The Seven Deadly Sins of Statistical Analysis

Kuzon, William M.- Urbanchek, Melanie G.- McCabe, Steven (1996) *Annals of Plastic Surgery*, 37:265-272.[1]

In a pedantic but playful way, we discuss some common errors in *the* use of statistical analysis that are regularly observed in our professional plastic surgical literature. The seven errors we discuss are (1) the use of parametric analysis of ordinal data; (2) the appropriate use of parametric analysis in general; (3) the failure to consider the possibility of committing type II statistical error; (4) the use of unmodified f-tests for multiple comparisons; (5) the failure to employ analysis of covariance, multivariate regression, nonlinear regression, and logistical regression when indicated; (6) the habit of reporting standard error instead of standard deviation; and (7) the underuse or overuse of statistical consultation. Confidence and common sense are advocated as a means to balance statistical significance with clinical importance.

I count religion but a childish toy,
And hold there is no sin but ignorance.
—Christopher Marlowe, *The Jew of Malta* (cl589 prologue [1, p. 183]

Plastic surgical research has become increasingly sophisticated. The use of statistical analysis for the interpretation of research data has grown from an occasionally to a commonly utilized tool. In general, however, plastic surgeons lack a strong background in statistics. Errors in both the employment and the interpretation of statistical analysis are frequent in the plastic surgical literature [2, 3]. The purpose of this paper is to address what the authors perceive as the most common errors in the use of statistical analysis made by plastic surgical researchers. We have chosen to portray these errors as "seven deadly sins" mainly for literary effect and to lighten reading about this dry topic. No tone of self-righteous indignation is intended: "He that is without sin among you, let him first cast a stone at her" (The *Holy Bible,* John 8:7). The focus of our discussion will be the principles of statistical analysis of data. The details of particular computational methods can be found in numerous standard statistical textbooks. We gratefully acknowledge the many who have emphasized proper use of data treatment principles, especially those listed in the references [4-9].

Sin 1: Using Parametric Analysis for Ordinal Data

Be sure your sin will find you out.
— The Fourth Book of Moses, Called Numbers 32:23 [1, p. 11]

The basis of both Sin 1 and Sin 2 is disregarding specific conditions about the parameters of the population being studied. Sin 1 is the use of a parametric statistical test for ordinal data analysis. Although Sin 1 is really a subset of Sin 2, in our opinion it merits discussion first and separately as the most common statistical error found in the plastic surgery literature. The two critical definitions here are *ordinal* and *parametric*.

Measurement scales can be nominal, ordinal, interval, or ratio. Nominal scales simply categorize data without assigning any hierarchical order. An example of nominal data would be compiling a list of complications from a particular surgical procedure. Although this nominal data allows one to distinguish between different complications, in the absence of additional information it does not allow the complications to be ranked in order of gravity.

[1] The authors would like to acknowledge Paul Corey, PhD, for his teaching and collaboration concerning statistical analysis over many years. Dr.Matthew Witten is gratefully acknowledged for his review of this manuscript for accuracy in its depiction of statistical principles.

Ordinal scales are used to rank data points hierarchically. A familiar ordinal scale is the ubiquitous ranking scale for outcomes of aesthetic or reconstructive surgical procedures: poor, fair, good, excellent. The order of the various levels is well defined (excellent >good > fair > poor), but the interval between each level is not certain.

An interval scale has discrete, defined levels and, in addition, the interval between each of the levels on the scale is well defined (and usually equal). The number of positive lymph nodes in a neck dissection is an example of interval data. A given patient cannot have 2.6 positive lymph nodes. However, four positive nodes are twice as many as two positive nodes, so the interval between levels of the scale is clearly defined. The level of variation being measured is usually scaled with equal units.

