USE AND MISUSE OF STATISTICS IN ODONTOLOGY

M. D., Odont. Dr. Professor Emeritus
University of Lund, Sweden

Hilding Björn

Odontological research is essentially concerned with biological questions. Statistics deals with numbers which are nothing more than limited and often oversimplified abstractions of complicated biological phenomena and processes. This calls for a cautious attitude towards statistics.

A computer with a good program can produce statistics, distributions, tables, statistical tests etc. No computer can make decisions or draw conclusions, in other words transform numbers to knowledge. Like a computer statistics is a useful tool—but just a tool. To make statistics meaningful some other ingredients are needed: experience, judgment, common sense.

Naturally, not every dentist can become a statistician. But most dentists read odontological reports, which these days contain many conclusions based on statistical analysis. As a consumer of such literature every dentist should be able to judge if the statistics, which he is reading, is reliable. He should not believe that what he reads is the final truth. Instead he should try to evaluate the statistics presented in reports!

For this lecture I need examples from odontological journals. The reports I am going to cite may contain valuable information in spite of fallacies on one point or the other. What I criticize is not articles or their authors but mistakes and errors that may occur to anyone of us.

As summary of this introduction I will cite the following advice given in 1907 by Edward H. Angle:

"Read it at least three times; first, to see what it is all about; secondly, to see what it says; and thirdly, in an attitude of friendly hostility."

POPULATIONS and SAMPLES

A study of a large population would be expensive in time and money. Therefore, an investigation is usually made on data collected from a miniature representation of the population, a sample.

To be useful for drawing of true inferences about the population, the sample must be obtained in such a way that:
1. every single individual of the population of persons, animals, suture needles, cylinders and ampoules with local anaesthetics or whatever variable the investigation is about, has the same chance of being drawn and
2. that the individuals are drawn independently of each others.

This results in an independent random sample. A sample of this kind allots to each individual unit of the population an equal probability of being included in the sample. Thereby bias (= partiality, favoritism, distortion) is avoided, that else could lead to false conclusions. Most statistical operations presuppose this type of sampling.

Suppose that some samples of individuals are drawn from a population in order to find out the average number of teeth per person. We will find that the means will not be the same for the different samples and not the same as for the total population. The difference between a sample mean and the corresponding population mean is called the sampling error. One of the aims of statistical analysis is to calculate this error. Random independent sampling makes this possible and makes it possible to draw conclusions about the population on the basis of data collected from the sample.

"Eight adult patients were chosen at random. They exhibited advanced destructive periodontitis...". (Tanner et al. J. of Clinical Periodontology, 1979 : 6 : 278)

The term random sampling is often mistaken for haphazard, accidental, casual selection. In statistics it means the contrary. It means selection by use of a random number table or a random number program in a computer or by some other method that excludes bias.

The first thing to do, when authors state that they used a random selection or sample, is to look for the population from which the selection was made. In this report no population at all is mentioned and for good reasons. A random sample of persons with advanced periodontitis must come from a population of individuals with advanced periodontitis. However, populations are composed of persons without periodontitis and those with slight, moderate or severe periodontal disease. It is of course possible to obtain a subsample of individuals with advanced periodontitis, but then the original population and the criteria for selecting the subsample should be stated. We must conclude that this sample was not a random but a haphazard one.

INDEPENDENCE

Violation of the rule of independence is rather common within odontology, specially within cariology and periodontology as well as in research involving experimental animals. Experimental animals such as dogs and monkeys, are expensive. Therefore, the scientist tries to keep the number of animals low. The temptation lies near at hand to compensate this by an inflation of the number of units, N, on which statistical calculations are based, by using teeth, which are many, or tooth surfaces, which are still more numerous, or-still worse-so-called “sites”, that is any number of parts of the structures
around the circumference of the teeth. In this way the N may be raised to entirely un-
realistic levels, which is a sabotage to the reliability of the statistics.

What is wrong with this method is, that individuals and not teeth or tooth surfa-
ces or “sites” are the actual sampling units. Individuals react in an individual way, which
means that teeth and periodontal tissues in one and the same individual are related and
tend to react in the same way. They are not independent of each others.

The Nov. 1982 issue of the Journal of Clinical Periodontology contains eight origi-
ginal articles. No less than five of these are ruined by one and the same statistical error-
violation of the rule of independence.

In the first report three dogs constitute the experimental sample. The total number
of items in the calculations is, however, no less than 58!

