



Statistical Methods

Mark Vollrath

Contents

- 12.1 Basic Questions: Distribution
Vs. Examination of Differences – 654**
- 12.2 Expression of Characteristics:
Confidence Intervals – 655**
 - 12.2.1 Methodology: Sampling – 655
 - 12.2.2 Statistics: Determination of Characteristic Value – 656
- 12.3 Differences Between Conditions:
Significance Tests – 659**
 - 12.3.1 Methodology: Experimental Designs – 659
 - 12.3.2 Statistics: Significance Tests – 663
 - 12.3.3 Statistics: Presentation of Results – 669
- 12.4 External and Internal Validity – 672**
 - References – 673**

12.1 Basic Questions: Distribution Vs. Examination of Differences

In the field of vehicle ergonomics, two main questions can be distinguished, which can be solved with the help of statistical methods:

- Determination of the expression of certain characteristics in the relevant population
- Examination of the difference between different conditions

Both questions have in common that on the one hand a certain methodical procedure is necessary for the answer (sampling procedure, design of experiments). On the other hand the appropriate statistical procedures have to be chosen (confidence intervals, significance tests, see [Fig. 12.1](#)). This will be explained first.

The first question is always important when certain customer characteristics have to be taken into account when designing control elements or displays. Controls in the vehicle shall be so placed that they can be reached by the driver's hand without the driver having to change his sitting position. Here the essential characteristic is the arm length. A display should be placed at a height that can be easily seen by any driver. In this case, the eye height above the seat is a relevant property.

The methodological approach in this case focuses on the selection of a representative sample. Usually you want to make a statement for a certain driver population, e.g. for the German driver. Since it is not possible to examine all persons in this population, a sample must be taken. Which concrete persons have to be examined in order for the results to

be representative for the German driver population?

On the one hand, the statistical methods for this question are concerned with how best to describe the distribution of the property. The mean value can be searched for in order to find the best case for most people. However, minimum or maximum values may also be meaningful to demonstrate that the solution chosen is also appropriate for people with the corresponding extreme characteristics. On the other hand, it is a question of the accuracy of the estimation of these parameters. This depends essentially on the size of the sample: The more persons examined, the more accurately the conditions in the population can be estimated.

The second question about differences is relevant when comparing variants or design alternatives and when examining to what extent certain conditions (e.g. a warning system) lead to changes compared to control conditions (e.g. a drive without a warning system) (e.g. a faster braking reaction of the driver). In each case, it is used to compare groups of people with each other, which can also be the same people at two different times (repeated measurement). In terms of the methodological approach, the focus here is on experimental design. How are the different groups "treated" so that a possibly found difference can actually be attributed to the interesting variation of influencing variables?

Another very important aspect here is the sample size. The smaller the difference, the more volunteers are needed to actually detect it in the test. Depending on the experimental

| | Question | |
|-------------|-------------------------------|--------------------------------|
| | Expression of characteristics | Differences between conditions |
| Methodology | Sampling procedure | Design of experiments |
| Statistics | Confidence intervals | Significance tests |

Fig. 12.1 Basic questions with the associated methodological and statistical aspects. For further explanation, see text

design and the quality of the data, different methods are used to prove the effect. The selection of the appropriate and sensitive procedure is the essential point of the statistics for these differential questions.

12.2 Expression of Characteristics: Confidence Intervals

12.2.1 Methodology: Sampling

The aim of sampling for this question is to obtain a sample that is as representative as possible, i.e. a sample that reflects the conditions in the population as well as possible. On the one hand, this concerns the procedure for drawing the sample and, on the other hand, the necessary number of persons.

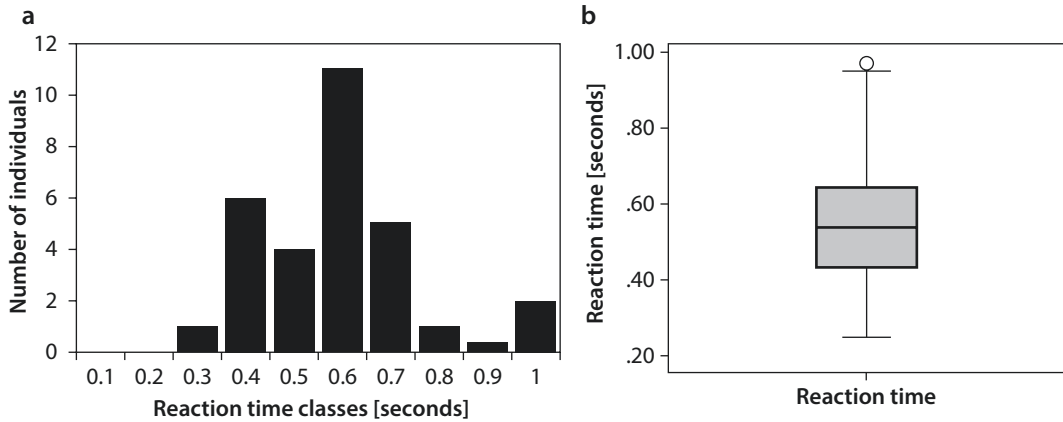
The best method of sampling is *random selection*. With this method, each person in the population has the same chance of being included in the investigation. Thus, all influencing variables on the characteristic to be measured are distributed in the sample in the same way as in the population. This is all the more the case the more people are drawn. For example, many characteristics depend on the sex of the person. If 1000 persons are randomly drawn from the German population, the gender ratio in such a sample will most probably correspond to that in the population as a whole. If, on the other hand, only a small sample of two persons is drawn, the probability is about 50% that only men or only women will be examined and the sought-after characteristic, e.g. the mean body size, incorrectly estimated.

However, the random selection of a sufficiently large sample from the population is rarely possible for practical reasons in the field of vehicle ergonomics. The examinations must be carried out e.g. with a certain vehicle at a certain place. It is possible that the results should be treated confidentially, so that only employees of the company can be considered as participants. And the cost of the examination is so high that only about 30 people can be examined. Against this background, the question arises as to how to arrive at the best samples under these circumstances.

