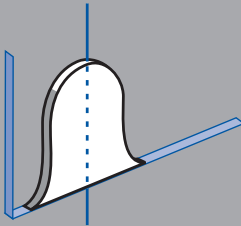
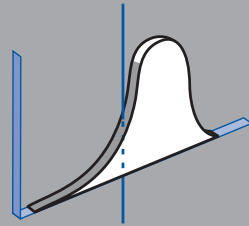


# 5 How to analyze your data

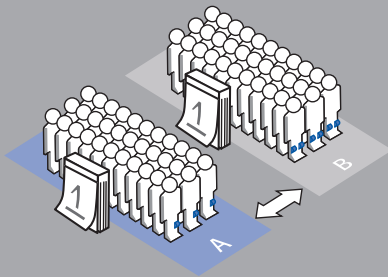


Parametric

or

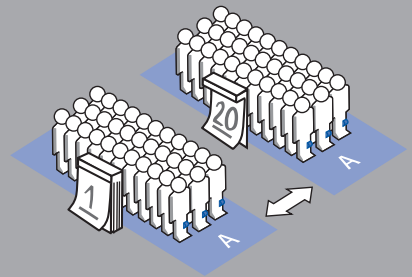


Nonparametric



Unpaired

or



Paired



Categorical

or



Continuous

## 5 How to analyze your data

<b>1</b>	<b>Statistical tests: the basics</b>	<b>95</b>
<b>2</b>	<b>How to choose the appropriate test</b>	<b>96</b>
<b>3</b>	<b>Binary or categorical data</b>	<b>98</b>
<b>4</b>	<b>Ordinal data</b>	<b>100</b>
<b>5</b>	<b>Group comparisons involving continuous data</b>	<b>102</b>
<b>6</b>	<b>Comparison of more than two groups</b>	<b>106</b>
<b>7</b>	<b>Analysis of paired data and other extensions</b>	<b>107</b>
<b>8</b>	<b>Summary</b>	<b>109</b>

## 5 How to analyze your data

### 1 Statistical tests: the basics

Statistical methods for data analysis have a long history, and biostatistics has emerged as an own, prospering field of research. Powerful statistical software has been developed that allows for thorough calculations on large datasets that would be impossible to be performed manually. Some of the available methods are complex and difficult to apply, while others can easily and successfully be employed by researchers without a strong statistical background. Luckily, the most common statistical analysis methods in biomedical research belong to the latter category. Don't let statistical software seduce you to do computations you cannot understand—if in doubt keep it simple, since you are responsible for the interpretation of the results.

The most common statistical approach in biomedical research is the analysis of differences that may exist between two or more groups of patients: Is postoperative pain intensity lower in patients after minimally invasive versus conventional total hip replacement? Are women more likely than men to develop anxiety disorder after severe injury? Are extraarticular type A fractures associated with better functional prognosis than intraarticular type B or complex type C fractures? You may argue that, if the difference is large enough, statistical tests are dispensable—and you are right. However, in case of moderate or small differences, the key question is, whether the observed effect has occurred simply by chance. This is what statistical tests are made for.

*Statistical tests are tools that distinguish between results compatible with chance, and those that no longer can be explained by chance.*

Also, all statistical tests share the same principle—they compare the observed results with an expected value, based on your dataset, and come up with a so-called test statistic. This statistic is compared to a tabulated value derived from the underlying distribution. If the statistic is higher than a certain critical or threshold value, the

difference between observed and expected results is no longer a matter of chance. All of these different steps of computation and comparison are nowadays made by statistical software.

*In research practice, it is important to know which tests should be used for which kind of data, and why a particular test may or may not apply to your research question.*

## 2 How to choose the appropriate test

The good news is that the choice of the appropriate statistical method for comparing groups is often straightforward. In most cases, the suitable method depends only on two criteria—the number of groups (two versus more than two) and the data type (binary, categorical, ordinal, continuous) involved in the comparison. Only when group differences with respect to a continuous variable are analyzed a third aspect, namely the choice between “parametric” and “nonparametric” methods, plays a role. When the data can be assumed to follow a normal (Gaussian) distribution in each group, a parametric method is appropriate. Nonparametric, or so-called distribution-free methods can be used in those cases where this assumption does not apply (in fact, they can be used for all types of data, but this is a little too simple).

