Statistical Analysis for Clinical Studies

Jürg Hüsler

10.1 Introduction

A quantitative research study has to be carefully planned with the design of the animal experiment or the clinical trial with patients held in mind. One has to formulate the research question and find existing published information on this question. The aim of the study is to gain new meaningful answers on the research question. One has to describe the population, the variables, measurements, outcomes, endpoints or parameters, and the factors that have or might have an impact on the primary endpoints. One has to fix the randomization method, which should be applied, and the blinding, if important and necessary to prevent biases, to determine how to collect the data, and how one can analyze the data with the appropriate statistical methods. One has to plan the statistical analyses in advance as per a good scientific practice, where one determines the appropriate statistical methods to be applied. This also prevents too many statistical analyses being performed purely in search of some significant findings, which are meaningless or statistically incorrect. It should also prevent the situation arising of the investigator or the statistician not being able to analyze the collected data with suitable statistical methods.

In this chapter, we discuss the most important issues on the statistical analyses related to a study or experiment. Further information can be found in the huge literature on statistics in medical research. We also mention some helpful textbooks in the references.

10.2 Development of the Data Analysis Plan

The design of the study is directly related to the statistical analyses or data analysis plan. Hence, the treatment of data and the intended statistical analyses are discussed at the beginning of the study and form part of the study protocol.

Box 10-1 Elements of a data analysis plan

- Primary and secondary endpoints to be measured
- Descriptive statistical methods to be used
- Graphics to be applied
- Statistical hypotheses
- Fixing the significance level $\alpha = 5\%$
- Statistical tests and models for answering the hypotheses

When designing the study, one has to think about how the questions can be answered with suitable and appropriate statistical methods. There exist a huge number of available statistical methods, which are very different. One is not always applying only the *t* test or the chi-squared test. The analysis plan depends on the variables to be observed. One has to formulate the primary endpoint(s) which is (are) answering the primary questions or hypotheses. Endpoints can be of different qualities; they are categorical, ordinal or metric. If the endpoint is (for example) ordinal, one has to apply appropriate methods (usually nonparametric ones), which are different from the ones used if the variable is metric or categorical. Means and standard deviations are not appropriate for ordinal data or scores, as well as for asymmetrically distributed data, e.g., survival data. In such cases medians and interquartile ranges should be applied. Categorical data are usually represented by frequencies and proportions.

The selection of the statistical methods also depends on the number of study arms, on the possible impact factors (possibly used for stratification), and the dependence relationship between the considered variables. For instance, if several implants are placed into each of the rabbits or dogs, one observes typically dependent (clustered) data of the implants within an animal. One cannot state that the data are independent without additional arguments. Then the statistical analysis has to be adapted and made appropriate for such correlated data. Note also that if several sites are used, the sites should be randomized to the implant types to prevent biased or confounded measurements and results. Simple tests are usually not possible for the comparison of different implants in the case of dependent records. If there are only two implant types or grafting materials investigated in each animal, then the statistical analysis is based on simple well-known tests; one would have to apply the paired *t* test or the Wilcoxon signed rank test depending on the type of endpoint (metric or ordinal) and whether the assumptions of the two tests are holding.

One distinguishes the endpoints by priority. The primary endpoint(s) should answer the main question and hypothesis. For this part, one has to clearly state the intended confirmatory statistical methods in the statistical plan. For the secondary endpoints, the intended statistical methods are more to explore further insights into the posed questions. Quite often these secondary answers lead to further research and studies.

Example 10.1

Statistical method section of a publication is based on the statistical analysis plan. This part in the publication mentions briefly the most important part of the statistical analysis plan. Usually it does not include everything.

For instance, we read for our mentioned grafting example:

Mean values and standard deviations as well as the 25th (lower quartile), 50th (median), and 75th (upper quartile) percentiles were calculated for each outcome variable. The main interest was focused on vertical and horizontal bone resorption. Some corresponding primary variables were stated.

Differences between test and control sites were analyzed using the Wilcoxon (signed rank) test for paired observations. The level of significance was set at $\alpha = 0.05$.

Often one also mentions the statistical software, which is used for the analysis of the data. Note that this part does not mention whether and how the randomization is applied. The hypotheses are