It is important, especially for small samples, that as many characteristics as possible that may influence the relevant characteristic are taken into account in the sample selection. This is called a *stratified sample*. Central characteristics are certainly age and sex. One would include approximately the same number of men as women and persons of different age groups in the study in order to specifically consider the influence of these characteristics. Other relevant features in the field of vehicle ergonomics are body height and weight, as well as driving performance for many questions. In order to map the range well with respect to these characteristics, one would like to try to map the possible expressions of the characteristics and their combinations with at least 3-10 persons. However, this leads to very large samples for only a few characteristics. If one takes both sexes into account, three age groups, three classes of body size and three groups with different driving experience are formed, one would have $2 \times 3 \times 3 \times 3 = 54$ combinations. If you want to examine 10 persons per combination, you would need 540 persons. From a practical point of view, this access with the help of stratified samples, which take into account the combination of characteristic values, is usually only possible with the inclusion of 2-3 characteristics.

In addition, the inclusion of different characteristics makes the sample more heterogeneous. The relevant properties thus scatter more, which makes reliable estimation more difficult. From this point on, it may make sense to first work with a *homogeneous sample* in order to obtain a good estimate for at least this type of test person even with a relatively small number of test persons, and then to extend this to other, again homogeneous samples in further steps.

In summary, the significance of the survey with regard to the population as a whole depends to a large extent on the sampling. If relatively homogeneous, locally limited samples are examined, the extent to which these results can be transferred to other groups of the population must be considered when interpreting the results. If statistical methods are applied to these results, they may be able to estimate relatively accurately the "true" val-



■ **Fig. 12.2** Distribution of reaction times for a lane change task. Shown in **a** as histogram the number of persons in the classes of reaction times displayed on the x-axis. In **b**, the same data is displayed as a box plot

ues in the population. However, this estimate only applies to the part of the population corresponding to the sample. Therefore, in order to assess the relevant results, it is important not only to have information on the sample size and the resulting estimated distributions, but also on the type of sampling and the main characteristics of the sample. This is the only way to assess the extent to which the results are not only accurate but also representative.

12

12.2.2 Statistics: Determination of Characteristic Value

After the data collection, it is advisable to first present the collected data descriptively. This can be done for example as frequency distribution or histogram (see ■ Fig. 12.2a). Here, meaningful categories of the characteristic are created and the number of values per category is displayed. One recognizes thereby very well the kind of the distribution (e.g. symmetrical or oblique, single or multiple peaks) and receives a first impression of the measured orders of magnitude. Also “outliers” can be recognized quite well (see, for example, the two large reaction times in ■ Fig. 12.2, left).

A more compressed way of displaying the data is the boxplot (see ■ Fig. 12.2b). The grey box contains the mean 50% of the measured values. The black horizontal line represents the median (see below). With the vertical

strokes the 1.5-fold of the range of the box is applied upwards and downwards, but only up to the last available value. All values outside this range are drawn as individual points and are therefore easy to identify as outliers. This type of display is also very suitable for comparing several conditions with the help of box plots arranged next to each other.

The next step is to try to describe the relevant properties of these distributions using characteristic values. As ■ Table 12.1 shows, this describes the typical values on the one hand, and the width or dispersion of the distribution on the other. As the column on the right shows, the interpretation of the individual measurements is slightly different. In addition to this different information content, these measures are also best suited for different types of data.

Four scale levels are distinguished according to the information content of the numbers used (see also ► Sect. 11.3.1.3). For calculations, the gender is often classified into the values “1: male” and “2: female”. For these two numbers, it is only useful to interpret the equality of the measured values. The fact that the 2 is twice as large as the 1 is correct for the numbers, but not for the categories for which the numbers stand. Such a classification is called a *nominal scale*. A useful parameter is the mode (or modal value). A mode of 2 means in the example that women are more frequently included in the sample than men. The range can also be useful

Table 12.1 Overview of essential characteristic values of distributions

| Characteristic value | Computation | Interpretation |
|------------------------|---|---|
| Mean | $\bar{M} = \frac{\sum_{i=1}^n x_i}{n}$ <p>x: measured values n: number that measured values</p> | Typical value of the sample. Sum of deviations from this value is minimal |
| Median | The value above which 50% of the measured values lie (interpolation for categorical values) | Typical value of the sample. 50% of the values are below / above |
| Mode | Most frequently measured value | Typical value of the sample. The value that occurs most frequently |
| Standard deviation | $S = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{M})^2}{n - 1}}$ <p>X: measured values M: mean N: number of values</p> | Deviation of the values Mean deviation from the mean value |
| Interquartile distance | Range between the value below which 75% of the values lie and that below which 25% of the values are | Deviation of the values. Width of the range in which the middle 50% of the values lie |
| Range | Maximum – minimum | Distribution of the values. Width of the range in which the values lie |

here to describe the number of categories used. Finally, it is possible to indicate the percentage of occurrence of the different categories (“The sample contains 45% men”).

The second scale level is the *ordinal scale*. Here the order of the numbers can be interpreted. If a font size is judged as “1: small” and “2: large” by test persons for example, the statement that judgement 2 is larger than judgement 1 is significant – a font size judged as “large” is larger than one judged as “small”. Again, not significant is the statement that this large font size is twice as large as the small one, since 2 is twice as large as 1. For data at this level, the median is a meaningful description of the typical value and interquartile distance for variance.

The *interval scale* is the third scale level. The distances between the measured values can also be compared here. If you judge the volume of a warning tone with “1: very quiet”,

“2: quiet”, “3: medium”, “4: loud” and “5: very loud”, you can first interpret the difference as well as the order of the numbers. In addition, it makes sense to say that the difference between 3 (medium) and 1 (very quiet) is greater than the difference between 2 (quiet) and 1 (very quiet). Whether on the other hand 2 (quiet) is twice as loud as 1 (very quiet) is doubtful. Here, too, this relation of the numbers must not be interpreted. Meaningful characteristic values for the interval level are the mean value and the standard deviation.