Many popular statistical tests for analyzing differences between groups, such as the t-test, analysis of variance (ANOVA), or the chi-square ( $\chi^2$ ) test can be integrated in a framework of analysis methods based on our three decision criteria—number of groups, data type, and assumption of normal distribution. An overview of some common methods for analyzing group differences based on these criteria is shown in [Table 5-1](#).

		<b>Data type</b>			
		<i>Binary or categorical</i>	<i>Ordinal</i>	<i>Continuous</i>	
<i>Number of groups</i>				<i>Normal distribution assumed</i>	<i>Normal distribution not assumed</i>
2	<i>Descriptive statistic significance test</i>	<i>Proportion chi-square test</i>	<i>Median Mann-Whitney U test</i>	<i>Mean value t-test*</i>	<i>Mean value Mann-Whitney U test**</i>
3+	<i>Descriptive statistic significance test</i>	<i>Proportion chi-square test</i>	<i>Median Kruskal-Wallis H test</i>	<i>Mean value ANOVA (F-test)</i>	<i>Mean value Kruskal-Wallis H test</i>

**Table 5-1** Appropriate methods for statistical analysis of differences between groups.

\*) sometimes referred to as "Student's t-test".

\*\*) also known as "Wilcoxon rank sum test".

When only two groups will be compared a simple example can illustrate the use of this table:

**Example** Suppose a fictitious randomized clinical trial of conservative versus operative treatment of fractures of the scaphoid.

Only patients who are working at the time of the injury are enrolled in the study. Duration of sick leave represents the primary endpoint. A total number of 60 patients are randomized to receive either conservative (short-arm cast) or operative treatment (Herbert screw fixation).

The occurrence of complications such as malunion, nerve compression, or wound infection (yes/no) represents the primary endpoint. Secondary endpoints comprise ratings of pain and discomfort at 6 months after the injury (no pain or discomfort, pain or discomfort at strenuous exercise, pain or discomfort at minor exercise, continuously). Also, patients are followed-up and the time until return to work will be recorded.

*The onset of a complication is binary, or dichotomous—a patient will or will not encounter an adverse event. Hence, we would compare proportions (or percentages) and use the chi-squared test.*

*Pain and discomfort is an ordinal variable comprising only three categories. Here we would compare the medians and use the Mann-Whitney U test.*

*Duration of sick leave is a continuous variable (it may, in theory, range from zero to hundreds of days). The difference between the two treatment groups can be analyzed by comparing the mean values of this variable. If the data are normally distributed, the t-test would be the appropriate statistical test. Otherwise, the nonparametric Mann-Whitney U test would be the relevant method.*

### 3 Binary or categorical data

The primary endpoint in our example, occurrence of complications, was measured in two categories (0: no complication, 1: one or more complications). This binary variable can easily be analyzed using proportions or percentages (percentages are proportions multiplied by 100). In [Table 5-2](#), the fictitious results for this binary variable are shown.

	Treatment group		Chi-square test
	Conservative (N = 30)	Operative (N = 30)	P value
Percentage with complications	10.0	20.0	0.278

**Table 5-2** Statistical comparison of occurrence of complications in patients after conservative and operative treatment.

It turns out that the incidence of complications in the conservative group (10 %) was lower than in the group with operative treatment (20 %). While this difference of 10 % in favor of the conservative group may be relevant from a clinical point of view—depending on the kind and severity of complications—a statistical test should indicate whether this difference can have been produced by chance. The null hypothesis in this case is that the percentage of complications is the same in both groups and equal to the marginal percentage of complications when both groups are combined (ie, 15 %). The chi-squared test (denoted by the Greek letter “chi” to the power of 2:  $\chi^2$ ) can be used to test this null hypothesis.

Results of this test are also included in [Table 5-3](#). Because the  $P$  value is greater than the prespecified significance level of 0.05 we conclude that the null hypothesis, the proportion of complications is equal in both groups, can not be rejected. Hence, it is decided that the observed difference was produced by chance.