In another article, the regeneration of the interdental tissues following surgical de-
nudation of the root surfaces was studied in seven patients. So the sampling unit was
patients. However, the statistics was performed on 33 teeth! The prize is, however, won
by an investigation on “Patterns of progression and regression of......periodontal disease
in 22 patients.” Here the statistics is made on in all 1155 “sites”!

DESCRIPTIVE STATISTICS

The task of so-called descriptive or describing statistics is to present data in a con-
densed form which is at the same time easy to understand, e.g. by use of tables or dia-
grams. The present fashion among editors of scientific journals seems to favour diagrams
at the cost of tables. There is a wide-spread belief that diagrams are more easy to read
and more informative.

There are reasons to question this belief. Some diagrams may make things clear
at a glance—but certainly not all!

The diagram (Fig. 1) illustrates changes with time of the maximum mouth open-
ing which could be attained in cases of osteoarthrosis of the temporomandibular joint.
It is not possible to get much information from this diagram. How much more accurate
and informative would not a table containing the actually observed values have been.

However, also the informative value of tables may be sabotaged: The following two
tables (Table 1, Table 2) are from an article on the effect of grafts on the contraction
143)

(Table 1) Percentages without statement of N are meaningless. The reader has to
search the text in the hope of finding the missing figures. In this case, it is not possible to
find any definite values of N. It is a pity not to be able to get reliable information from
a report, which obviously has cost much time and money.

(Table 2) The only information given is, that the measurements are in mm. No state-
The first number must be an average but the second, is it standard deviation or standard error?

It is necessary to tell what the values in a table mean! A table should contain all information needed to decipher it!

FREQUENCY DISTRIBUTIONS

Frequency distribution tables and diagrams are convenient means of summarizing data to get a survey of populations and samples.

The diagram (Fig. 2) (a so-called histogram) demonstrates the percentage distribution of 797 twenty year old recruits who have 0 to 28 intact teeth. Such a diagram displays some interesting features: (Björn, A. -L. & Halling, Arne, Swed Dent J 1983: 7: 129)

1. Form

A distribution usually has two tails, one to the left and one to the right. Sometimes there is just one tail.

A frequency distribution may be symmetrical as is case with the normal distribution.
The distributions encountered in odontological research are very seldom normal. One exception is shown in this diagram—a practically normal one. However, the distributions we usually see, are more or less skewed. Most often the right tail is longer than the left one. Then we say that the distribution is skewed to the right.

2. Central tendency

Variable values have a tendency to collect towards the center of the distribution—the so-called central tendency. Means of different kind are used to express this tendency. Commonly used are the arithmetic mean—in every day talk just the mean—and the median, i.e. the middle variable value (or the mean of the two middle values if N is an even number). For symmetrical distributions mean and median coincide—as in the above diagram. It is easy to see, that mean as well as median of this distribution lies around 13 intact teeth.

Why use the median at all? The mean has certain limitations. Among other things it is very sensitive to extreme values. The median is a better measure of the central tendency when a distribution is strongly skewed. As an example (simulated) may be used the income of 9 persons in 1000 $: 7.2 9.0 10.3 12.2 15.1 17.0 18.6 22.4 245.0 Mean = 39.6 Median = 15.1

The diagram (Fig 3) demonstrates a one-tailed distribution skewed to the left, while the diagram (Fig 4) shows a two-tailed distribution skewed to the right. These diagrams also
transfer an idea of the degree of peakedness of the distributions. The left-hand distribution is extremely peaked (leptokurtic). In other cases distributions may be flat (platykurtic). Especially skewness of higher degree may affect the outcome of parametric statistical tests. (Björn & Halling Swed Dent J. 1983: 7: 129)

3. Deviations from the mean

The normal distribution is completely defined by two parameters, the arithmetic mean and the Standard Deviation. The S. D. is a measure of the deviations from the mean. We can often get a pretty good impression also of this parameter by studying frequency distribution diagrams.

Many distributions—such as those of number of intact and DMF-teeth, of microbial counts, saliva secretion rate etc., etc.—tend to be extremely skewed. In such cases the standard deviation is not a satisfactory measure of the dispersion. Here the use of percentiles is to be preferred. The meaning of this concept is easily understood by way of an example:

<table>
<thead>
<tr>
<th>Table 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ml/min</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>0.50-0.99</td>
</tr>
<tr>
<td>1.00-1.49</td>
</tr>
<tr>
<td>1.50-1.99</td>
</tr>
<tr>
<td>etc</td>
</tr>
</tbody>
</table>

Table 3 shows the first part of the cumulative frequency distribution of test persons on secretion rate (ml/min) of stimulated saliva.