In a ratio scale, there is no restriction of a data point to a discrete level. Any value is permitted, including fractions. Ratio data meet all the qualifications of the previous levels of measurement scale, with the additional requirement that there must be a meaningful zero point representing complete lack of the characteristic. Scale values may be multiplied, added, and divided. Examples of ratio scales include temperature in degrees Kelvin, size in millimeters, molar concentrations, and weight in grams.

In sampling theory, a parameter is a variable that expresses some property of the entire population (from the Greek parametrein: to measure one thing by another). Population mean, variance, and standard deviation are the parameters most commonly used to describe a population. Sample mean, standard deviation, and variance are the corresponding descriptive statistics for a sample of data drawn from that population. However, parameters and statistics are calculated. That is, multiplication and division are used to compute the mean and variance. In order for these mathematical operations to be valid, the data must be expressed using an interval or a ratio scale.

Therein lies *the* sin: simply expressing ordinal data using integers does not justify the use *of* parametric statistics. Just as it is invalid to rank the results of a given surgical procedure as poor, fair, good, or excellent and state that the average result is "fair and a half," it is invalid to rate those same outcomes as 1, 2, 3, or 4 and state that the average result is 2.5. The Baker scale for capsular contracture, Clarke's levels for melanoma thickness, the Stass-Lowenstein scale for staging pressure sores, Lovet's scale for rating muscle power, and other analogous ordinal scales all fall under this restriction. *To avoid committing Sin 3, for nominal or ordinal scaled data, use nonparametric statistical analysis (see Sin 2).*

Sin 2: Inappropriate Use of Parametric Analysis

0! What authority and show of truth
Can cunning sin cover itself withal.
—William Shakespeare, Much Ado About Nothing, IV, I, 35 [1, p. 209].

The most "common" statistical tests (Student's t-test, analysis of variance, and least-squares regression analysis) are parametric tests. This means they use interval or ratio scaled data to calculate sample statistics with direct inference to population parameters. Before parametric statistical analysis is appropriate, certain sampling criteria must be met: (1) the study sample must be randomly drawn from a normally distributed population and (2) the sample size must be large enough to be "representative" of the study population. Although several common parametric tests (the t-test in particular) are "tolerant" of relaxation of these additional two criteria, in strict terms, parametric analysis should only be employed if they can be fulfilled.

The assumption of normality of the sample data can be tested with straightforward statistical methods (e.g., chi-squared goodness-of-fit test or the Kolmogorov-Smirnov test). The minimum acceptable sample size for parametric analysis is a more subjective issue. A strict attitude toward sample size would be that parametric analysis is appropriate only if N>10 (or even N>30) for each group. A more prevalent opinion is that sample sizes sufficient to achieve statistically significant differences between groups or, if there is no significant difference between groups, to achieve an acceptable statistical power (discussed later) are generally acceptable.

Regardless, unless sufficient justification for the use of parametric analysis can be provided, non-parametric analysis (often called "distribution-free" techniques) should be employed. For most of the common parametric tests, an equivalent nonparametric approach is available (Table). Nonparametric methods generally involve frequency analysis or the ranking of data and performing analysis on the ranks. Nonparametric methods can be used with ordinal data, do not require normally distributed data, and can be used with small sample sizes.

Sin 3: Failure to Consider Type II Statistical Error

There are worse things than a lie . . . I have found ,. . that it may be well to choose one sin in order that another may be shunned.
—Anthony Trollope, *Doctor Wortle's School* (1879) Ch 6 [1, p. 555]

The following discussion will consider the case of determining whether two sample (or experimental group) means are "statistically" different. Statistical analysis is actually the computation of probabilities. When using the calculation of a probability to decide whether two means are different" or "the same," a widely accepted significance level of 0.05 or 5% is used. If we compute that the likelihood of two samples being drawn from a single population is less than 5%, we conclude that the two means are "statistically significantly different." The interpretation of this conclusion is clear. Our null hypothesis was that there is no difference between two means. Rejection of this null hypothesis signifies that there is less than a 5% chance that our conclusion (that the means are different) is erroneous. We accept a 5% chance of incorrectly rejecting the null hypothesis. This wrongful rejection of the null hypothesis when it is true is referred to as a type I error. Alpha is the probability associated with committing a type I error. By preselecting α (usually 5%), rejection of the null hypothesis is associated with a known and controllable chance of error.