As the name suggests, you can use the *ratio scale* and interpret the numbers as ratios. This is often the case with physical data. A measurement of reaction times is an example of data at the ratio level. A reaction time of 500 ms is twice as long as one of 250 ms. Also for this scale level, mean and standard deviation are a good description of the typical value and dispersion.

Overall, at a certain scale level only certain interpretations of the numbers and thus only certain characteristic values are meaningful. At the higher scale level, the characteristic values of the lower levels can also be used and interpreted and sometimes provide interesting additional information.

In addition to describing the data of the sample by specific characteristic values, it is often a matter of estimating the conditions in the population with the help of the sample. Especially for the mean value, but also for the percentage share of certain categories, the question arises as to how precisely the calculation of the sample characteristics reflects the conditions in the population. This is answered with the help of confidence intervals. A confidence interval indicates the range in which 95% (sometimes also 99%) of the population characteristics that may have generated the sample characteristic lie. Or: With a probability of 95% (or 99%), the true value in the population lies in this range. From a methodological point of view, this of course only applies if the sample is a representative selection of the population. Confidence intervals only indicate how accurate the estimate based on the sample is, but not how well the sample reflects the population.

The formula of the confidence interval for a mean is as follows:

$$\Delta_{crit} = M \pm z_{\alpha/2} \cdot \widehat{\sigma}_M \tag{12.1}$$


The following is included

$$\widehat{\sigma}_M = \sqrt{\frac{\sum_1^n (x - M)^2}{n \cdot (n - 1)}} \tag{12.2}$$

\widehat{M} in the formulas stands for the mean value. $\widehat{\sigma}_M$ is the standard error of the mean value. The z value stands for the corresponding value of a standard normal distribution that includes the mean 95% (99%) of this distribution. Formally, this corresponds to the z-value with $\alpha = 5\%/2$ or $\alpha = 1\%/2$. The following numerical values can be found in the corresponding tables:

95%: $z_{\alpha/2} = 1.96$

99%: $z_{\alpha/2} = 2.58$

For the data from  Fig. 12.2, the mean value is $M = 0.53$ seconds with $n = 30$ persons and a standard error of the mean value of 0.03. This results in $\Delta_{crit} = 0.53 \pm (1.96 \cdot 0.03) = 0.53 \pm 0.06$. The 95% confidence interval therefore ranges from 0.47 to 0.59 seconds. Thus, a satisfactorily accurate estimation of the mean population reaction time is already achieved with 30 test persons.

Eq. 12.2 also makes it immediately clear which role the sample size n plays. The larger the sample, the smaller the standard error, which in turn directly determines the width of the confidence interval. This can be used to calculate the required sample size for a given accuracy if the mean value and standard deviation are known e.g. from a pilot study. The formula is as follows:

$$N_{necessary} = \frac{1.96^2 \cdot sd^2}{Accuracy_{requested}^2} \tag{12.3}$$

In the above example, the standard deviation was $sd = 0.17$. If the mean value of the population is to be estimated with an accuracy of ± 0.1 second, then according to the formula the result is a $N_{necessary} = 11$.

You can also calculate confidence intervals for percentage values. The basic formula is comparable:

$$\Delta_{crit(\%)} = P \pm z_{\alpha/2} \cdot \widehat{\sigma}_{\%} \tag{12.4}$$

Where P is the empirically calculated percentage. The standard error results as:

$$\widehat{\sigma}_{\%} = \sqrt{\frac{P \cdot (100 - P)}{n}} \tag{12.5}$$

In the above example, the reaction time of 11 of the 30 = 37% of the subjects was in the category between 0.5 and 0.6 seconds. What is the confidence interval of this percentage? The standard error is:

$$\widehat{\sigma}_{\%} = \sqrt{\frac{37 \cdot 63}{30}} = \sqrt{77.7} = 8.8$$

This calculates

$$\Delta_{\text{crit}(\%)} = 37\% \pm (1.96 \cdot 8.8) = 37\% \pm 17$$

The 95% confidence interval thus ranges from 20% to 54%. Here, too, it is possible to specify by conversion which sample would be necessary to achieve a certain accuracy.

$$N_{\text{necessary}} = \frac{1.96^2 \cdot (P \cdot (100 - P))}{\text{Accuracy}_{\text{requested}}^2}$$

So if you want to estimate the percentage with an accuracy of $\pm 5\%$ the formula above calculates $N_{\text{requested}} = 358$.

12.3 Differences Between Conditions: Significance Tests

12.3.1 Methodology: Experimental Designs

The second type of question compares at least two conditions. The general question is to what extent certain influencing factors systematically change the measured values. The scientific concern here is the search for causal laws, i.e. for cause-effect relationships. To make this clear, a distinction is made between independent variables (IV, causes) and dependent variables (DV, measured values, see also ► Sect. 11.1.3). The relationship is represented schematically as follows:

$$\text{IV} \rightarrow \text{DV} \quad \text{or} \quad \text{DV} = f(\text{IV})$$

The IV systematically causes certain changes in the DV, the measured values. The measured values are therefore a function of the independent variables. In ergonomics, it can be demonstrated for example that a certain display variant in the head-up display with a warning tone leads to faster reaction times than the conventional display with a warning tone in

the combined display. Here IV is the type of display in two steps (HUD vs. conventional display), DV is the reaction time. It is assumed that the mean value of the reaction times of a group of drivers with the HUD is smaller than in a group with a conventional display.