The diagram (Fig 5) shows a percentage frequency distribution diagram. Each percentage corresponds to 1% of the number of test persons and each corresponding secretion rate is a percentile. The illustration demonstrates the 80th percentile, showing that 80% of the test persons had a secretion rate of about 2.2 ml/min or less. The remaining 20% had secretion rates between 2.2 and 5.5 ml/min.

The diagram (Fig 6) shows the 25th, 50th and 75th percentiles, also called the first, second and third quartiles. The interquartile distance is a convenient measure of the dispersion of values around the median. The second quartile=the 50th percentile=the median. An alternative way of describing a skew distribution is by way of giving the values at a number of percentiles, e.g. the 10th, 20th, 30th...etc. percentile.

SCALES

Little is known about the effect on speech of excising segments of the tongue.
Twenty patients—who had been subjected to glossectomy and, thus, had had more or less of their tongue excised—participated in a study of this problem.

One of the methods used was to ask these patients to read aloud twenty simple everyday words. The results were recorded on tape. The tapes were then played back to a panel of five listeners, who were asked to write down each word they thought that the test subjects had pronounced. If the word was correctly recognized a score of 1 was given, if not a score of 0.

Further, the patients were ranked according to the amount of tongue tissue removed, from the least to the most extensive damage.

Finally, the rate of speech was assessed by counting the words spoken per minute by each patient when reading a simple standard text.

This report offers examples of three of the most common measuring scales. First the zero-one scale used for the word tests. The scores could as well have been “yes” for “I do understand” and “no” for “I do not understand”. This is an example of the simplest scale of all, a nominal scale (from latin: nomen = name). Further examples of this type of scale are: male-female, cold-warm, light-dark, plus-minus etc. It is no real scale, more a division into categories.

Next the scale used for estimating the removed amount of tissue. This is called an ordinal or rank order scale, because the values can be ranked in order from the smallest to the largest. There are no fixed intervals between the scale divisions. It is a rubber band scale. Consequently, arithmetic means have no meaning. Another mean must be used—the median. Examples are the scales used in periodontal indices such as Russell’s periodontal index, Löe & Silness’s gingival index etc. Last but not least the interval scales, which are characterized by equal intervals between the scale divisions. Here belong the scales used for distances, weights, speeds, temperatures, areas, age etc. In the present report this type of scale was used for assessing the rate of speech.

**PARAMETRIC and NON-PARAMETRIC STATISTICS**

Parametric statistics is based on the parameters of some frequency distribution—in most cases the normal distribution. Parametric tests are, among others, Student’s t-test; analysis of correlation and regression (including multiple regression); analysis of variance etc. They require, among other things, that an interval measuring scale has been used.
Non-parametric or distribution-free statistics does not rely upon any special distribution and does not presuppose an interval measuring scale. Examples of nonparametric statistical tests are the Sign test, the Wilcoxon-Mann-Whitney tests, the Spearman rank correlation, the Friedemann analysis of variance, the Chi-square test and many others.

**SIGN TEST**

The main basis for statistical analysis is the knowledge of the everywhere present random variation of events. On the other hand the main task of statistical analysis is to find an answer to the question: is what I have found due to this random variation? Or can it with some degree of confidence be ascribed to the factor tested, e.g. the outcome of operations; or to the difference in effect of two drugs; etc, etc. An example of the simplest sort may serve to elucidate this point.

Suppose that we are walking down an avenue with pairs of trees growing on the right and left sides. Further imagine that we get the impression that the right hand trees are on an average somewhat higher than the left hand ones. We cannot measure the heights but we can compare the heights of the trees in each pair and judge whether the right-hand or left-hand tree is the higher one.

Assume that there were in all 58 pairs of trees. We judged that in 33 pairs the right-hand tree was the higher one. These pairs we gave a plus sign. There were 12 pairs where we judged the two trees to be of the same height. These were marked with a zero. Finally, we found the left-hand tree to be higher in 13 cases. These we gave a minus sign: (Table 4)

<table>
<thead>
<tr>
<th>+</th>
<th>0</th>
<th>−</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>12</td>
<td>13</td>
</tr>
</tbody>
</table>

Assume that we name the frequency of plus signs A and that of minus signs B. Thus, A=33 and B=13. Now the question is: Can the observed difference A-B simply be ascribed to the random variation? Or can we with some degree of confidence say that it signifies a real difference?