Examples of Nonparametric Analog of Common Parametric Statistical Methods[a,b]

Type of Problem	*Type of Data*	*Parametric Methods*	*Nonparametric Methods*
Comparison of groups	One group (compared to a reference value)	z-test, t-test	Chi-squared test, Kolmogorov-Smirnov test
	Two independent groups	t-test, z-test, analysis of variance	Wilcoxon's signed rank test, median test, chi-squared test, Kolmogorov-Smirnov test, Mann-Whitney test
	Two paired or related groups	Paired t-test, z-test	Wilcoxon rank sum test, sign test
	Three or more groups	Analysis of variance, z-test	Kruskall-Wallis test, Friedman two-way analysis of variance by ranks
Association	One sample	Least-squares correlation analysis	Spearman rank correlation coefficient, Kendall's rank correlation coefficient (tau)
	More than one sample[b]	Regression analysis or logistical regression	Chi-squared test of independence

[a] Note that for each row, all the tests listed in the nonparametric column are similar in approach to all of those in

the parametric column.

[b] Note that the chi-squared test can be applied to frequency data only. There is no direct nonparametric analog of least-squared regression analysis.

A further consideration arises when we *accept* the null hypothesis, concluding that we fail to find a real difference between the two sample means. A type II error occurs when the null hypothesis is false and *we* fail to reject it. The probability of committing a type II error is termed beta. Alpha and β are inversely related and vary inversely with sample size. To decrease the probability of committing both type I and type II errors, the sample size is increased. When comparing two sample means given an α level and sample size (N), the estimated β depends on the magnitude of the difference between the two means and on sample variance. The relationship among these statistics for comparison of two groups using a z-test is given by Equation 1:

$$Z_\beta = \sqrt{\frac{ND^2}{\sigma^2}} - Z_{\alpha/2} \tag{1}$$

where:

Z_β = standart normal value associated with β
$Z_{\alpha/2}$ = standart normal value associated with α
N = sample size
D = difference between the two sample means $\left(\mu_1 - \mu_2\right)$

and σ^2 = the variance of the differences between the sample means.

Most commonly, a is set at 0.05, so $Z_{\alpha/2} = 0.84$. Using Equation 1, it is possible to determine β and the power expected when testing the statistical significance of the difference between the two means. Beta can be used to express the power of a statistical test. Power is denned as (1-β). Typically, an acceptable β level is 0.20 ($Z_\beta > 1.64$). If no difference between the means is detected, this translates into a 20% chance of incorrectly concluding that the means are similar.

Closer examination of this equation reveals its implications. As the difference between the two means (*D*) or the sample size (*N*) increases, the calculated Z_β increases and the chance of making a type II error decreases. Conversely, as the difference between two means or the sample size decreases, Z_β decreases and the chance of making a type II error increases. As *D* approaches zero, the chance of making a type II error becomes infinite. Because of this, it is generally agreed that acceptance of the null hypothesis cannot be proven. Rather, one can only compute the β and power associated with rejection of the null hypothesis. Therefore, whenever a conclusion of "no significant difference" between groups is reached, a calculation of the associated β and resulting power using Equation 1 is needed. Note that the principles discussed here apply when there are more than two experimental groups, but the computation of β is more complex, often involving reiterative estimation