Even more complex questions can be represented in this scheme. It can be assumed that, in addition to the location of the display, the presence of a warning tone is also essential for the effect of the warning. To check this, the second IV would be to introduce the warning tone in the steps “without” and “with”. In order to investigate the effect of both IVs alone and in combination, both IVs would now have to be combined, resulting in four experimental groups. Independent variables are also referred to as “factors” in order to distinguish experimental designs according to the number of IVs studied. This leads to the description as “single factorial”, “two factorial”, etc. experimental design (see ► Table 12.2).

The number of levels of the factors must be distinguished from the number of factors. An influencing factor is often examined in two levels (e.g. without vs. with). But also the comparison of several levels (warning in the conventional display, in the HUD, in the centre console) is not unusual. For each design, therefore, the number of levels shall be given in addition to the number of factors. This is often done in the form “a × b × c × ... factorial experimental design”, where a, b and c are the levels of the respective IV. The above example with the two factors “location of the display” and “warning sound” could be described as 2 × 2 factorial experimental design.

The experimental designs also differ on the DV side. Here it is very important how many DVs are examined. If reaction time is the only measured DV, it is a univariate plan. Mostly, however, several DVs are recorded, e.g. also subjective evaluations. If only one global judgement (“How good was the display?”) is recorded as the only additional DV, it is a bivariate plan. In general, multiple DVs are called multivariate plans.

The scale level of the measurements is also important for further evaluations (see

■ **Table 12.2** Overview of essential aspects of experimental designs. For further explanation, see text

| Criterion | Significance | Description of the |
|------------------------|--|---|
| Number of IV | How many influencing factors are investigated? | Single factorial Two-factor ... Multifactorial |
| Number of levels of IV | Which aspects of the influencing factors are investigated? | 2 levels 3 levels ... |
| Number of DV | How many parameter are measured? | Univariate Bivariate ... Multivariate |
| Scale level of DV | What is the meaning of the measured numbers? | Nominal scale Ordinal scale Interval scale Ratio scale |
| Measurement repetition | Does each test person receive only one or more levels of IV? | Independent Mixed Dependent |

► Sect. 12.2.2). As described above, the question arises as to which of the information contained in the figures can also be interpreted. This leads to the different types of characteristic values described above, but also has further evaluations in the statistical comparisons (see ► Sect. 12.3.2).

The last criterion is the distinction between between-subjects and within-subjects designs. In the case of between-subjects designs, each subject receives only exactly one level or combination of the independent variables. In the case of within-subject designs (also called repeated measurements), each subject provides measured values in all conditions. With multi-factor plans, it is also possible to examine individual IVs with repeated measurements, while others can be examined with independent groups. The decision to do so shall be taken on the basis of substantive considerations. It can be assumed that experienc-

ing one condition influences the reactions in another condition, independent plans should be chosen. So if the different warnings are to be investigated with the help of an unexpectedly occurring critical situation in the driving simulator, a dependent plan makes little sense, since the subject already knows this critical situation after the first condition and this would no longer be unexpected for him in the second condition. Whenever learning effects are assumed, repeated measurements should be avoided. The same applies if the tests are very tiring for the test person, so that a decrease in motivation and performance is to be feared. This could also distort the results, making independent plans more useful.

Then why repeat measurements at all? The advantage of within-subject designs is that each subject can be compared with him/herself. Especially when people react quite differently, such a change can always be expressed in terms of the individual typical value. Small effects can thus be discovered with just a few test persons, since the differences between persons that are not of interest are eliminated in this way. Plans with repeated measurements are therefore advantageous both in terms of the number of test persons required and their sensitivity to discover effects.

However, in these plans possible time effects have to be controlled. Since both learning and fatigue can in principle never be completely ruled out, it is important to ensure that this works evenly in the various conditions. This control of time effects is achieved by varying the order of conditions for each subject. There are essentially two possibilities: The complete permutation of all possible sequences and the technique of the Latin square. ■ Figure 12.3 provides an overview of this.

Under two conditions, i.e. only one IV, two groups of subjects are formed, randomly assigned to the two groups to ensure that the groups are comparable. The first group receives treatment A at first, then B. In the second group, the order is reversed.

Under three conditions, a multiple of 6 test persons is required. Each of these 6 subjects receives a different sequence of conditions. The smallest meaningful number for

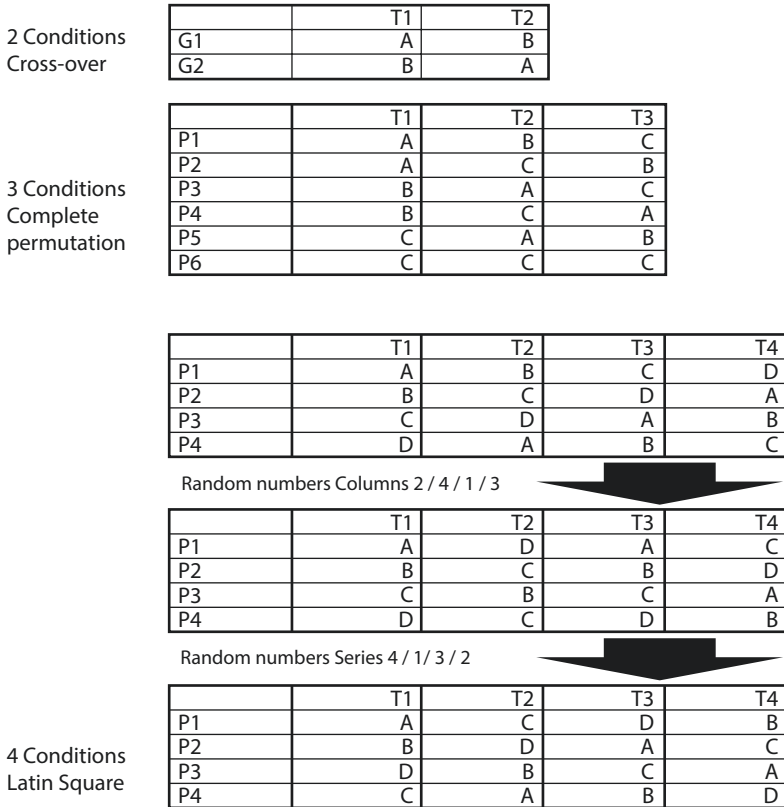


Fig. 12.3 Overview of techniques of control of time. The respective test conditions (A..D) for subjects (P) or groups (G) are shown. The different times are designated as T1..T4

this case are 12 test persons with whom one can already prove effects by the dependent plan.