**NULL HYPOTHESIS**

To solve this problem we put up the hypothesis that there is in fact no difference, i.e. A-B=0 or null. This is a so-called null-hypothesis, H₀. An equivalent hypothesis is of course A=B.

The zeros, which cannot contribute to the solution of our problem, are ignored. There remain 33+13=46 differences with a sign. If right and left hand trees on an average had been of the same height, we would have expected an equal distribution of plus and minus signs, in other words, that the expected frequencies of each sign had been 46/2 = 23.

The probability that the observed difference was not merely due to random variation
can be looked up in sign test tables found in most elementary statistical text-books. In this case we find \( p \) (for probability) < 0.01, i.e. less than 1%.

**ALTERNATIVE HYPOTHESIS**

This probability we find so small that we prefer to reject our null-hypothesis and rather accept the alternative hypothesis, \( H_1 \), \( A<>B \) and conclude that the probability speaks for there really being a difference in height between the right- and left hand trees. (For samples exceeding the highest \( N \) of these tables, one can calculate Chi-square and use a Chi-square table).

The sign test is simple, easy to understand and to apply to a variety of problems but not as sensitive as some other tests.

**IMPORTANT NOTE**

It is essential to understand that the results of so-called significance tests do not tell the eternal truth. A statistical test cannot prove any hypothesis to be 100% true or 100% false. Statistical tests can only show degrees of probability. In the present example a 95% probability was found that there is a difference. At the same time a 5% risk of making a false conclusion was accepted.

**A HOLY COW**

There is a superstitious belief that \( p < 0.05 \) solves all problems. This is not so. You should never believe one single investigation to be conclusive. There is need of several control investigations before you accept anything as sure. Odontologists tend to be somewhat easygoing with respect to this.

It is also essential to understand, that a difference, even if small, will become significant if the number of cases is great enough. However, a small difference—even if statistically significant—may have no practical significance at all. The clinical significance cannot be assessed by significance tests alone. All results of statistical analysis must be evaluated against the background of the collected experience of the results obtained by the researcher and those arrived at by other scientists. Finally, clinical experience, and, last but not least, common sense must have their say.

**OTHER APPLICATIONS OF THE SIGN TEST**

The sign test can be applied not only to variables assessed by nominal scales. It can also be used in connection with ordinal as well as interval scales. Here an example with interval scale:

Suppose that we wish to compare the systolic blood pressure in 12 test persons before and after injection of an adrenergic drug. (Table 5)

Is there a significant difference in median blood pressure before and after the injec-
To calculate the medians we rank the variable values in order from the lowest to the highest: (Table 6)

<table>
<thead>
<tr>
<th>Before</th>
<th>147</th>
<th>125</th>
<th>132</th>
<th>113</th>
<th>132</th>
<th>136</th>
<th>110</th>
<th>124</th>
<th>113</th>
<th>119</th>
<th>125</th>
</tr>
</thead>
<tbody>
<tr>
<td>After</td>
<td>145</td>
<td>134</td>
<td>148</td>
<td>135</td>
<td>129</td>
<td>136</td>
<td>155</td>
<td>136</td>
<td>186</td>
<td>145</td>
<td>153</td>
</tr>
</tbody>
</table>

The median is the middle value of a distribution if the number of values is uneven. If the sample has an even number of values, the median is the mean of the two middle values. In this case the medians are 125 and 145 respectively.

So there seems to be— as we have all reason to believe—a rise in blood pressure. The question is now, if this difference may be due to random variation alone or if it signifies a difference in median blood pressure caused by the drug.

To answer this question we can use the sign test. We simply walk down the avenue of blood pressure pairs and note which one of the values in each pair is the highest. If the pressure before is the higher one, we give a minus sign, if that after is higher we give a plus sign. (Table 7)

<table>
<thead>
<tr>
<th>Before</th>
<th>147</th>
<th>125</th>
<th>132</th>
<th>113</th>
<th>132</th>
<th>136</th>
<th>110</th>
<th>124</th>
<th>113</th>
<th>119</th>
<th>125</th>
</tr>
</thead>
<tbody>
<tr>
<td>After</td>
<td>145</td>
<td>134</td>
<td>148</td>
<td>135</td>
<td>129</td>
<td>136</td>
<td>155</td>
<td>136</td>
<td>186</td>
<td>145</td>
<td>153</td>
</tr>
<tr>
<td>Sign</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Thus 2 minus and 9 plus signs. The sign test table gives the two-tailed significance level, i.e. the probability of our null hypothesis being true, i.e. the probability that there is no difference in blood pressure before and after the injection, to 10% or less but >5%.