If the sin of failing to report β or the chance of making a type II error is serious, the sin of failing to compute sample sizes based on a reasonable β is fatal. When designing a study, it is possible to calculate the sample sizes needed to achieve an acceptable statistical power. When comparing two group means, Equation 1 can be rearranged as Equation 2:

$$N = \sqrt{\frac{\left(Z_\beta - Z_{\alpha/2}\right)}{D^2}} \tag{2}$$

The between-group difference (*D*) and the sample variance (α^2) can be obtained from pilot data, from similar data in the literature, or through educated guessing. By predetermining acceptable values for both α (usually 0.05) and β (usually 0.20), N can be estimated for each group from Equation 2. This should be completed before experiments are started and once or twice as data are collected. This will ensure that, if the null hypothesis is accepted once data collection is complete, the probability of a type II error will be "acceptable." Note that once a statistically significant difference between groups is achieved, type II error becomes less of a concern. Therefore, when collecting data, an interim analysis will determine either that a significant difference between groups has been obtained or that the selected β has provided the desired protection against accepting the null hypothesis when it is false. One of these criteria should be fulfilled before data collection is completed. If it is not, more data points should be added to the data set. Needless to say, grant applications that indicate an understanding of this problem by including sample size calculations as part of the experimental design are more likely to succeed than those that ignore type II error entirely.

Sin 4: Using Unmodified t-Tests for Multiple Comparisons

The sin ye do by two and two ye must pay for one by one.
—Rudyard Kipling, Ib 1.60 [1, p. 708]

It has been gratifying to observe a general decline in the frequency with which this sin is committed. Nevertheless, it is still a common error. This problem is again related to the calculation of probabilities and, specifically, to type I error. Consider the comparison of group means in an experiment with three groups: A, B, and C. If a two-group comparison test (e.g., the t-test) is employed, we accept, as discussed earlier, a 5% chance of being in error when concluding that there is a statistical difference between any two groups. To compare all three groups to each other, we must perform three pairwise comparisons: A vs. B, B vs. C, and A vs. C. Therefore, the cumulative probability of erroneously rejecting the null hypothesis is 5% (for A vs. B) + 5% (for A vs. C) + 5% (for A vs. C) = 15% overall. As more groups are compared, this cumulative chance of type I error is compounded. Therefore, multiple, unmodified pairwise comparisons are not valid.

A strategy to diminish the chance of reaching invalid conclusions when comparing multiple group means is analysis of variance (ANOVA). Although ANOVA designs can become complex, the basic principle is very simple. ANOVA asks the question: Is the variation within the data set due to differences between groups greater than the variation due to differences within groups? This determination is made by computing an F ratio, which is an expression of between-group variation divided by within-group variation. The probability associated with the F ratio can then be determined from standard F distributions. If this probability is greater than 0.05, we conclude that the variation in our data set is due to random sampling and not to any effect of our group assignments. Group assignments are also called "main effects." In the case of a non significant overall F ratio for an ANOVA, it is not valid to perform pairwise comparisons of individual group means. If the F ratio is associated with a probability less than 0.05, we conclude that differences between groups explain a significant portion of the variation within the data set (i.e., that there is a significant influence of our main effects on group means). In this latter situation, pairwise comparisons of multiple individual group means are permissible, but two criteria must always he fulfilled in order for these comparisons to be valid.

The first of these we have already stated: The overall ANOVA F ratio must be significant. For this reason, the comparison of multiple group means after an ANOVA has been found to be significant is referred to as post hoc testing. For maximum validity, all multiple comparisons of group means must be post hoc. The second criterion relates to the cumulative type I error associated with the multiple pairwise comparisons discussed earlier.

Because the cumulative error associated with multiple pairwise testing using an a of 0.05 for each individual comparison will result in an increased probability of detecting a significant difference between groups, post hoc pairwise comparisons must be modified in some way to ensure an overall α of 0.05 once all the comparisons are completed.This is referred to as the Bonferroni principle or the Bonferroni correction. In the case of our three-group example, we would modify the significance level for each individual comparison, setting α= 0.05/3. Then our three comparisons would result in a cumulative α error of 0.05/3 (A vs. B) + 0.05/3 (B vs. C) + 0.05/3 (A vs. C) = 0.05 overall. There are numerous post hoc tests (Duncan's, Tukey's, and others), each with relative advantages and disadvantages in particular circumstances. However, they all employ this same principle: modification of the probability distribution or the significance level for each individual comparison to ensure an overall α level of 0.05.