Combining several IVs with each other quickly creates more conditions. For a 2×2 plan with complete repeated measurement, four conditions are present. If a third factor with also 2 levels is introduced, the number increases to 8. A 3×3 plan contains 9 conditions and so on. Here it is usually no longer possible to present all conceivable sequences in a complete permutation. An alternative here is the technique of the Latin square, in which each subject receives all conditions and the sequences are chosen in such a way that each condition is equally frequent at all times across the various subjects. Not all sequence effects can be controlled, but the simple time effects. This technique also requires multiples of the number of conditions. With four conditions (as shown in Fig. 12.3), at least 4 subjects are required. The base square for this

case is shown in the middle of the figure. In order to introduce a certain randomness here, a random sequence of the four points in time is generated, in the example 2/4/1/3. The columns in the second square are then sorted in this order. The original column 2 becomes the new column 1 and so on. The same is done with the lines where the order 4/1/3/2 was drawn here. So the new line 1 is the old line 4 and so on. This procedure is repeated for all groups of four subjects to be examined. Again, 12 subjects are the lower limit of what appears to make sense in this dependent plan. For 5 or more conditions, additional squares can be created accordingly.

The advantages of between-subjects designs and designs with repeated measurements are shown in Table 12.3. Independent plans are insensitive to learning and fatigue effects. Due to the greater differences between subjects, effects cannot be discovered as easily as in within-subjects designs. On the other

hand, these effects are more robust, can be better replicated and are more significant. With repeated measurements, considerably fewer test persons are required and even small effects can be discovered. However, time effects may influence the effects of UV and the effects may be limited to the specific sample and thus poorly transferable to other individuals. Depending on the issue at hand and the practical framework conditions, the plan to be chosen must be weighed up accordingly.

In order to make the procedure and the experimental design easy to understand for the readers of the corresponding reports, a schematic representation is recommended. Figure 12.4 shows an example of a three-factor plan. As the first IV1, the driver age is examined in two levels, distinguishing young and old drivers. Further there is as IV2 the warning tone with the conditions “without” and “with”. Finally, as the third IV3, the location of the warning is investigated in three lev-

els, with each subject experiencing all three locations (repeated measurement). In addition to the IV, the number of test persons can be seen in the cells. The numbering makes it clear at which point a repeated measurement was introduced and where independent groups are examined. Since the IV3 is examined with repeated measurements, a complete permutation with 12 test persons was used here.

The presentation of the experimental design is also so important because the questions of the experiment can be derived directly from it. As shown above, the aim is to examine the extent to which the IVs lead to a systematic change in the DVs. If one compares a IV with two levels (without and with warning tone), the question arises whether the characteristic values of the corresponding two groups differ (see Table 12.4). At three levels of IV, one can examine whether IV leads to differences between the three groups at all. Further one is interested in which of the groups differ from each other.

It gets more complex with two or more IV. With two IVs, one is interested on the one hand in the effect of each individual factor, and on the other hand in the interaction of the factors. Does the warning sound have a different effect when combined with the HUD than with the conventional display? Such an effect is called interaction. The same applies to three-factor plans, where the interaction between all three factors is added to the individual effects and two-fold interactions. This increasing complexity leads to the fact that already the results of four factor plans are dif-

Table 12.3 Advantages of independent plans and plans with repeat measurements

| Between subjects design | Within-subjects design (repeated measurements) |
|--|--|
| Robust to learning and fatigue effects | Only few test persons are needed |
| When an effect is discovered, it is more replicable and more significant | Even small effects can be detected |

| | | IV1 “Driver’s age” | | | |
|---------------------------|----------------------|---------------------|-----------|-------------|-----------|
| | | Young drivers | | Old drivers | |
| | | IV2 “warning sound” | | | |
| | | | | | |
| IV3 “Location of warning” | HUD | VP 1..12 | VP 13..24 | VP 25..36 | VP 37..48 |
| | Conventional display | VP 1..12 | VP 13..24 | VP 25..36 | VP 37..48 |
| | Centre console | VP 1..12 | VP 13..24 | VP 25..36 | VP 37..48 |

Fig. 12.4 Example of a three-factor experimental design with repeated measurements on the “Warning location” factor (VP stands for subject)

Table 12.4 Experimental designs and related questions. For further explanation, see text

| Experimental design | Questions |
|-------------------------|--|
| 1 IV, 2 levels | Do the two groups differ? |
| 1 IV, 3 and more levels | Does the IV work? Which groups are different? |
| 2 IV | Does the IV 1 work? Does the IV 2 work? Does IV 1 have a different effect depending on the characteristics of IV 2 and vice versa? (interaction) |
| 3 IV | Are IV 1, IV 2 and IV 3 working? Is there an interaction between IV 1 and IV 2, IV 1 and IV 3 or IV 2 and IV 3? Is there an interaction between the three IVs? |

difficult to interpret in practice and from there one can only recommend to concentrate on the most relevant three IV per investigation and in case of doubt to carry out several investigations. These problems of interpretation are described in more detail in ► Sect. 12.3.2. Before doing so, however, it is important to present the statistical validation of the results under the keyword “significance tests”.

