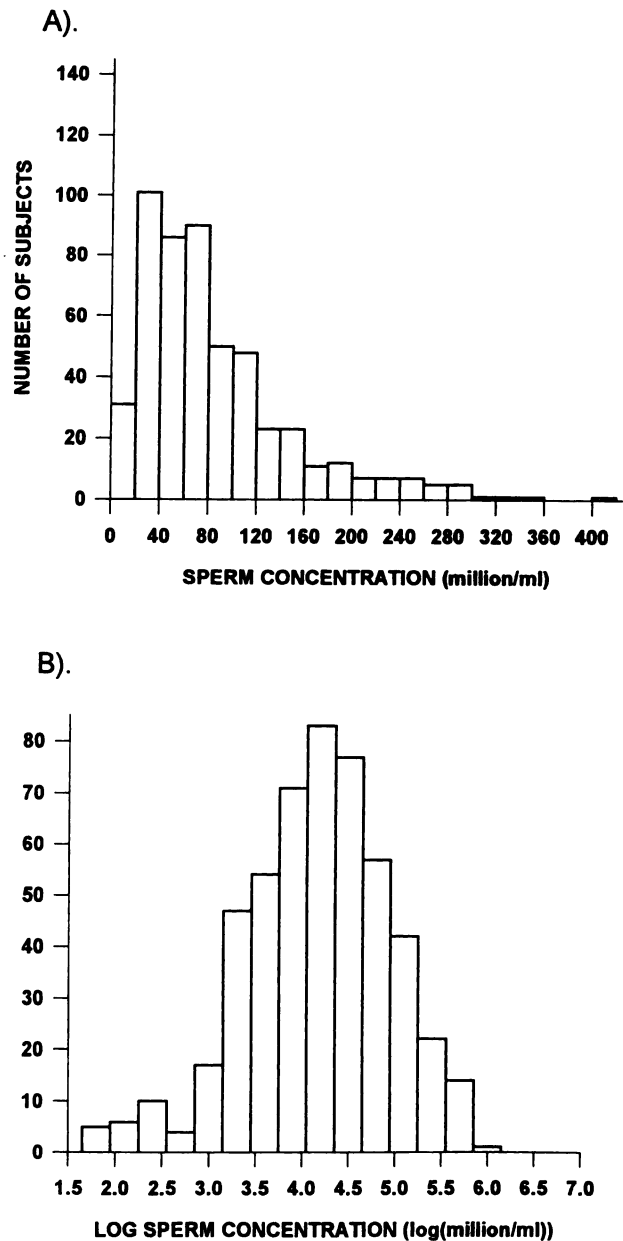


FIG. 2. Distribution of sperm concentration and log sperm concentration from 510 normal men. (A), Distribution of raw values of sperm concentration. (B), Distribution of the log-transformed values from A.

Discussion

We examined two issues in the analysis of sperm concentration data and illustrated them with samples from a large database. The first issue, the use of log-transformed data in analysis, should not be controversial. Bartlett (1947), in a definitive publication, set out four criteria that would ideally be met by sample data: 1) that the variance be constant and independent of the mean level, 2) that the transformed variable be normally distributed,