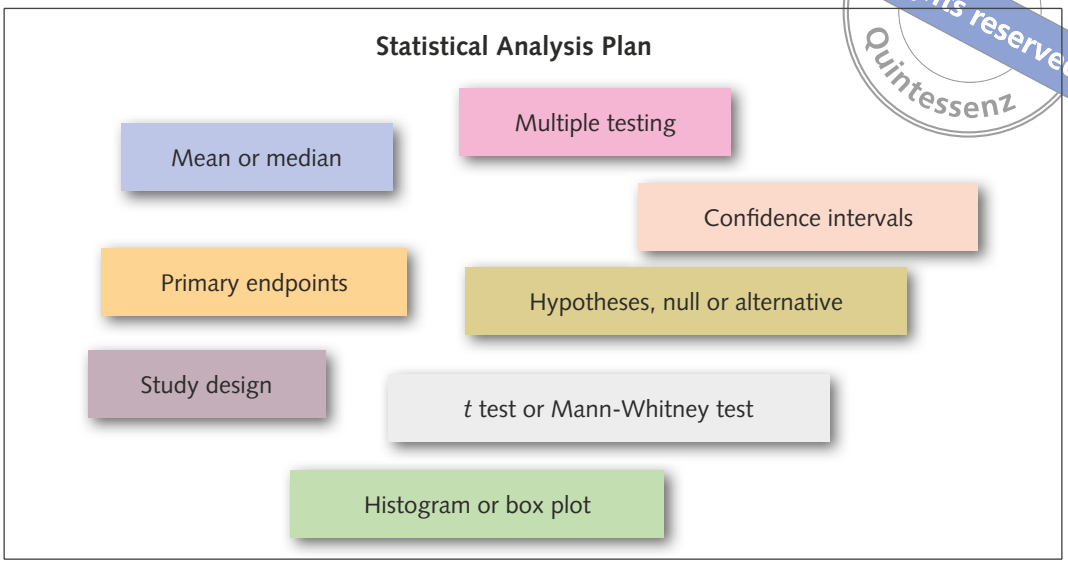

Fig 10-1 Randomization procedures.

stated typically earlier in the publication. They should be formulated as precisely as possible in the statistical sense, stating whether they are two-sided or one-sided. Usually one investigates two-sided hypotheses. The possible dependence of data is usually also mentioned earlier in the publication, or one has to derive it from the description of the study plan in the material and method section. The nonparametric tests are applied without applying a test for the normality of the data, which is reasonable because of the small sample. We note that no correction for the multiple testing of the selected endpoints is applied. When means and SDs are derived, one is assuming that all the outcome variables are metric and symmetrically distributed.

10.3 Randomization Procedures

Randomization is a principle in a trial to allocate the patients or animals to the treatments (Fig 10-1). The randomization is needed to ensure that the characteristics of the patients or animals do not confound with the treatment outcomes. It ensures that the outcome is not biased. This randomization is certainly needed in a randomized controlled clinical trial. One expects that in a randomized study, the subject characteristics are more or less balanced, e.g., that the ages or gender distributions of subjects are comparable in the treatment groups. Hence, we expect unbiased results on the treatment effects, on the primary endpoints, if the randomization is correctly applied.

Box 10-2 Randomization:

- Simple randomization
- Stratified randomization
- Balanced randomization
- Computer generated random numbers or codes

Other sampling methods are also known, for instance the convenience sampling (posing the problem of generalizability), the quota sampling (providing no good estimate of the true parameters), or cluster sampling (which is less efficient than random sampling). These sampling methods should not be used in scientific osteology studies.

Hospital numbers, birth dates or social security numbers are, for example, inappropriate numbers for the randomization.

Usually, one applies blinding of the treatments in a randomized controlled study to prevent biased treatment effect data. The randomization in osteology studies is usually used for assigning a patient or an animal to a certain treatment, using one of the possible implants or one of the possible grafting materials. It is also used, for instance, to assign the grafting material to the planned sites within an animal.

Example 10.2

If one compares two treatments, one has to allocate a patient to group 1 or 2 by a random number, as with a coin. This may result in unbalanced group sample sizes. For instance, if we throw a coin 10 times, one may get the sequence of head and tails, coded as 1 and 2, respectively – 1, 2, 2, 1, 1, 1, 2, 1, 2, 1 – hence 6 times a "1" and 4 times a "2." If one wants a balance with 5 times a "1" and 5 times a "2," one has to randomly permute the 10 digits, consisting of 5 "1"s and 5 "2"s, with the help of computer software. This may result in 2, 2, 1, 2, 1, 1, 2, 1, 1 and 2. We would get rarely the systematic series with 1, 2, 1, 2, 1, 2, 1, 2, 1 and 2.

If grafting is applied on each side of the tibia of a minipig, then the treatment and the control should be randomly assigned to the side. This should prevent a confounding bias. Hence, one selects with a random number generator or a coin the "1" or head for a left side, and "2" or tail for the right side.

The randomization with more than two treatment groups is little more complicated, but easily done with statistics software. Randomization is performed nowadays with a computer software pseudo-random number generator. One should not perform the randomization by hand.

One may select the randomization sequence with the software during the planning of the experiment and fix this derived randomization sequence in a list, which is used during the trial when a new patient or animal is entering the study. For instance, one may put the randomization code in a sealed envelope for every patient; the envelope is opened when a patient enters the trial. The random code is indicating the treatment. If possible and if needed, the treatment is blinded to the operator and patient. It is best that the code is only known to a person independent from the study team. The code is kept secret until the end of the trial.