One other critical feature of ANOVA should be emphasized. "Standard" ANOVA calculations rely on equal numbers in all cells of the design (i.e., in all groups). When there are unequal numbers of data points in each group, the ANOVA is said to be "unbalanced." There are several ways to deal with unequal sample sizes in an ANOVA design. The important point is that many statistical software packages do not ideally provide for the analysis of unbalanced designs. In particular, the common strategy of simply filling in missing data points in a given group with the arithmetic mean for that group is not recommended. Most quality statistical software packages will indicate that analysis of an unbalanced data set is either not possible or that some compensatory strategy has been used. However, several packages that we have seen will simply ignore the unbalanced nature of the data and will compute *F* ratios without correction. This can give erroneous results. When evaluating statistical software, the ability to deal accurately with unbalanced ANOVA designs is an important feature for which to look.

Sin 5: Underutilization of Analysis of Covariance (ANCOVA), Multivariate Regression, Nonlinear Regression, and Logistic Regression

The worst sin towards our fellow creatures is not to hate them, but to be indifferent to them: that is the essence of inhumanity.
—George Bernard Shaw, The Devil's Disciple, 1901, act II [1, p.680]

While most plastic surgery research is conducted using relatively straightforward experimental designs that are adequately handled with pairwise comparisons, ANOVA, or standard least-squares regression analysis, there are numerous circumstances in which more sophisticated statistical methods should be considered. Several hypothetical examples can illustrate such circumstances. Consider an experiment in which nerve regeneration after peripheral nerve repair is studied in young, middle-aged, and old rats. Routine variables to document axonal regeneration (axon histomorphometry, electron microscopy (EM], electrophysiological evaluation] are recorded for animals in each group and the group values are compared using ANOVA with appropriate post hoc pairwise testing. Significant differences are noted for young vs. old animals, therefore the investigator concludes that axonal regeneration differs between young and old animals. However, young animals are both lighter and smaller than old adults. Consequently, body weight or, more likely, body size (i.e., distance needed to achieve functional axonal regeneration) may differ substantially between groups. If this type of confounding influence could affect conclusions, ANCOVA is a useful technique.

ANCOVA asks the question: For our target (dependent] variables, is there a difference between groups if we adjust our data, taking into consideration differences between groups with regard to possible confounding variables (covarivates]? In our example, body size is a potential confounding variable. ANCOVA would examine differences between groups after accounting for the effect of the covariate (body size) and would determine whether or not body size had a significant relationship to the outcome variables. Because experimental variables are often interrelated, ANCOVA is frequently indicated, but rarely utilized by statistical sinners.

When there is more than one important covariate that could affect a particular outcome, the use of more complex regression analysis should be considered. In a multivariate regression, a least-squares computational method is employed for any number of variables in an attempt to account for the variation observed in the dependent variable. The variance due to an individual, independent variable is compared to the total variation in the data set and an F ratio is computed. If the probability of a larger F ratio is less than 5%, that variable is considered to be "significant" in explaining the variation in the outcome measure. For example, consider an experiment to determine the factors that determine the breaking strength of a wound after repair. Our outcome or dependent variable is breaking strength. We also record the age of the animal, the time since repair, the collagen content of the wound, and the time of day (he wound was created and repaired. Using multivariate regression, the "significance" of each of these independent variables in accounting for the variation in breaking strength could be tested. A predictive equation could be derived. Most likely, in this case, we would discover that the time since repair and the collagen content of the wounds were highly significant in explaining the breaking strength, the age of the animal less significant and the time of day of no significance at all.