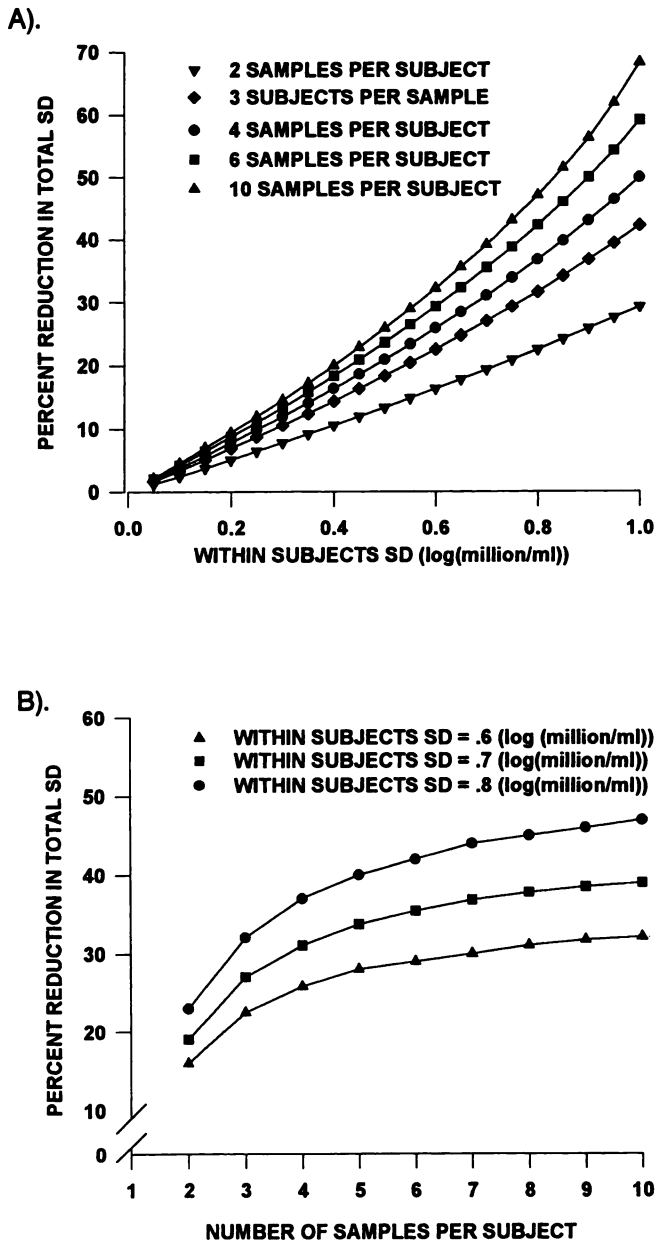


FIG. 3. Percent reduction in variation of sperm concentration due to multiple sampling. (A), Percent reduction in total SD as a function of within-subjects variation for 2, 3, 4, 6, and 10 samples per subject. (B), Percent reduction in total SD as a function of number of samples per subject for within-subjects variation of 0.6, 0.7, and 0.8. It is assumed that the total SD for a single sample is $1 \log(1 \times 10^6/\text{ml})$ in both panels.

3) that the transformed variable be an efficient estimator of the true mean level, and 4) that real effects should be linear and additive. The first assumption is required for the use of analysis of variance (ANOVA) models; that is that the groups being analyzed have equal variance. The second assumption, that the data follow the normal distribution, is basic to many procedures. Distributions of test statistics, such as the *t* statistic and the *F* statistic, are

mathematically derived from the fact that these statistics are functions of normally distributed variables. This means that if the sample data are not from a normally distributed population, then, although test statistics can be computed, the probability statements about the results are not necessarily true. If assumption 3 is violated, then the arithmetic average may not be a reasonable estimate of the true central tendency of the population, and confidence intervals may be misleading. Additionally, in an ANOVA framework, effects are assumed to be additive, and a transformation may be required to correct for non-additivity (assumption 4). When the variation of the raw data is proportional to the mean, as is true in most samples of sperm concentration, the log transformation satisfies all of these criteria. Moreover, the method is easy to understand, easy to perform with a calculator or computer, and not affected by extreme observations (Berry, 1987). Our examination of the data from 510 men healthy validates this theoretical observation, and it has been shown to be true in other data sets (Paulsen et al, 1984; Wang et al, 1985; Nagao et al, 1986; Douglas et al, 1988; Bromwich et al, 1994; Farrow, 1994; Auger et al, 1995). If the inclusion criteria for a sample restrict the range of sperm concentrations or, alternatively, an extremely heterogeneous group of subjects is included, then the distribution may not be lognormal. The investigator can evaluate his or her sample by examining the relationship between the mean and variance of the raw data and using the Shapiro and Wilk test (Shapiro and Wilk, 1965) to test the distribution of the raw and transformed data.

There is some disagreement between investigators whether the actual arithmetic average (arithmetic mean) or the anti-log of the average log value, i.e., the geometric mean, should be reported in publications. We recommend the use of the geometric mean, which provides a closer estimate of the true center of the distribution. Confidence intervals must be computed from the log-transformed data and then converted to anti-logs; otherwise the probability statements associated with them are false.