Another randomization method is possible by using computer software, which presents the code interactively. A new code is created or derived when a new patient or animal is included in the study to be assigned to one of the treatment groups. Also the performed randomization has to be kept secret (if possible) until the end of the study, when the codes are broken for the statistical analyses. The recruitment team should not know the next randomization code, because it might have an impact on the selection of patients.

In large studies, one stratifies the important prognostic or impact factors. In such a case, one uses the aforementioned simple randomization for each stratum of such an impact factor. By this stratified procedure, one expects to have allocated a balanced number of patients to the treatments within each stratum. This is shown in Example 10.3.

Example 10.3

A dentist wants to randomize 20 patients to a grafting treatment and a control group. For example, if the gender is an impact factor, the randomization by strata guarantees that roughly the same number of women as men receive the different treatments.

Since one does not know in advance how many men and women will be recruited, one needs a list of random numbers for the men and the women that is sufficiently long. Certainly if each list contains 20 random numbers, they are sufficiently long.

Furthermore, one will stop recruiting after the 20th patient. This would eventually result in an unbalanced number of treated and control

patients. To prevent this outcome, one balances the random number in each group so that the sequence of random numbers is always balanced after, for instance, four numbers. One may also select a different number for the balancing of the randomization.

For instance, one would use the following random sequences with "1" indicating the treatment and "2" the control. For example:

- For men: 1, 2, 1, 2, 2, 2, 1, 1, 2, 1, 2, 1, 2, 1, 1, 2, 1, 2, 2 and 1.
- For women: 2, 1, 1, 2, 1, 1, 2, 2, 1, 2, 1, 2, 2, 1, 1, 2, 2, 2, 1 and 1.

Please note that in each block of 4 numbers, one has twice a "1" and twice a "2." If one stopped at the 20th patient, having recruited 7 men and 13 women, one would have 3 men in the treatment group and 4 men in the control group, and 6 women in the treatment group and 7 in the control group. This is a sufficiently balanced sampling with stratification. Figure 10-2 shows an example with three treatments.

10.4 Power Analysis and the Derivation of the Sample Size

If the design is feasible, one should also know in advance the necessary effort to get a good reliable answer on the research question. The effort is related to the sample size. If the sample size selected is too small, hence the effort is too little, then one risks not getting a clear answer. If the sample size is large or too large, then one finds a clear answer, but the effort is too large and one includes too many animals or patients. This method is also ethically not correct. Thus the sample size should be selected with large chance, so with the so-called statistical power, one finds a medically meaningful answer or decision. This answer is based mostly on a statistical test result, usually a significant one.

Let us discuss the derivation of the necessary sample size for a medical study. Several statistical terms are involved in this derivation. Therefore, we need to understand the statistical terms: the hypotheses; the significance level α (usually α = 0.05); the power of the test; the statistical test to be used for the primary endpoint and the primary question; the clinical relevant effect which gives a medically meaningful answer; also the distribution of the data of the primary endpoint. If certain factors have an impact on the endpoint, this should also be known and considered in the power analysis.

Box 10-3 Sample size and power determination depends on:

- Null and alternative hypotheses (H_0 and H_1)
- Medically relevant effect to be detected
- Variability of the measurements
- Statistical test to be applied
- Error probabilities and power

10.4.1 Hypotheses

One starts with the statement or (statistical) hypothesis that we wish to show. Since we can only falsify and never prove a hypothesis, a claim is proposed as an alternative hypothesis, a so-called H_1 hypothesis, as explicitly as possible in statistical terms. For example, we can suppose that the buccal bony wall thickness has an impact on the hard tissue changes at implants after three months. Two sites with different thicknesses (1 mm or 2 mm) are selected. The statistical alternative hypothesis is now formulated: the mean changes at the two sites are different.

The contrary is the so-called null hypothesis H_0, which we want to falsify or to reject with a statistical test using the data. For this example, we would formulate the null hypothesis as: "The mean changes at the two sites are not different."

The H_1 hypothesis can be stated as two-sided (e.g., the mean BIC (bone-to-implant contact) $\mu 1$ of one implant type is different to the mean BIC $\mu 2$ of another implant type) or one-sided (for example, the mean BIC μ is after a certain time larger than $\mu 0 = 50\%$ for a certain implant type).