12.3.2 Statistics: Significance Tests

Why is an examination with several subjects necessary for the investigation of differences between different conditions? The reason lies in the diversity of persons described above. Not everyone reacts with the same speed, so that a group of people always get a distribution of the measured values, although they are examined under the same conditions. However, it also follows from this that differences will always occur when comparing two groups, even if the groups are treated equally. If the groups are treated differently, the question arises as to whether the differences found can be explained by the random error or

whether they arise systematically as an effect of the influencing factors investigated. This question examines the statistics with the help of significance tests. Essentially, these questions answer the following question:

- Question of statistics: Are the differences found between the various conditions so great that it can be assumed that the influencing factors investigated have an effect?

To answer this question, it is formulated somewhat differently:

- How probable are the differences found under the assumption that this is only due to the (random) differences between persons, but not the factors examined?

The advantage of this formulation is that the assumption it describes can be converted into a statistical model. If you assume that only random differences are the cause, you can create a distribution of possible differences (“What results would you find if you repeated the investigation 100 times?”). If, for example 10 test persons were examined in two groups and their reaction time measured, these 20 measured values can be randomly assigned to two groups of 10 measured values each and the difference between the mean values of the two groups calculated. If this is repeated frequently, a distribution of possible differences is obtained on the assumption that only random differences (in this case distributing the 20 values randomly into two groups) had an effect. Such a distribution is shown in ► Fig. 12.5. It can be seen that under random conditions the same results are relatively frequent in both groups (reaction time difference = 0), while large deviations in positive and negative direction are less frequent.

With the help of this random distribution, the above question can now be answered by indicating how likely it is that the difference actually found in the experiment will occur under random conditions. One takes the probability of this and more extreme differences, which corresponds to the area under the curve. This is shown in the ► Fig. 12.5 filled out accordingly. In order to decide whether this result indicates that the influencing vari-

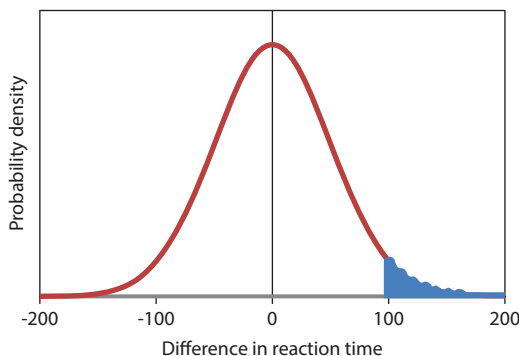


Fig. 12.5 Example of a random distribution of reaction time differences. The area in which the actually found reaction time difference and more extreme differences lie is shown in full

able actually acted, a decision rule is introduced:

- If the difference found is extremely unlikely under random conditions, then it is assumed that the influencing factor (IV) has worked, i.e. the difference is systematic and not random.

To decide whether something is improbable, a so-called significance level α is defined:

- “Unlikely” or “significant” usually corresponds to a significance level of $\alpha = 5\%$.

You can also find the convention to use a $\alpha = 1\%$. This is often referred to as “highly significant” results. The various significance tests (see below) now indicate the probability of the found result under random conditions, usually as a relative frequency of e.g. $p = 0.023$. Since this p is smaller than the significance level α (which is usually given in percent), it is decided that the result cannot be explained well by random differences, i.e. the influencing factor has worked or “the result is significant”. In summary, a significant result means that this is very difficult to explain by random differences. A proof of the effect in the very strict sense is of course not possible, because always a certain amount of uncertainty remains – even under random conditions this result could have occurred, albeit very rarely.

Statistically, the assumption that only random differences had an effect is considered to

be a null hypothesis (“ H_0 ”). The alternative hypothesis (“ H_1 ”) assumes that there is an effect. A distinction is made between a specific and an unspecific alternative hypothesis. The specific alternative hypothesis indicates the direction of the effect, e.g. that the reaction times become shorter with a new warning system. With the unspecific alternative hypothesis, on the other hand, one suspects a difference, which can, however, be in both directions. Herefor instance is formulated: “The reaction times in the experimental group are different than those of the control group”. The significance test can also be understood as a decision about these hypotheses. If the result is significant, the null hypothesis can be rejected. This indicates the presence of an effect. If no significant result is found, the null hypothesis must be maintained. So you couldn’t see any effect. It is important to note that this does not automatically mean that there is no effect. This is related to two types of errors shown in **Fig. 12.6** and explained below.

The first type of error is the alpha error. If one finds a significant result in an experiment, although the null hypothesis actually applies, i.e. in reality (i.e. in the population) this difference is not present, one makes an erroneous decision: On the basis of the result of the significance test one concludes that there is an effect, which is not true. This is shown graphically in **Fig. 12.7**. If the null hypothesis is correct, the possible results of studies are distributed according to the red curve. If one finds in a study a reaction time difference of e.g. +150 ms that is significant according to the statistical test, one decides to reject the null hypothesis, although in reality it is true in the population. One therefore concludes that an effect exists even though none exists.

To protect yourself from this alpha error, you can choose a lower significance level, e.g. 1% instead of 5%. This reduces the probability of wrongly choosing an effect. Another possibility is to replicate the effect under conditions as similar as possible. If a significant result is also found in the repetition, the probability of a wrong decision is significantly lower overall. If, for example, two studies are carried out and a significant result is found at

| | Reality/total population | |
|-------------------|---|--|
| | No difference Null hypothesis is correct | Effect existing Alternative hypothesis is correct |
| Significance test | | |
| Significant | Alpha error | Correct decision |
| Non-significant | Correct decision | Beta error |

Fig. 12.6 Correct and incorrect decisions in the significance test. For further explanation, see text