*The chi-squared test can be used in situations where binary data for just two groups are being compared.*

*In this case the chi-squared test is equivalent to a significance test of the odds ratio (OR) or the risk ratio (RR). When testing the odds ratio or the risk ratio the null hypothesis is that  $OR=1$  and  $RR=1$ , respectively.*

*The chi-squared test can also be used in situations where binary data for more than two groups are being compared. The previous comments about difficulties with pairwise comparisons using nonparametric methods apply here also.*

*Application of the chi-squared test requires the sample size to be “large enough”. A rule of thumb states that this condition is met when none of the expected frequencies calculated according to the null hypothesis is smaller than 5. If this is not the case it is recommended that either a corrected version of the chi-squared value (applying the “Yates’ correction”) or a method known as “Fisher’s exact test” is used. Most statistical programs provide results for all three tests, P values for the chi-squared test, the corrected chi-squared test, and for Fisher’s exact test.*

#### **4 Ordinal data**

In our example, pain and discomfort was assessed using a discrete variable with a small number of categories (A: no pain or discomfort, B: pain or discomfort at strenuous exercise, C: pain or discomfort at minor exercise or continuously). With only three levels the ratings of pain and discomfort are not really a continuous variable. It must further be assumed that the distances between adjacent response categories are not equal. However, the categories are clearly ordered. For analyzing the results for this endpoint, the Mann-Whitney U test is appropriate (see [Table 5-1](#)).

When using the U test, instead of comparing values of the arithmetic mean, we rank order the entire data set and compare the mean ranks obtained for the two groups. The distribution of the pain and discomfort ratings and respective results from the Mann-Whitney U test are displayed in [Fig 5-1](#).

The marked differences in the pain and discomfort ratings between the two groups are in favor of the operative treatment. This difference in the rank ordered data is also statistically significant ( $P=.033$ ), demonstrating that not only the sick leave data but also the patient reported outcomes show differences in the same direction.

For comparisons of more than two groups the nonparametric equivalent to the F-test used in the analysis of variance is the Kruskal-Wallis H test. Again, a significant result of this test only indicates that the data are not compatible with the null hypothesis of no differences between groups. The test result will not tell us which of the groups differ significantly from each other. In contrast to the parametric analysis of variance, where a number of methods for pairwise comparisons are available which avoid overadjustment for multiple testing, nonparametric tests of pairwise differences mostly rely on Bonferroni correction or a modified, less conservative method, the Bonferroni-Holm correction.

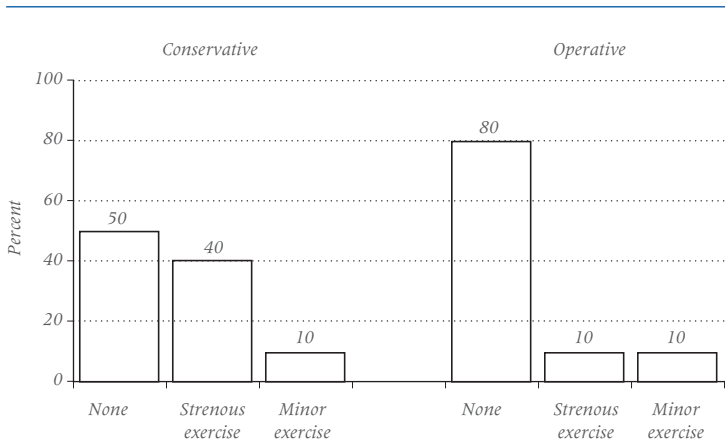


Fig 5-1 Pain and discomfort ratings of patients after conservative and operative treatment.



The average number of weeks until return to work was 11 weeks in the conservative group and 9 weeks in the surgical group. The difference of 2 weeks between the two regimens does not seem to be very impressive. Yet, given the standard deviation of 3, this difference is two thirds of the standard deviation.