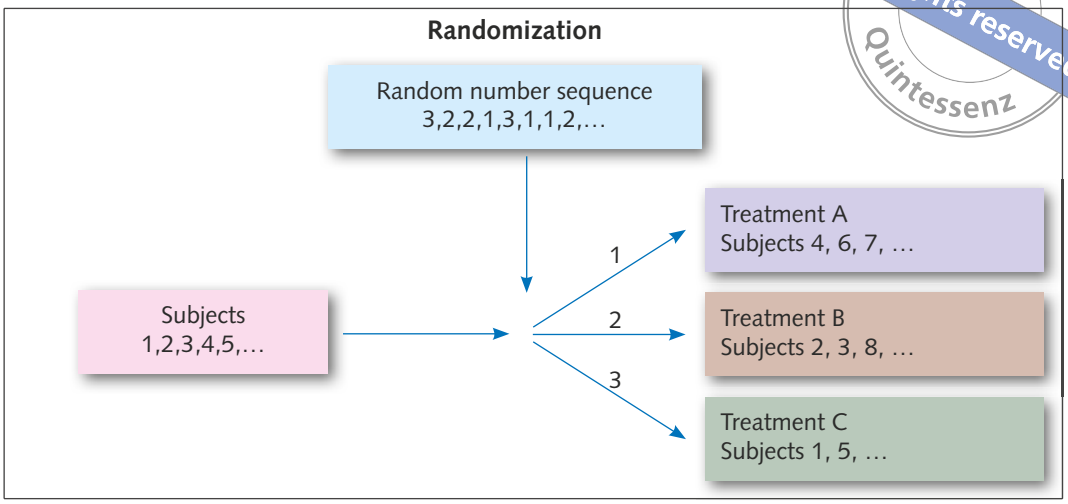

Fig 10-2 Elements of a statistical analysis plan.

Example 10.4

Let us consider the simple two group comparisons to investigate a new grafting material. One starts with the main question to be investigated.

- *Motivation:* The grafting treatment in defects has an impact on the primary endpoint "new bone volume in relation to the total volume" (BV/TV). This variable is a metric one. The new material is compared to a control treatment. One has two groups (treatment and control) of data to compare.
- *Statistical hypotheses:* H_0: $\mu T = \mu C$ versus H_1: $\mu T \neq \mu C$ where μT indicates the mean (or median) of BV/TV in the treatment group and μC indicates the mean (or median) in the control group. Note that statistically one compares means (or medians). This comparison is more stringent than the formulated statement of the motivation. One does not compare the whole distribution.

10.4.2 Decisions and Errors

The decision of a statistical test either rejects or cannot reject the null hypothesis. Since we do not know the truth, namely the true hypothesis, this decision can be correct or false.

If the null hypothesis is true, we may reject this hypothesis by chance. The probability of doing this false decision is called α, and the error as type-1 error. For example, we may conclude by chance that the means BIC of two implants are different but the truth is that they are equal. This result is also called a false positive one.

Contrary, we may not reject the null hypothesis, in case the alternative hypothesis is true. In this case, we would conclude that the two means BIC are not different, but the truth is that they are different. This result is a called a false negative. The probability of this false decision, called type-2 error, is denoted by β.

In the statistical work we wish to perform only seldom an error or false decision. Thus, the error probabilities α and β should be small. Typically, we select $\alpha = 0.05$. Also β should be small, too, and surely not larger than 0.2.

Note also that one cannot reject the null hypothesis, it does not mean that the null hypothesis is true, because of the falsification principle. This erroneous interpretation is unfortunately often communicated. In such a case, the decision is not conclusive.

10.4.3 Power

The determination of the probability β depends on further ingredients and can be rather complicated. The probability of a correct decision, if H_1 is true, is called the power of the test; so the power is 1-β. This value should be at least as large as 0.80. It should also be noted that selecting a larger power (e.g., 90%) needs a larger sample size.

10.4.4 Sample Size Derivation

The determination of the sample size, the goal of this section, is based on formulaes for the different study investigations and statistical tests or methods. The most common cases are investigations of means or of proportions. We may present our result as an estimate with a confidence interval or we may compare estimates of different groups by a statistical test. Both methods depend on the sample size. The power and sample size derivations depend on additional items, as follows.

For the case of statistical comparison of means, one has to select the medically relevant effect (difference of two means) of the primary endpoint. If one does not know it, it may eventually be found in existing scientific literature or in a pilot study (up to a certain accuracy). One has also to know the variability of the primary endpoint. This value can possibly also be found in the literature (up to a certain accuracy) or by performing a pilot study to get an estimate of this variability. Furthermore, the distribution of the data should be known approximately. This could also be found in a pilot study. The latter is important to select the appropriate statistical test procedure, which is fixed in the statistical analysis plan (Fig 10-2). For example, normally distributed data allow us to use a parametric ANOVA in a small sample case. If this is not holding one has to apply so-called nonparametric statistical methods. With these items, one uses statistical software to derive the sample size needed to have a power of 0.80, for example.

Example 10.5

We consider the following simple statement in a publication. We have changed some of the given formulations to avoid the particular case and to make it more general.