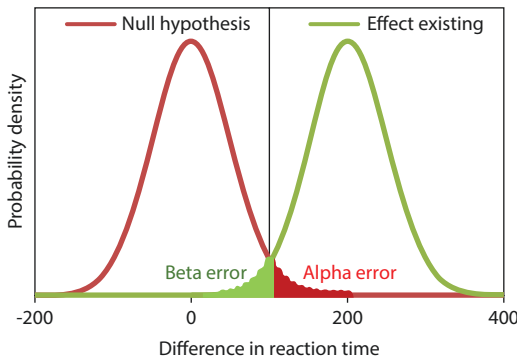


Fig. 12.7 Alpha and beta errors for the example of reaction times. The red curve shows the distribution of the possible results (difference between the mean values of the two investigated groups, “reaction time difference”) under the assumption of the null hypothesis, the green curve the distribution of possible results in the “reality”, i.e. the population in which an effect is present

$\alpha = 5\%$, the probability is that both studies will become significant, even though in reality there is no difference is $p = 0.05 \cdot 0.05 = 0.0025$, i.e. only 0.25%. In this way one can minimize the probability of an erroneous interpretation in the sense of claiming that there is a difference, although this is not the case in the population. If, for example, you want to introduce a new warning system in the vehicle, which is, however, associated with considerable costs, you can be sure in this way that there is actually a benefit for the driver.

However, this minimization of the alpha error should be seen in the light of the fact that it is associated with an increase in a second type of error, the so-called beta error. This is illustrated in Fig. 12.7 with the help of the green curve. This shows the distribution of test results in the event that an effect is

actually present in the population, which can be described in the example as an extension of the reaction times around 200 ms. Since only a part of the population is examined in each study, it is possible that a random sample is examined in which this effect does not show up, so that the reaction time difference e.g. is only +50 ms. According to the decision criterion shown in the figure, the null hypothesis cannot then be rejected, since this result is not quite probable even if there is no effect. This erroneous decision is referred to as a beta error. Here it is concluded on the basis of the test result that there is no effect, although in reality there would actually be a difference in the population. The more you try to minimize the alpha error, the bigger the beta error will be, the less you will discover an effect even though it is present.

Therefore, if you want to be sure not to overlook any effect, it makes sense to select the significance level alpha relatively large. This is particularly the case if one wants to prove that two variants are equivalent. If a new warning concept for example has been developed, which is associated with significantly lower costs than an older, relatively expensive solution, the equivalence of the variants is to be demonstrated in the experiment. In this case, the interest is to confirm the null hypothesis. Here it is important not to overlook it if there is a difference between the variants. Usually, the significance level is then set at 25%. In addition, there are also special types of significance tests, the so-called equivalence tests, which can be used to prove equality. A detailed description can be found at Wellek (Wellek 2010).

Statistical testing for difference thus assumes that distributions of possible differences are present under random conditions. One could now create the corresponding random distributions for each experimental design with the parameters measured there in order to achieve this statistical estimation. As this would be relatively time-consuming, either the measured values are transformed into ranks or categories and corresponding tables are drawn up for the various experimental designs. In each case, it must be taken into account how many independent variables were examined in which steps and how many test persons were involved. Or one transforms the measured values in such a way that they correspond to certain statistical distributions, which are then used to read the probabilities in a comparable way as in the example in [Fig. 12.5](#). In the latter case, one speaks of distribution-based methods, while the first are called distribution-free or non-parametric methods. “Distribution-free” means that no theoretical distribution is referred to. This designation is therefore not completely correct, since an empirically provided distribution (e.g. of ranks) is used. From this point of view, the term non-parametric procedure is preferable, since “parameters” refers to the essential parameters of the theoretical distribution (e.g. mean value and standard deviation for a normal distribution). Such parameters are not required for non-parametric procedures.

Data quality shall be taken into account when deciding on the procedure to be used. Distribution-based methods are only really meaningful from the interval level, since differences are calculated in the calculation of the characteristic values that are only significant at this scale level. Furthermore, a certain sample size is required (e.g. more than 30), since the values are distributed sufficiently similar to these theoretical distributions only for larger samples. Finally, the question always arises as to whether the measured values are actually distributed sufficiently similarly to the theoretical distribution. It can be partly checked whether the empirical values correspond to certain prerequisites. However, these tests are often relatively sensitive and indicate

deviations from assumptions which, however, practically do not lead to any substantial change in the statistical evaluation. The non-parametric methods are also applicable at the lower scale levels, but not quite as sensitive to detect significant effects. Non-parametric evaluations are, however, difficult, especially in the case of multifactorial plans, since the assessment of the interaction of several factors is often made by adding effect estimates, which in turn does not appear to be useful at the ordinal level. Therefore, parametric methods are also frequently used in practice, although the prerequisites are doubtful.

In the presentation of the results of the statistical tests, the corresponding test quantities into which the empirical characteristic values have been converted are given on the one hand in order to be able to reproduce the calculation. The value of the test variable itself also includes an indication of the general conditions, which essentially corresponds to the number of test persons examined or the measured values used. This is hidden (in a slightly transformed form) in the so-called degrees of freedom (df). On the other hand, the result of the statistical test is reported as a p-value (as shown above).

Special programs such as SPSS, R (as open source variant) or toolboxes in Matlab are used for the calculation. A detailed description of the individual tests would go beyond the scope of this chapter. A detailed description can be found in corresponding textbooks of statistics, e.g. at Bortz und Schuster (2010) or Sedlmeier und Renkewitz (2008). An overview of the most important procedures can be found in [Table 12.5](#). The parameters listed in the right-hand column can be found in the corresponding editions of the statistics programmes. When selecting the individual tests, it is important to consider whether a test plan with repeated measurements is available or whether independent groups have been compared. In the first case, the measured values must also be arranged in such a way that measured values for the various conditions (columns of the data matrix) are available for each test person (row of the data matrix). In the case of independent groups, group membership is coded