**ONE- and TWO-TAILED TESTS**

Testing the significance of the difference in means ($\bar{X}_A$ and $\bar{X}_B$) of two distributions (A and B). Null-hypothesis: $A < B$. The one-tailed test so to say looks at one tail of each distribution. In this example the probability that $H_o$ is true is 2.5%.

Null-hypothesis: $A = B$. The two-tailed test compares both tails. The probability that $H_o$ is true is $2.5 + 2.5 = 5\%$. (Fig. 7)
The choice of one- or two-tailed test should be made before any statistical analysis is performed and should be based on previous knowledge and/or rational reasoning. If there is doubt, a two-tailed test is the safer choice.

In a scientific report always state if a one- or two-tailed test has been done as well as the reasons for the choice. Then a reader may judge for himself and have the opportunity to correct the conclusions accordingly. Therefore, the most important is perhaps not the choice but to make clear which type of test has been used.

Now back to the blood-pressure experiment! Before the analysis we selected a suitable significance level, say the 5% level. We knew already before our experiment that adrenergic drugs have the property to rise the blood pressure. This beforehand knowledge makes us anticipate a rise even in this case. The question was if the specific dosage we used was enough to cause a rise in this trial also. So we need to compare only one tail of each distribution. Therefore we use a one-tailed test, i.e. we divide the level found by two.

The probability that there was no difference is \( <0.05 \) or \( <5\% \). This probability we consider as so low, that we reject our null hypothesis that there in reality was no difference and accept the alternative hypothesis that there in fact was on an average a higher blood pressure after the injection. It is important to understand that at the same time a risk of 5% of being wrong is accepted.

The rule for selection of one-or two-tailed tests can be formulated like this: If you are sure—by previous knowledge—that a difference between means or medians can go only in one direction, then you use a one-tailed test, else select a two-tailed test. If you are in doubt it is safer to use the more conservative two-tailed test.

In our first example, that with the trees in an avenue, we might have discussed the influence of such factors as a different quality of the soil on right and left sides, different incidence of the daylight, different water supply and such factors—but, since we have no knowledge of such things, we are forced to use a two-tailed test. In the second example we could predict that under the conditions of the test procedure the effect of the adrenergic drug—if any—could only be a rise in blood pressure. Hence a one-tailed test was appropriate. Hypothesis: \( A < B \).

Why is this so important? Naturally, because the one-tailed test is more sensitive. If you do not use it in appropriate cases you may miss a really existing difference. If you use it when it should not be applied, you may accept a difference as statistically significant, when in fact it is not. Therefore, it is important to consider this point and also to state in your report that you have selected the one or the other—and why!

The sign test is a very convenient test for paired comparisons, easy to apply and fast. But it is not very sensitive. To be able to detect minor differences you may use another non-parametric test for paired comparisons, viz. the Wilcoxon signed-ranks test. If we find the conditions for a parametric test at least approximately fulfilled we may also —(663)—
WILCOXON MATCHED-PAIRS SIGNED-RANKS TEST

Like the sign test this is a method to test if there is a significant difference between the medians of paired samples.

The sign test utilizes information only concerning the direction, but not about the size of the differences between the two members of a pair of variables. The Wilcoxon test takes also the size into account, which makes it a more sensitive test.

It will now be applied to the blood pressure example. One begins by calculating the differences between the values for each person before and after the injection: (Table 8)

<table>
<thead>
<tr>
<th>Before</th>
<th>147</th>
<th>125</th>
<th>132</th>
<th>113</th>
<th>132</th>
<th>136</th>
<th>110</th>
<th>124</th>
<th>113</th>
<th>119</th>
<th>125</th>
</tr>
</thead>
<tbody>
<tr>
<td>After</td>
<td>145</td>
<td>134</td>
<td>148</td>
<td>135</td>
<td>129</td>
<td>155</td>
<td>136</td>
<td>156</td>
<td>145</td>
<td>153</td>
<td>150</td>
</tr>
<tr>
<td>Differ</td>
<td>+2</td>
<td>-9</td>
<td>-16</td>
<td>-22</td>
<td>+3</td>
<td>-19</td>
<td>-26</td>
<td>-32</td>
<td>-32</td>
<td>-34</td>
<td>-25</td>
</tr>
</tbody>
</table>