Having observed this difference of 2 weeks of sick leave between the treatment groups, the investigator will usually be interested in whether this difference is “statistically significant”. To answer the question whether a difference between groups is statistically significant is equivalent to answering the question whether this difference can be explained merely by chance. For testing statistical significance, we (hypothetically) assume that both treatments are equally effective with respect to duration of sick leave and that the observed difference simply occurred by chance. This is the “null hypothesis”. Then, based on the value of a relevant test statistic, the probability of obtaining the observed difference, or one more extreme, under this assumption is computed. This probability is called the  $P$  value. If the  $P$  value is less than or equal to a prespecified probability level, the so-called significance level, the result is said to be “statistically significant”—the occurrence of the observed difference just by chance would be so unlikely that we reject the hypothesis that both treatments are in fact equally effective. The significance level most often used in statistical tests is 0.05 or 5%. This value is no natural constant, but simply a convention (see also chapter 2 “Errors and uncertainty”).

*A significance level of 5% is only a convention, but reasonable and accepted in the scientific community.*

Depending on the circumstances, smaller (eg, 0.01, 0.001) or larger (eg, 0.10) significance levels can be chosen. Even though the particular choice of the significance level is a bit arbitrary it is very important that it is defined in advance before the test is conducted and not post hoc after the  $P$  value has already been obtained.

As can be seen from Fig 5-1, the data for the primary endpoint are not completely normal. The distribution of the variable in the conservative group is more compressed than a normally distributed variable would be and the operative group has two cases with a very short duration of sick leave. Many statistical methods exist for assessing whether empirical data are consistent with a normal distribution. However, this is rarely needed—a pragmatic way is to graph your data first to gain an impression of the underlying distribution.

*Do descriptive and graphical analyses first before proceeding with statistical testing.*

Our data seem to be sufficiently normal so that a parametric test for statistical significance of the difference between the two groups is justified. According to Table 5-1, the appropriate statistical test for comparison of mean values in two groups is the t-test. Alternatively, if we don't trust that the data come from normal distributions, the nonparametric companion of the t-test, the nonparametric Mann-Whitney U test can be used. Results of both analysis methods are presented in Table 5-3.

	Treatment group		t-test (parametric)	U test (nonparametric)
	Conservative (N=30)	Operative (N=30)	P value	P value
Mean value	11.0	9.1	0.016	0.024
Standard deviation	3.0	3.0		

**Table 5-3** Statistical comparison of duration of sick leave in patients after conservative and operative treatment.

With the usual significance level of 0.05 both the parametric t-test and the nonparametric U test indicate that the difference is too large to be attributable to chance alone. We therefore reject the null hypothesis of equal outcomes in both treatment arms.

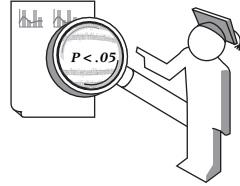
A respective manuscript summarizing the results of the study could read in the “Methods” and “Results” sections:

---

**Example**

*Methods: ... Differences between the two groups were tested by the t-test. The results were considered to be significant if  $P < .05$  ...*

*Results: ... The patients treated by surgery were on sick leave for an average of  $9 \pm 3$  weeks compared with  $11 \pm 3$  weeks in the patients treated conservatively ( $P = .016$ ) ...*




---

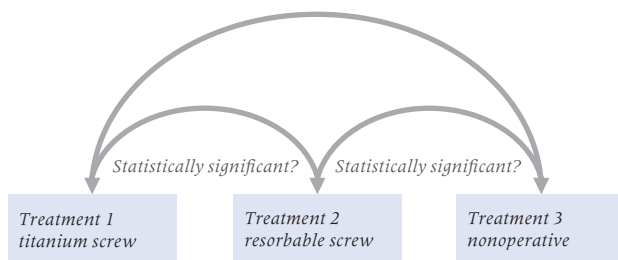
*We tacitly employed a so called “two-sided test”. This means that we expected that significant differences could occur in both directions, favoring either conservative or operative treatment. If a real difference can be assumed to occur only in one direction a “one-sided test” would have been the correct method. However, one-sided tests are rarely appropriate in medical research.*

*The standard t-test requires the standard deviations in both groups to be equal—an assumption which can also be statistically tested. If this assumption is violated an alternative t-test is available. Almost every statistical program will provide the user with the results of the standard and the alternative t-test.*