Assuming a standard deviation of the primary endpoint of 1.0 mm and an expected intergroup difference of 1.0 mm, a sample of 40 patients (20 patients per each of the two groups) had a power of 89% in detecting a significant intergroup difference (at significance level $\alpha = 0.05$) with a two-sided test.

Note that the statistical test of this two-group comparison is not mentioned. We have to guess that the investigators want to apply the common *t* test or the Mann-Whitney test for two independent groups. No word is given whether the primary endpoint is normally distributed or has another distribution. They did not fix the power to be 80% or 90%, as usually. It seems more that the sample size is fixed to be 20 per group. The standard deviation of the primary endpoint is given (SD = 1 mm) and the effect, which should be detected, also 1 mm. Is this effect of 1 mm medically relevant? It is as large as the SD. This does not mean that the effect is already medically relevant. The alternative hypothesis is stated two-sided, comparing the means of the two groups. Then some statistical software was used to derive the necessary sample sizes per group. Deriving the sample sizes needed for the power of 90%, we get n = 21 per group based on the Mann-Whitney test (using the logistic distribution) or n = 22 per group based on the *t* test. If we set the power to 80%, then the sample sizes are n = 16 for the nonparametric Mann-Whitney test and n = 17 per group for the parametric *t* test. We note that the distribution has an impact on the power derivation. Note that one selects α = 0.05, not *P* = 0.05.

10.5 Essential Statistical Analysis Tools

10.5.1 Descriptive Measures

Statistical analysis starts with the description of the data. The description involves summarizing the data into so-called statistical measures. For the location of the data, one usually uses the

Box 10-4 The major statistical tools can be grouped into:

- Descriptive measures
- Graphs
- Statistical estimates and confidence intervals
- Statistical tests
- Statistical models for more complex questions

mean, the median, the minimum and the maximum of the data, sometimes also quartiles or certain percentiles (e.g., the 5 and 95 percentiles). Then one describes the variability of the data with the standard deviation or the interquartile range (difference of upper and lower quartile). More measures are available, for instance for the skewness (asymmetry) or symmetry of the data. One should use not only mean and SD for metric variables.

One has to know the meaning of these measures, the properties and fallacies. Some are more robust than others. By robustness, one thinks that a descriptive measure can be strongly influenced by one or a few outliers. For instance, the mean is not robust, whereas the median is rather strongly robust. To be more explicit, if one value of the observations is large or an outlier tending to ∞ (say), then the mean tends also to this outlier and tends to ∞ if the outlier does, whereas the median is not at all changing in value. Many outliers have no impact on the median, contrary to the mean. With many it means that if less than 50% of the data tend to $\pm\infty$, they have in the worst case no impact on the median (see Example 10.6). It is similar to the situation for the standard deviation, which is not robust against outliers; however, the interquartile range is robust.

Example 10.6

Let us consider a sample of n = 11 data of Bayesian information criterion bone-to-implant measurements. The BIC values are shown in Fig 10-3 as black dots on the upper axis. We plot the mean (circle) and the median (being equal to one of the observations, namely the 6th smallest one) also. Then we select the four largest observations and move them to become outliers (by adding 35 units). This data set is shown on the lower axis together with the new mean and median. This shows clearly that the median is not affected by the four outliers, but the mean moved away to larger values following the outliers.

Note that in the first case the median (= 37.5) is almost equal to the mean (= 36.0). We note that the distribution is quite symmetrical. In the second case, the mean (= 48.7) is larger than the median. More than half (7/11) of the data are smaller than the mean. The median is still the same value. The mean has moved by 12.7 units (12.7= (4 x 35)/11).

10.5.2 Graphs

To describe the data, one also applies graphs that visualize the distribution of the data (see Fig 10-3) or one may show only a few aspects of the data. Again graphs that are related to the types of data are applied. One typically uses bar charts and rarely pie charts for categorical data and histograms, box plots, and mean-SD plots (Fig 10-4) for numerical data. Note that Fig 10-4 is much less informative than the scatterplot of Fig 10-3. A table with means, medians and SDs would often be sufficient, instead of such a simple graph with little information.

Scatter plots are applied when one is interested in the relationship between two numerical variables. But we can also use a scatter plot for one variable alone as in Example 10.6.

Some of the graphs are not appropriate for small samples e.g., a box plot should not be used for sample sizes less than 25 and histograms not for sample sizes less than 50. In small samples it is better to show the data in scatter (or dot) plots.