Now the differences are ranked without regard to sign and then the signs given back to them: (Table 9)

<table>
<thead>
<tr>
<th>Differ</th>
<th>2</th>
<th>3</th>
<th>9</th>
<th>16</th>
<th>19</th>
<th>22</th>
<th>25</th>
<th>26</th>
<th>32</th>
<th>32</th>
<th>34</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Sign</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Finally, the sums of the ranks with + and those with − signs are calculated: 3 and 63 respectively.

If the medians of samples A and B did not differ, one would expect, that the rank sums for + and − differences would be the same. If, instead, the rank sums differ significantly, this means that the two medians differ significantly.

To find the significance the lower rank sums is compared with a table, where the critical values for samples of different size and different significance levels are tabulated.

For this example we read the critical value for a sample size of 11 and one-tailed test to be 5 at a significance level of 0.005. Since our test result, 3, is lower, the conclusion can be drawn, that there is a difference in blood pressure significant at the 0.5 % level. (As already explained a one-tailed test is appropriate for this blood pressure experiment).

This result can be compared with the one-tailed significance obtained by other tests for paired variables. The sign test gave p<0.05, the rank sum test p<0.005, and a t-test would give p<0.0005.

In case the numbers of signed differences exceeds the tabulated sample sizes, there is a formula for an approximation to the normal distribution, which is tabulated and to find in —(664)—
any textbook on elementary statistics.

Since blood pressure is measured in an interval scale, and since there is no significant difference in variances one would, of course, prefer the t-test in this case. Suppose that instead the extent of loss of tongue tissue in glossectomized patients had been judged twice. Then, because of the ordinal scale, the nonparametric Wilcoxon test had been the method of choice.

**CORRELATION and REGRESSION**

The meaning of these terms is not always quite clear to odontologists. Both describe the mutual relationship between two variables. The meaning of these concepts may be elucidated by simple examples.

Suppose that we wish to examine the dependence of loss of weight of six test pieces of alginate impression material upon the relative humidity of the air in which the test pieces are kept.

We have two variables: Humidity, X, and loss of weight, Y. (table 10)

<table>
<thead>
<tr>
<th>Humidity (%)</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>70</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss of weight (mg)</td>
<td>250</td>
<td>200</td>
<td>110</td>
<td>100</td>
<td>60</td>
<td>20</td>
</tr>
</tbody>
</table>

We can illustrate these data by plotting a scattergram (Fig 8).

Already at first sight this scattergram indicates a relationship—a low value of X is on the whole accompanied by a high value of Y even if there is a certain spread. The relation may be illustrated by the line that best fits the points of the scattergram, the regression line.

It is obvious that the humidity of the air cannot depend on the loss of weight of the test specimens, but that instead the weight loss depends on the humidity of the air. Therefore, we may call the X-variable the independent variable and the Y-variable the dependent one. Regression is, thus, characterized by one independent and one dependent variable.

The regression line can be calculated by means of the method of least squares, which is described in any elementary textbook on statistics. The simplest regression line is the straight one according to the well-known formula \( Y = a + b \cdot X \), where \( b \) is the regression coefficient, an expression for the slope of the regression, and \( a \) is the so-called in-
tercept value. Both a and b can take on any value. Many modern pocket calculators have a built in program for calculating regression according to this formula.

Regression is not always rectilinear. The line that best fits the points of the scattergram may well be curvilinear. With a minicomputer it is easy to try different formulae to find a line with good fit. In the present case logarithmic transformation of the X-variable renders a good fit (Fig. 9).

A logarithmic transformation can also easily be made with aid of a pocket calculator with built in regression program. The degree of fit can be indirectly assessed by correlation analysis.

Extrapolations are common in commercial life in order to predict the development in the future. A scientist should be very careful in using extrapolations:

In an article in British D. J. a prediction was tried about the number of teeth extracted per person/year in the future (Fig. 10).