A limitation of multivariate regression is that the variables must be continuous. In plastic surgery, we frequently consider categorical or non-continuous variables to be of great significance. Such categorical variables could be sex (male vs. female), diabetes (yes vs. no), craniosynostosis (syndromic vs. nonsyndromic), and so on. In order to consider the effect of independent categorical variables such as these on a given dependent variable, logistical regression should be employed. The concept is identical to multivariate regression. Both continuous and categorical variables can be included as independent variables in the same analysis. In clinical outcomes research, logistical regression should be considered a pre eminent tool for determining the importance of the various factors that could affect a clinical outcome.

Although many experimental situations in plastic surgery could benefit from ANCOVA, multivariate and/or logistical regression analysis, o nonlinear regression, our observation is that they are seldom employed. As laboratory and clinical research in plastic surgery grows in sophistication, this sin of omission will require rectification.

Sin 6: Reporting Standard Error Instead of Standard Deviation

Sin is whatever obscures the soul.
—Andre Gide, La Symphonie Pastorale(1919)[1, p. 727]

Reporting standard error of the mean is perhaps sin at all; but, reporting standard error understanding- its meaning is a serious transgression. We all know that standard error is as the standard deviation divided by square root of N, but this equation does not the meaning of the standard error of the mean. Challenge yourself now; give a definition of the standard error of the mean without reading any further. You should have immediately stated that the standard error of the mean is the square root of the variance (i.e., the standard deviation) associated with the distribution of sample means that would be derived by repeatedly sampling n data elements from the study population. Now give a definition of standard deviation. Standard deviation is the square root of the sample variance and is, therefore, a direct measure of the spread of data in a sample. It is well known that two-thirds of the sample data points fall within one standard deviation of the sample mean and that 94% of data points fall within two standard deviations of the mean. This direct, easily conceptualized meaning of standard deviation makes it preferable when reporting descriptive statistics. The meaning of the standard error of the mean is far more difficult to conceptualize and therefore more difficult to interpret directly. *The practice of reporting standard error because it "looks better" is a statistical sin.*

Another argument sometimes advanced for reporting standard error is that one can easily determine, by looking at the overlap of standard error bars on a graph, whether or not two means are significantly different. *This belief is incorrect*. It is easy to construct scenarios wherein two means will have values within one standard error of each other, yet they are significantly different statistically. It is also easy to construct the alternative scenario (bars don't overlap, means are not significantly different. It is not possible to determine whether two means are significantly statistically different simply by looking at either standard deviation or standard error bars on a graph, therefore, because of its direct and easily understood meaning, we advocate the reporting of standard deviation as the parameter indicating the spread of the sample data.

Sin 7: Failure to Rely on a Statistician or Relying Too Much on a Statistician

'What is the Unpardonable Sin?' asked the lime-burner ... 'It is a sin that grew within my own breast.' replied Ethan Brand .... The sin of an intellect that triumphed over the sense of brotherhood with man and reverence for God.
—Nathaniel Hawthorne, Ethan Brand (1850) [1, p, 503]

"My statistician says ..." This statement is a double-edged sword. In its positive connotation, it indicates that the researcher has sought the expertise of a statistician to assist with the interpretation of data, an obviously desirable maneuver. It may, however, indicate that the researcher has little or no concept of the statistical methods being employed for the analysis of the data, preferring to abdicate all responsibility to a third party. While there are times when statistical analysis may become extraordinarily complex, it is our opinion that it is the responsibility of the primary author or investigator to understand and to agree with the statistical analysis utilized. This may seem unfair, since one cannot become an expert on everything. Nevertheless, if statistical analysis is to be used as our means of evaluating our research results and thereby used to validate important decisions regarding patient management, we submit that it is a sin simply to "give data to the statistician" and then to get back "results."