There are many other plots that try (for instance) to visualize certain aspects of higher dimensional data. However, do not use 3D effects if the third dimension is meaningless, since it manipulates the graphical impression negatively. Also be careful with the use of colors and the shading of bars. A

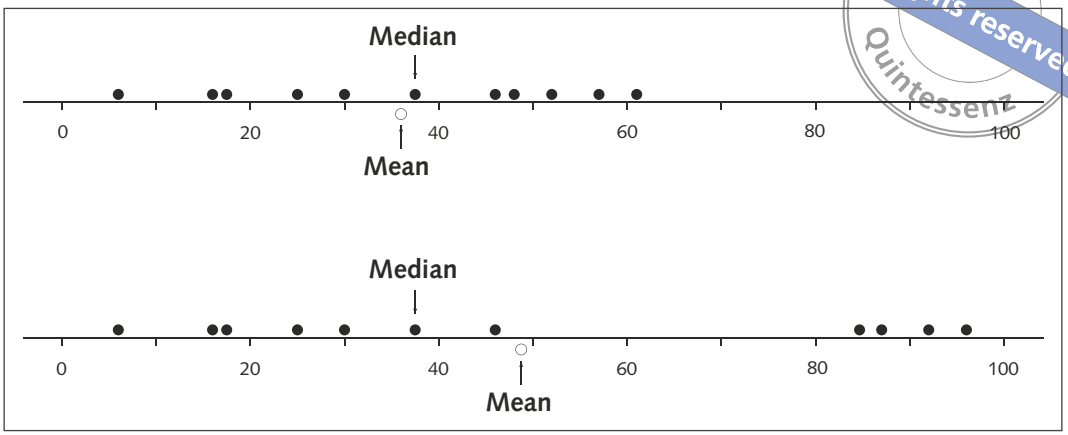

Fig 10-3 If mean and median are different in a sample, it tells us that the distribution is possibly not symmetrical or that there are a few large (or small) values or outliers; hence, it cannot be a normal distribution (Example 10.6).

graph should communicate interesting information on the data and visualize clearly the results of the main question. It should be as concise as possible.

10.5.3 Statistical Estimates and Confidence Intervals

Statistical answers are usually combining the data, filtering certain information out of the data, and presenting this information with the descriptive measures as means, medians and standard deviations (SDs). Since the data in the sample are collected at random, the results are also random, and are not equal to the true population values. The sample mean, for instance, estimates the population mean. Depending on the sample size, this estimate is more or less accurate. One knows that the sample measures (mean, median, SD) are rather close to the population measures (mean, median, SD) if the samples are very large. However, in the usual experiments and clinical studies in osteology, one does not have very large samples. Hence, one should know the accuracy of these statistical measures, or the so-called variability, measured often by the so-called standard error. If this standard error is small, one has more confidence that the sample measure is close to the population measure.

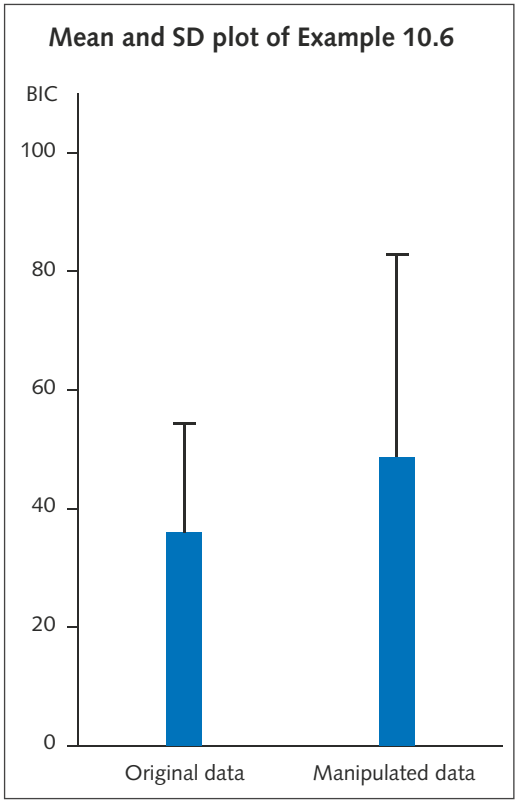

Fig 10-4 Mean and SD plot. One cannot detect possible outliers or an asymmetry of the data distribution.

A good statistical principle is to determine so-called confidence intervals, which combine the estimate of a population measure with its variability to show its accuracy. In particular, if a test result is non-significant, maybe resulting from a low power, one should present the possible difference of means or proportions by a confidence interval.

This estimation is expressed as an interval meaning that any of the values in the interval might be the population measure with good confidence (typically 95% confidence probability). These values are in good accordance with the sample information. One says also that the values outside the confidence interval would be rejected by the corresponding test if one states this value as the hypothetical value. To explain this more explicitly, let us consider a simple example.