An extension of the line beyond the observed X, Y-values is a dubious operation. You can never be sure of the form of the regression line beyond the observed X, Y-values. In this case a very good fit is obtained by a curvilinear regression: (Fig. 11)

**CORRELATION**

Correlation is also an expression for the mutual relationship between two variables which, however, are in a sense equal: none of them depends on the other; there is no independent and no dependent variable.

Suppose that we wish to assess the relationship between the number of present maxil-
lary teeth on the right and left sides of the jaw in 5 persons. For each individual we have two variables, A and B (Table 11).

Table 11

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>7</td>
</tr>
</tbody>
</table>

It is obvious that it is not the number of teeth on the right side that determines the number of teeth on the left side or vice versa.

As before we can illustrate this relation by a scattergram and by a line, a correlation line, calculated in the same way as the regression line (Fig. 12, 13).

This may seem confusing. What then is the difference? From a practical point of view the difference is, that by regression analysis a prediction can be made of the probable value of the Y-variable corresponding to an X-value. Say that we would like in the regression example to find out what loss of weight can be expected at a humidity of 50%. This can be done either by the graphical method or by mathematical calculation. We will find that the weight loss can be expected to be about 80 mg (by the curvilinear regression).

To come back to correlation: As an expression for the closeness of the coordinate points of the scattergram to the correlation line, a correlation coefficient, r, may be calculated. This r may take values from +1 to -1, the sign being dependent on the direction of the correlation line. If all points lie on the line r=1. The more scattered the points, the lower the coefficient. r=0 of course means that there is no correlation at all.

Correlation and regression may be illusory, just due to the random variation. To avoid drawing conclusions from illusory correlations one must assess the significance of r. If the variable pairs are not too numerous this can be found out in an r-table, available in every elementary textbook, or else by applying a simple formula.
A common mistake is to believe, that a significant correlation also means a causal relationship. However, significance of r—if ever so high—proves only that there is a linear relation between the variables. Cause-effect relation must be assessed in other ways. It is clear that e.g. the highly significant correlation between the number of remaining teeth on the right and left sides of the jaws does not prove that tooth loss on the left side causes tooth loss on the right one or vice versa.

**RANK CORRELATION**

When conditions for parametric tests are not fulfilled, the correlation may still be estimated. A nonparametric analogue to the usual correlation analysis is Spearman’s rank correlation, which estimates the rank correlation coefficient $r_s$. This can be calculated in the same way as the parametric r using the rank values instead of the observed ones.

Example: Ten dental students were ranked (thus an ordinal scale) with respect to two variables, viz. theoretical knowledge and practical skill. The best student was given rank 1, the next best rank 2, etc. (Table 12)

<table>
<thead>
<tr>
<th>Knowledge</th>
<th>7</th>
<th>6</th>
<th>3</th>
<th>8</th>
<th>2</th>
<th>10</th>
<th>4</th>
<th>1</th>
<th>5</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skill</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>1</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>10</td>
</tr>
</tbody>
</table>

Was there any significant correlation between theoretical knowledge and practical skill? The rank correlation coefficient was found to be 0.85. By comparing this value (with n-2 d.f.) with the values listed in a rank correlation table it was found that there was in this case a correlation between knowledge and skill significant at the 1% level.

Then there is the question of nonsense correlations. A significant correlation between two variables may simply be due to both being correlated to a third variable, one that may be more difficult to detect than that in the following example.

A study comprising a number of American towns showed a highly significant correlation between alcohol consume and number of practising doctors. The explanation is, of course, not that doctors cause people to drink, but that both alcohol consume and the number of doctors are related to the number of inhabitants.

**HANDLING of pH-DATA**

As everyone knows, pH is the negative common logarithm of the hydrogen ion concentration, symbolized as $H^+$.

$$\text{pH} = -\log H^+; \quad H^+ = 10^{-\text{pH}}$$

The logarithmic scale in which pH is measured is an ordinal or rank order scale with no fixed distances between the scale divisions. $H^+$ is, of course, in an interval scale and thus with fixed distances between the scale divisions: (Fig. 14)
INDEPENDENT SAMPLES

Suppose that we have saliva samples from two groups of test persons. We want to know if there is any significant difference in central tendency between the pH values of the two samples (Table 13).