An appropriate analogy may be a prospective cosmetic surgery patient. It is certainly not necessary for the patient to understand the multitude of details involved in the execution of their surgical procedure. On the other hand, most surgeons would likely agree that a patient who wanted no knowledge of the nature of the procedure at all [e.g., incision placement, amount of time off work, likely benefits, and so on] would make a very poor surgical candidate for many reasons, not the least of which is that they would likely be unrealistic about what the procedure could accomplish. In this same way, a researcher with data requiring interpretation should not entrust everything to a statistician, but rather should become informed to the extent that he/ she can actively participate in the interpretation of the data in a meaningful way. Very few of the common errors in the use of statistical methods that can be found monthly in all of our plastic surgery journals would occur if all researchers accepted this responsibility.

The question then becomes: How can an average plastic surgery researcher get to know something about statistics without becoming a statistician? An effective strategy to accomplish this goal is to develop a long-term relationship with a statistician. Choosing the right statistician is analogous to choosing a lawyer, doctor, or hair stylist. It will take some trial and error, and require some investment of time. After identifying someone who is a potential collaborator, it is absolutely necessary to consult with them prior to commencing experiments. It is critical that they achieve an understanding of your experimental goals and of the technical methods employed. Just as we have stated that the investigator must acquire statistical sophistication, it is essential that the statistician acquire a working knowledge of the field of research the data addresses. Data cannot be interpreted in a vacuum. The "lies, damn lies, and statistics" attitude springs largely from statistical interpretation of data out of the context of its experimental milieu. If statistical analysis is performed without an understanding of the underlying biology, whatever "my statistician says" becomes completely irrelevant.

Statistical Significance vs. Clinical Importance

In medicine, sins of commission are mortaJ, sins of omission venial.
—Theodore Tronchin, quoted in *Bulletin of New York Academy of Medicine,* V (1929), 151 [1, p. 357]

Like any technique for the reduction of data, statistical analysis can be used to distort the truth. Even if statistical methods are employed and interpreted correctly, statistical analysis is still merely the computation of probabilities that will not overcome problems in methodology and, 1996 in fact, may give a false sense of security. The statistical probability that two means are similar or different, that variables are interrelated, and so on cannot be used as "proof" of a biological or surgical hypothesis. Informed reasoning when structuring hypotheses and designing experiments, and the thoughtful interpretation of the data and its statistical analysis, are required to decide what is "right." In this endeavour, it is sometimes appropriate to decide that statistical analysis is not needed or that statistical findings should be ignored. Although they may not be common, we would submit that these situations are at the core of the pursuit of science where startling or unexpected advances occur. Rather than interpreting statistical analysis as a "final answer," we should think of the result of statistical analysis as another piece of data that helps us decide whether our conceptualization of biological mechanisms is correct or incorrect. In order to achieve that level of scientific sophistication, we all need to confess and be absolved of \'m our statistical sins.

References

1 Bartlett, J. (1980) *Familiar quotations*. Beck E.M., et.al., eds., Boston: Little, Brown.

2 Altman, D. (1994) The scandal of poor medical research, *British Medical Journal,* 308:283-284.

3 Velanovich, V.- Robson, M.C.- Heggers, J.P. et al. (1987) Statistical analysis and study design in plastic and reconstructive surgical research. *Plast. Reconstr. Surg.*,80:308-313.

4 Cohen, J. (1977) *Statistical power analysis for the social science*, New York: Academic Press: 462

5 Feinstein, A. (1977) *Clinical Statistics*, St. Louis: CV Mosby.

6 Freeman, L. (1965) *Elementary Applied Statistics*, New York: Academic Press,

7 Guilford, J.- Fruchter, B. (1988) *Fundamental Statistics in Psychology and Education*, 6th ed. New York: McGraw-Hill.

8 Siegel, S. (1988) *Nonparametric Statistics for the Behavioural Sciences*, 2nd ed. New York: McGraw-Hill.

9 Winer, B. (1991) Statistical principles in experimental design, 3rd ed. New York: McGraw-Hill.