Example 10.7

Let us use the data from our Example 10.6 above without outliers (upper axis). We note that the sample is small (n = 11) and that the data are not normally distributed. Hence it is appropriate to apply the nonparametric test, here the Wilcoxon signed rank test for the one sample case. One would expect that the median is for instance larger than 40%. Then we describe the median with the corresponding nonparametric 95% confidence interval, which is (21.75, 50). It indicates that any value between 21.75 and 50 is in accordance with the data as a population median. If we pose the hypothesis that the population median is 40, say, then the Wilcoxon signed rank test would not reject the null hypothesis (with $\alpha = 0.05$), telling us that this value is in accordance with the data information. However, if we pose the hypothesis that the median is 60%, then the Wilcoxon signed rank test would reject the null hypothesis and tell us that the data are not in accordance with the null hypothesis. The same holds for any other value outside the confidence interval (21.75, 50).

This principle holds for any other test and its corresponding confidence interval, as for instance the *t* test and its most often applied parametric confidence interval based on mean and standard error.

10.5.4 Statistical Tests

A major part of the statistical analysis consists of statistical tests. A statistical test weighs the observed result of the sample to give some evidence on the findings in relation to the stated hypotheses. Above, we mentioned already some well-known and often-used tests. In the statistical literature, there exist many tests for different comparisons and hypotheses. One most often compares means or medians of a metric endpoint or compares frequencies or proportions of a categorical variable. The selection of the appropriate test for the posed question or hypothesis is often not a difficult task when one applies certain principles.

Firstly, one should know how many groups are compared and if these groups somehow depend on each other. For instance, if one has repeated measurements from the same subject (e.g., in time), then the data of the groups are dependent. This happens also in split-plot animal studies where the data within the same animal are dependent.

The selection of the appropriate test for the posed question or hypothesis also depends on the sample size and the underlying distribution of the data. If the normal distribution holds for the data, then a parametric test (e.g., the *t* test or the F test of an ANOVA [analysis of variance, which compares more than two means]) can be used for the comparison of the groups. If the data of two groups are pairwise dependent, one applies the paired *t* test, if they are independent, then the unpaired *t* test. In addition, if the data are not normally distributed, one cannot apply the *t* tests in small samples. If the sample sizes of the data in each group are large, bigger than 50 or more, and if the data are not strongly asymmetrically distributed, then one may apply the appropriate *t* test because of theoretical reasons (so-called central limit result). In the other cases, it is more appropriate to apply the nonparametric tests, such as the

Box 10-5 Selection of an appropriate test is based on:

- The hypothesis and the measures to be compared
- The type of the variable or endpoint (categorical, ordinal or metric)
- The number of groups to be compared
- The sample sizes
- The independence or dependence of the data
- The distribution of the data

Box 10-6

Pitfalls: Do not use many statistical tests without any correction of the type-1 error rate α, since one increases the number of false positive results in a paper. Fix the primary endpoints, which should be only few, one or two, where one uses eventually a correction of the type-1 error rate. For secondary endpoints one does not correct the type-1 error rate, because one accepts such significant results in the explorative meaning indicating interesting further research.

so-called Wilcoxon signed rank test for pairwise dependent data, or the Mann-Whitney test for two independent data groups. They are based on some mild assumptions on the shape of the distribution of the data. They can be used also for large samples. They can be better and more powerful than the *t* tests, depending on the type of the data distribution. Similar is the discussion with more than two groups. One may apply parametric or nonparametric ANOVA methods: the parametric F test and the nonparametric Kruskal-Wallis test for the comparison of more than two independent data groups. All these tests can be easily determined with the help of most statistical software.

A good conservative principle is applying a nonparametric test, if one is not sure on the normal distribution of the data. Validation of the normality assumption is a bit unreliable or impossible in small samples. Applying a test for normality, as for instance the Kolmogorov-Smirnov or even the more powerful Shapiro-Wilks test, is not really helpful because of their low powers in small samples. Do not apply a statistical method and test if their assumptions are not holding. Note that if the test for normality does not reject the null hypothesis, it does not mean that the data distribution is a normal one. It depends on the power of the normality test which is small in small samples.

Example 10.8

Let us consider the two-sample comparison. We compare the means or the medians if the data are scores or are not symmetrically distributed. We assume that the variable is a metric one. We have to decide whether the two samples are independent or dependent (in this case pairwise matched).

We consider first the case of two dependent samples, which are pairwise matched. For instance, one observed bone values at baseline and 4 months after implantation for each of 23 patients. The sample size n = 23 is not large. Now we have the choice of applying the (parametric) paired *t* test or the (nonparametric) Wilcoxon signed rank test. This choice depends on the distribution of the data, in this case on the differences of the paired values. If this distribution is a normal one, one may apply the *t* test. Otherwise we have to apply the Wilcoxon signed rank test (if the differences are somehow symmetrically distributed around the median of the differences). If one is not sure about the normality of the data, it is better to apply the nonparametric test. Since n = 23 is a small sample size, it is safer to apply the nonparametric test. Often the parametric and nonparametric tests show the same decision; however, the parametric one might be incorrect.