With respect to the second issue, we have also demonstrated the advantages of using the geometric mean of multiple samples from individuals as the analysis variable. We have shown that the variation of a sample population is reduced when multiple samples are used and that this leads to narrower confidence intervals for the parameters and more power for any testing hypotheses. We have substantiated the theoretical development with illustrations from our database, where the use of multiple samples produced a 19% reduction in SD. This means that if the total variation in a sample were $1 \log(1 \times 10^6/\text{ml})$ unit, then use of six to seven samples would reduce it to $0.81 \log(1 \times 10^6/\text{ml})$ units. This results in narrower confidence intervals for group means and more powerful tests. We have also shown that the individual variation

in normal men can be very high, so that multiple samples are required to provide a precise stable value for individual subjects. Other investigators have recognized the utility of multiple samples for better data accuracy (Sherins et al, 1977; Paulsen et al, 1984; Nagao et al, 1986; Douglas et al, 1988; Cooper et al, 1991). Baker (1989) notes that, in designing clinical trials to test the efficacy of an agent to treat male infertility, regression to the mean is a potential problem in the analysis of sperm parameters. Newell and Simpson (1990) note that use of multiple samples will correct for this effect. It is also recognized that tight control of the abstinence time, e.g., 3–4 days, will reduce the variability of sperm parameters (Freund, 1962; Amann and Howards, 1980). In practice, this is sometimes difficult to achieve. Use of multiple samples and allowing a wider abstinence interval, e.g., 2–7 days, as recommended by the World Health Organization (1992), would also reduce variance. A single sample may be sufficient under exceptional circumstances, e.g., if a catastrophic effect such as induced azoospermia is expected. However, for most applications, if for some reason it is impossible to obtain multiple samples from each subject, then it should be recognized that less reliable estimates are used when the results of his studies are interpreted.

Figure 3 may be used to estimate how many samples per subject are required for a particular study. If the values of within-subject SD are <30% of the total variation, the reduction in variation for 3, 4, 6, or 10 samples is similar. There is noticeably less reduction in variance for only two samples, even for a small within-subjects SD. For within-subject SD >30% there is not much gain for more than six samples. Because the assumed total SD for this chart is 1, these values on the x axis can be interpreted as the within-subjects variation expressed as percent of the total variation. If the investigator can estimate the within-subject SD as a percent of the group SD, either from other studies or pilot data, then he or she can determine the best number of subjects per samples for required power and/or accuracy. If prior information is not available, then six samples per subject is an optimal choice for most studies, and three samples per subject may be considered the minimum acceptable number to provide significant reduction in variation.

We have not addressed the analysis of longitudinal studies, in which time itself is an important effect. Although the multiple samples will be obtained over a time interval that may span several weeks, it should be considered as a single period with no time effects. Indeed, it is important that there are no events or changes in the sampling period that are known to influence the data.

In summary, based on our data and those of others, we strongly advocate the use of the log transformation in the analysis of sperm concentration data. We also recommend that the analysis be based on the average of multiple sam-

ples from each subject, rather than a single value, and that confidence intervals be reported rather than simple standard deviations.

Acknowledgments

We gratefully acknowledge the help of Connie Pete and Elaine Rost of the University of Washington, who performed the laboratory analysis over the many years of study and organized the data for these analyses.