*The printed output of statistical test results often show a magic quantity: the “degrees of freedom”. This is simply a technical number that is a function of either the number of cases in the sample and the number of groups or, when contingency tables are analyzed, of the number of rows and columns in that contingency table.*

## 6 Comparison of more than two groups

Consider now a similar experiment in which two surgical treatments (say, a titanium Herbert screw and a new bioresorbable screw) have been included in the study protocol in addition to the conservative treatment arm. Now the comparison involves three instead of two groups. According to the overview in [Table 5-1](#), analysis of variance (ANOVA) and the associated F-test would be the relevant statistical analysis method. The null hypothesis to be tested with the F-test is that the mean values of all three groups are equal. If the statistical test gives a significant result this only tells us that this null hypothesis is not consistent with the data. Sometimes this will be exactly what we wanted to know. Yet, we would still not know which of the differences—between the titanium and resorbable screw, between the titanium screw and conservative treatment, or between the resorbable screw and conservative treatment—are unlikely to occur when in fact no differences exist ([Fig 5-3](#)).



**Fig 5-3** Pairwise comparisons of three groups.

Generally, if  $k$  groups are involved, a total number of  $\frac{(k-1) \times k}{2}$  pairwise comparisons are possible—in our example with three groups the number of comparisons is  $\frac{(3-1) \times 3}{2} = 3$ . To answer the question which of the differences are statistically significant it may be tempting just to conduct t-tests for all pairwise differences. Such

multiple t-tests are usually a bad choice. The reason is that multiple testing is associated with a “true” significance level that is larger than the nominal value of, say, 0.05. In this situation, a null hypothesis of no difference will be rejected even if the probability that the difference occurred by chance is larger than the prespecified significance level. Another choice, the so-called Bonferroni correction by dividing the nominal alpha level by the number of comparisons ( $0.05/3 = 0.017$ ) and rejecting the null hypothesis if the  $P$  value is less than or equal to the corrected significance level avoids the problems of the inflation of the significance level. However, this method is likely to overadjust the significance level and hence to miss existing differences. Statisticians call this property of the Bonferroni correction “conservative”.

*As a rule of thumb, the Bonferroni method is better than ignoring the implications of multiple testing.*

If Bonferroni-corrected results are statistically significant, the researcher is on the safe side because the difference is at least “as significant” as or even “more significant” than the nominal alpha level. More complicated situations may require the application of specific multiple-comparison methods which avoid overadjustment.

## **7 Analysis of paired data and other extensions**

Paired data arise when the observations are related in some natural way: an endpoint is measured before and after an intervention, presence of a certain disease is assessed in pairs of twins, cases and controls are matched according to relevant characteristics. It would be a mistake to use the methods described above for analyzing differences between groups with such paired data. Ignoring the process that generates the paired data (repeated measurements, matching) will almost always produce wrong results of statistical tests. Fortunately, for all examples described here, methods that account for paired data are available. When only two variables are paired, the paired t-test (continuous data), Wilcoxon signed-rank test (ordinal data), or McNemar test (binary data) can be used. With more than two related observations per case (measurement of outcome before and after

treatment and 6 months later, for example), repeated measurements ANOVA (continuous data), the Friedman test (ordinal data) and the Bowker test (categorical data) are appropriate.

Even though the methods for statistical data analysis presented in this chapter cover a wide range of approaches to answer specific questions of scientific interest, a researcher can be confronted with situations in which these methods are not sufficient. A common challenge in data analysis arises when the effects of more than one factor on outcome variables have to be taken into account simultaneously. Then, multivariable methods like linear or logistic regression are required. Application of multivariable methods is not necessarily much more difficult than using the methods included in this chapter, yet, they require—as can be expected with advanced techniques—a deeper understanding of the underlying statistical principles and at least some experience in their application. It is always a wise decision to seek advice from a statistical expert if you have any doubts about appropriateness of a particular statistical method.

## 8 Summary

- The results from statistical tests indicate whether an observation may have been produced simply by chance.
- Statistical tests compare the difference between expected and observed values.
- The choice of the appropriate test depends on the quality of data (binary, categorial, continuous), the underlying distribution (symmetric versus shewed), and the dependency of groups (paired versus unpaired).
- All statistical problems that go beyond need qualified assistance.