10.5.5 Statistical Models

What is a Statistical Model?

Finally we mention the statistical models as possible tools in the statistical analysis. With a model, one tries to analyze the relationship between observed variables. Quite often, one has a primary endpoint, which may depend on some factors. For instance, one investigates the ingrowth of an implant. The BIC is the primary endpoint. Influencing factors might be: the types of implant; the material of the implant; the surface treatment of the implant; the size (length or diameter). Typical questions to answer are whether the BIC depends on one or on several of these factors.

Another example: when one compares three different grafting materials in a study, where each material is placed twice (bilaterally) in a dog's mandible, one needs a model to answer the questions on the impact of grafting material and site (left or right) as the explanatory factors. In general, the model is assuming the type of influence, and the dependence of the measured bone values within a dog. The model depends on the type of the endpoint (metric, ordinal or categorical) and the influences of the explanatory factors (linear, nonlinear or a specified influence, interacting). All these ingredients have their influence on the selection of the appropriate model. One cannot apply simple tests.

Statistical models that are not as simple as two group comparisons should be better applied with an experienced scientist or discussed with an applied statistician.

Example 10.9

For instance, say we want to investigate the BIC, a metric variable, for the comparison of three different implant types applied in each dog bilaterally. Hence, we typically select an ANOVA (analysis of variance) model. However, because of the dependence within a dog, we have to apply a more complicated model, in this case a so-called mixed or repeated ANOVA model (a generalization of the simple ANOVA model), which is taking care of the dependence. However, one has to include the dependence appropriately. Furthermore, one has to decide whether a parametric or a nonparametric model is suitable. This means one has to argue whether the endpoint, for instance the BIC, a bone value or a score, has a normal distribution or not. In the latter case, the so-called nonparametric analysis would be based on ranks of the data.

Simpler is the ANOVA model analysis if independent data are collected, if one considers only one implant in each patient or dog. One calls the ANOVA model a one-way ANOVA, if one has only one impact factor (e.g., only implant type); a two-way ANOVA, if one has two impact factors (e.g., implant type and side); or a more general ANOVA, if one investigates more than two impact factors (implant types, age, gender, smoker status, diabetes status). In case of two or more impact factors in an ANOVA model, one has to also analyze possible interaction influences between the impact factors.

Multiple Testing

If the ANOVA analysis shows a significant result, saying that the three implant types have shown significantly different BIC means and no significant impact of the implant sites, one wants to answer which of the three implant types are significantly different with respect to the BIC mean. Then one applies the appropriate so-called post-hoc two-sample tests. In our example, three two-sample tests are used in total: the first test for the comparison of the first with the second implant type; the second one for the comparison of the first with the third implant type; and the third one for the comparisons of the second with the third implant type. This testing is called a multiple testing situation, since one uses the same data for multiple, here three, tests. However, we have to correct the significance level for these three two-sample tests in case of multiple testing. There exist several devices for such a correction, which depend on the assumed model. Common procedures are Tukey's correction (honest signifi-

Table 10-1 Probability of at least one false positive decision when performing k independent tests, if all null hypotheses are true.

Number k	1	2	3	4	5	10	20
Probability *P*	5%	9.75%	14.26%	18.55%	22.62%	40.13%	64.15%

cant tests), Bonferroni's correction, Newman-Keulbs, Scheffé, and Dunn devices. These corrections are usually not exact. They are correcting the probability of a type-1 error too much. But these devices are better than not correcting, and instead apply the so-called Fisher (LSD, least significant device) tests. The LSD procedure inflates the number of the false positive results in the study, since it does not correct the type-1 error rate.

To indicate this with some numbers, let us consider the case of k independent tests, which could be performed in the statistical analysis of the study data. Then the probability that at least one of these k tests shows a significant result, assuming that all null hypotheses are true, is larger than the intended $\alpha = 5\%$; one gets the following false positive probabilities *P* in relation to the number k of tests (Table 10-1). Because of the multiple testing in any study, one should present the *P* values of the performed statistical tests numerically and not logically, as $P < 0.05$ or $P < 0.01$. Only if the *P* value is less than 0.001, then it is adequate to write $P < 0.001$.

References

1. Altman D. Practical Statistics for Medical Research. London, UK: Chapman & Hall/CRC, 1991.
2. Armitage P, Berry G, Matthews JNS. Statistical Methods in Medical Research, ed 4. Oxford, Malden, MA: Blackwell Science, 2002.
3. Lwanga SK, Lemeshow S. Sample Size Determination in Health Studies. Geneva: WHO, 1991.
4. Peacock JL, Peacock PL. Oxford Handbook of Medical Statistics. Oxford, UK: Oxford University Press, 2011.