References

- Amann RP, Howards SS. Daily spermatozoal production and epididymal spermatozoal reserve of the human male. *J Androl* 1980;124:211–215.
- Auger J, Kunstmann JM, Czyglik F, Jouannet P. Decline in semen quality among fertile men in Paris during the past 20 years. *N Engl J Med* 1995;332:281–285.
- Baker HWG. Development of clinical trials in male infertility research. In: Serio M, ed. *Perspectives in Andrology*. New York: Raven Press; 1989:367–374.
- Bartlett MS. The use of transformations. *Biometrics* 1947;3:39–52.
- Berry DA. Logarithmic transformations in ANOVA. *Biometrics* 1987; 43:439–456.
- Bromwich P, Cohen J, Stewart I, Walker A. Decline in sperm counts: an artefact of changed reference range of “normal”? *Br Med J* 1994; 309:19–22.
- Cooper TG, Jockenhovel F, Nieschlag E. Variations in semen parameters from fathers. *Hum Reprod* 1991;6:859–866.
- Douglas JM, Davis LG, Remington ML, Paulsen CA, Perrin EB, Goodman P, Conner JD, King D, Corey L. A double-blind, placebo-controlled trial of the effect of chronically administered oral acyclovir on sperm production in men with frequently recurrent genital herpes. *J Infect Dis* 1988;157:588–593.
- Dunn OJ, Clark VA. *Analysis of Variance and Regression*. 2nd ed. New York: Wiley; 1987.
- Farrow S. Falling sperm quality: fact or fiction? *Br Med J* 1994;309:1–2.
- Freund M. Interrelationships among the characteristics of human semen and factors affecting semen specimen quality. *J Reprod Fertil* 1962; 4:143–148.
- Gordon DJ, Herrigel JE, Moore DJ, Paulsen CA. Efficacy of Coulter counter in determining low sperm concentrations. *Am J Clin Pathol* 1967;47:226–228.
- Johnson NL, Kotz S. *Continuous Univariate Distributions*. Boston: Houghton Mifflin Company; 1970.
- Nagao RR, Plymate SR, Berger RE, Perrin EB, Paulsen CA. Comparison of gonadal function between fertile and infertile men with varicoceles. *Fertil Steril* 1986;46:930–933.
- Newell D, Simpson J. Regression to the mean. *Med J Aust* 1990;153: 166–168.
- Paulsen CA, Enzmann GD, Bremner WJ, Perrin EB. Effects of cimetidine on reproductive function in men. In: Cohen S, ed. *Update: H2 Receptor Antagonists*. Proceedings of an International Symposium, Royal Society of Medicine in London, England, 1983. Supplement to *Drug Therapy*. New York: Biomedical Information Corporation; 1984: 169–178.
- Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). *Biometrika* 1965;52:591–611.
- Sherins RJ, Brightwell D, Sternthal PM. Longitudinal analysis of fertile and infertile men. In: Troen P, Nankin HR, eds. *The Testis in Normal and Infertile Men*. New York: Raven Press; 1977:473–488.
- Wang C, Chan SYW, Leung A, Ng RP, Ng M, Tang LCH, Ma HK, Tson WL, Kuan M. Cross-sectional study of semen parameters in a large group of normal Chinese. *Int J Androl* 1985;8:857–874.

World Health Organization. *Laboratory Manual for the Examination of Human Semen and Semen-Cervical Mucus Interaction*. Cambridge, England: Cambridge University Press; 1992.

Appendix I

Means and Variance of Data With a Lognormal Distribution

A variable x is said to have a lognormal distribution if

$$y = \log(x) \quad (1)$$

has a normal distribution.

Either the natural or base 10 log (or any other base) may be used. If μ and σ^2 are the mean and variance of y , respectively, then the true mean value of x is simply the geometric mean of x

$$\mu_x = \exp(\mu), \quad (2)$$

where $\exp(t)$ is the irrational number e raised to the power t . The variance of x is given by

$$\sigma_x^2 = \exp(2\mu) \cdot \exp(\sigma^2) \cdot [\exp(\sigma^2) - 1]. \quad (3)$$

If logs to base 10 are used, then substitute 10 for e when taking powers in equations (2) and (3).

Appendix II

Reduction in Variation Due to Multiple Sampling

To determine the expected reduction in the sample variance due to using the mean of multiple samples for each

subject, assume there are k subjects, each with n_i samples, and that x_{ij} is a sample for subject i , $i = 1, 2, \dots, k$, $j = 1, 2, \dots, n_i$. Denote the within-subjects variance for subject i by σ_i^2 and the between-subject variance as σ_b^2 . The total variance, σ_T^2 , is given by

$$\sigma_T^2 = \sigma_b^2 + (1/k) \sum \sigma_i^2, \quad (4)$$

where summation is from 1 to k . Let x_i denote the mean of the x_{ij} . The variance of x_i is equal to σ_i^2/n_i . The reduced total variance, σ_{Tr}^2 , is given by

$$\sigma_{Tr}^2 = \sigma_b^2 + (1/k) \sum (\sigma_i^2/n_i). \quad (5)$$

For simplicity, assume that all subjects have the same number of samples, n , and they all have the same within-subjects variance, σ_w^2 . The difference in total variance due to using the mean of n samples, D , is given by

$$D(\sigma_T^2) = \sigma_w^2 \cdot ((n - 1)/n). \quad (6)$$

The percent change in variance due to taking multiple samples is simply

$$\%C(\sigma_T^2) = 100 \cdot D/\sigma_T^2. \quad (7)$$

The reduced SD, σ_{Tr} , is simply the square root of the reduced variance, and the percent change in SD is simply

$$\%C(\sigma_T) = 100 \cdot (\sigma_T - \sigma_{Tr})/\sigma_T, \quad (8)$$

where σ_T is the SD corresponding to the total variance.