We cannot use the t-test because of the logarithmic measuring scale. To come round this we may think of transforming the pH-values to H⁺. Then we get stuck on another difficulty, viz. the highly significant inequality of variances: (Table 14)

By t-tests on pH and H⁺ values respectively, we would get entirely different results; and no one correct. The t-test is among other things based on equality of variances. We are forced to use nonparametric statistics. We decide upon the Wilcoxon rank sum test, to see if the medians differ significantly. To this end we rank both samples together: (Table 15)

The rank order is the same, only with reversed sign, which does not affect the test. Consequently, by the Wilcoxon rank sum test the same result is obtained whether based on pH or H⁺:

Lower rank sum=10; \( p < 0.02 \), i.e. \( < 0.05 \).

PAIRED COMPARISONS

Here the problem is principally different. As a non-parametric test we apply the Wilcoxon signed ranks test. Assume that we deal with double determinations of pH in saliva
samples, A and B, from each one of four individuals. For this test we first have to find the differences between the two values in each variable pair, then disregard the sign of the differences, rank them and, finally, give back the sign: (Table 16)

<table>
<thead>
<tr>
<th>pH</th>
<th>B</th>
<th>Diff</th>
<th>H⁺</th>
<th>A</th>
<th>B</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.7</td>
<td>6.5</td>
<td>0.2</td>
<td>1.995E-7</td>
<td>3.162E-7</td>
<td>-1.167E-7</td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>4.9</td>
<td>0.2</td>
<td>7.943E-6</td>
<td>1.259E-5</td>
<td>-4.646E-6</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>4.4</td>
<td>-0.2</td>
<td>6.310E-6</td>
<td>3.981E-6</td>
<td>+2.329E-6</td>
<td></td>
</tr>
<tr>
<td>6.3</td>
<td>6.1</td>
<td>0.2</td>
<td>5.012E-7</td>
<td>7.943E-7</td>
<td>-2.931E-7</td>
<td></td>
</tr>
</tbody>
</table>

Ranking of the differences: (Table 17)

<table>
<thead>
<tr>
<th>pH-diff</th>
<th>Rank</th>
<th>H⁺-diff</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>.2</td>
<td>2.5</td>
<td>-1.167E-7</td>
<td>-1</td>
</tr>
<tr>
<td>.2</td>
<td>2.5</td>
<td>-2.931E-7</td>
<td>-2</td>
</tr>
<tr>
<td>.2</td>
<td>2.5</td>
<td>-4.646E-6</td>
<td>-3</td>
</tr>
<tr>
<td>-.2</td>
<td>-2.5</td>
<td>+2.329E-5</td>
<td>+4</td>
</tr>
</tbody>
</table>

The differences between the pH values in this example have all exactly the same absolute numerical value and, consequently, the same unsigned rank, while all the differences between the H⁺ values have individual differences. Therefore, the rank sums and the results of the tests will differ. Only the Wilcoxon matched-pairs signed-ranks test gives a correct result.

**DENTOCULT**

There is a test called “Dentocult”, which is intended for use by practitioners in their offices. It is advertised as a means for dentists to assess in an easy way the approximate number of lactobacilli in saliva without the need of a laboratory.

A test tube contains a plastic slide covered with a culture medium specially suited for lactobacillus acidophilus. A saliva sample is spread over the surface of the slide, which is then incubated at 37°. The resultant growth is finally compared with a standard photographic reproduction of slides with 1000, 10000, 100000 and 1000000 colonies. The number of colonies on the test slide is assessed to the nearest one of these numbers or to 0 in the case of no growth. Obviously, it must be difficult in many cases to decide upon what score shall be preferred.

This test offers an excellent example for discussing a rather important statistical question:

Saliva samples from 162 persons were used for parallel Dentocult assessments and direct counting of colonies. The results are reproduced in the table 18.
The percentages of lactobacilli counts which correspond to each Dentocult level is shown in the table 19.

Except for the zero group the Dentocult test fell within the indicated range only in 17 to 34% of the cases. This means that an isolated test-and probably also a pretty large series of tests in the same patient-will convey only rather uncertain information about the lactobacilli count of the saliva. One could add, that the relation between lactobacilli counts and caries activity is of the same loose kind.

The important thing is sometimes said to be to distinguish between so-called "lactobacilli millionaires" from others. The chance in this sample of 162 persons was one to three.

A fallacious reasoning is common to a lot of statements in the propaganda for a variety of preparations, means and procedures. The results obtained for a big sample are supposed to be valid for the individual case. Unfortunately this kind of error is not rare even in scientific literature.

From the Dept. of Community Dentistry
University of Lund
School of Dentistry
S-21421 Malmö, Sweden