Table 12.5 Overview of important statistical methods and the presentation of test results

| Experimental design | Test | Scale level | Repeat measurement? | Result formulation |
|-------------------------|---------------------------------------|-------------------------------------|---------------------|---|
| 1 IV, 2 levels | Independent sample t-test | Interval level | Without | $t(18) = 2.3, p = 0.017$ |
| | Dependent sample t-test | Interval level | With | $t(9) = 3.5, p = 0.003$ |
| | U-test | Ordinal level Distribution-free | Without | $U(9) = 14, p = 0.009$ |
| | Wilcoxon test | Interval level Distribution-free | With | $T(10) = 10, p = 0.042$ |
| | Sign test | Ordinal level Distribution-free | With | $x(10) = 1, p = 0.011$ |
| | Chi-square test | Nominal level | Without | $X^2 = 5.05, p = 0.025$ |
| | McNemar test (2 Categories) | Nominal level | With | $X^2 = 4.5, p = 0.033$ |
| 1 IV, 3 and more levels | Single factorial analysis of variance | Interval level | Without | $F(2,27) = 4.7, p = 0.03$ |
| | H test | Ordinal level | Without | $H(1) = 7.06, p = 0.007$ |
| | Friedman Rank Variance Analysis | Ordinal level | With | $X^2_{r(2)} = 6.1, p = 0.047$ |
| | Chi-square test | Nominal level | Without | $X^2(3) = 12.5, p = 0.005$ |
| | Q test by Cochran (2 Categories) | Nominal level | With | $Q(3) = 8.3, p = 0.003$ |
| 2 IV | Analysis of variance (ANOVA) | Interval level | Without | HW IV 1: $F(1,30) = 4.7, p = 0.038$ HW IV 2: $F(2,30) = 5.2, p = 0.011$ WW $F(2,30) = 4.4, p = 0.021$ |

using a separate variable. The scale level of the measurements must also be observed. For smaller samples (per group $n \leq 10$), non-parametric testing should be used if possible, since violations of the requirements of distribution-based testing are very important for small samples. In order to assess the test variables, it is necessary to specify the degrees of freedom, since the significance (expression of the p-values) depends on them. For the degrees of freedom, either the number of subjects is relevant (marked “n” in the table) and/or the number of levels of the independent variable (marked “k” and “l” in the

table). The degrees of freedom can also be found in the corresponding editions of the statistical programs.

The interpretation of statistically significant effects is simple when comparing two groups – these two groups differ. If several levels of a IV are investigated, the test parameter indicates whether at least two of the investigated groups differ. You can then either decide graphically where the differences lie, or make corresponding comparisons in pairs to decide this statistically. These pairwise comparisons are usually performed automatically by the statistics programs.

Another interesting aspect is the interpretation of the results of the two- and multi-factor variance analyses. For the main effects, i.e. the effects of the individual factors examined, it is indicated in each case whether the corresponding factor leads to significant differences irrespective of the characteristics of the other factors. Furthermore, the interactions between two factors each and, depending on the experimental design, between three and more factors are examined. If there are interactions, the main effects can sometimes no longer be interpreted depending on the direction of the effects. This will be shown at following chapter.

A central property of the statistical tests can finally be described at the distribution in **Fig. 12.5**. The less the difference values scatter, the narrower this distribution is and the more likely it will be to decide that a particular difference is very unlikely under random conditions. The consequence of this is that smaller effects can already be discovered with larger samples. The larger the sample, the more reliably and accurately the true value of the group is estimated (see above). This means that the difference between the values of the two groups is also less prone to error, and thus spreads less in the distribution. In parametric tests, this is also taken into account by the fact that when calculating a corresponding test variable, a measure of the difference between the groups (e.g. the difference between the group averages) is usually placed in relation to the measurement errors (e.g. the pooled variance within the groups). This becomes particularly clear with the test variable of the variance analysis, where the F-value (test variable) is a fraction of primary variance and error variance. Primary variance describes the difference in the measured values resulting from the different independent variables, error variance the random differences between the test persons.

In addition to enlarging the sample in order to capture the differences with a lower measurement error, this provides a second possibility for experiment planning in order to better recognize significant effects. For this purpose, either homogeneous groups of test

persons are used in order to minimize the error variance, or each test person is compared to him/herself under different conditions (dependent test plans or test plans with repeated measurements). From a statistical point of view, this also explains why experimental designs with repeated measurements can very well detect even small effects, even if only a relatively small number of test persons are examined.

The consideration of the appropriate sample size is also related to the statistical validation of effects. The term “power” is used to describe how well a statistical test is suitable for statistically proving an existing effect. The power of a test is calculated as $\text{power} = 1 - \beta$ where β is the beta error described above. The greater the probability of incorrectly rejecting an effect, the smaller the power, i.e. the ability of a test to detect an actually existing effect. As an investigator, one is correspondingly interested in using a test that is as powerful as possible. For example, parametric tests are usually more powerful than distribution-free methods. However, the main determinant of power is the sample size. The more people examined, the more likely it is that smaller effects can be statistically proven. If you know how the measured values with which you decide on the effect are distributed and how large the examined effect is in reality, then you can estimate how many test persons are needed before the examination to be able to statistically prove this effect. As a rule, however, the effect size is not known before the start of a study (otherwise the study would not be needed either). One can then fall back on conventions and, with the help of corresponding programs, calculate estimated values for suitable sample sizes for small, medium and large effects, e.g. for the test corresponding to the experimental design. A program very frequently used in this context is the freely available G*Power (Faul et al. 2009). An estimation of a meaningful sample size before the start of the investigation is very useful to ensure that a relevant difference can be demonstrated at all with the selected number of subjects.

Table 12.6 Example of a tabular display of the results of a two-factor analysis of variance

| | UV 1 Modality: verbal vs. manual | | UV 2 System: navigation vs. telephone number | | Interaction modality x system | |
|---------------|-------------------------------------|--------------|---|--------------|-------------------------------|--------------|
| | F (1,29) | p | F (1,29) | p | F (1,29) | p |
| SDLP | 9.1 | 0.005 | 4.6 | 0.040 | 5.2 | 0.029 |
| Reaction time | 56.9 | 0.000 | 0.9 | 0.345 | 11.3 | 0.002 |

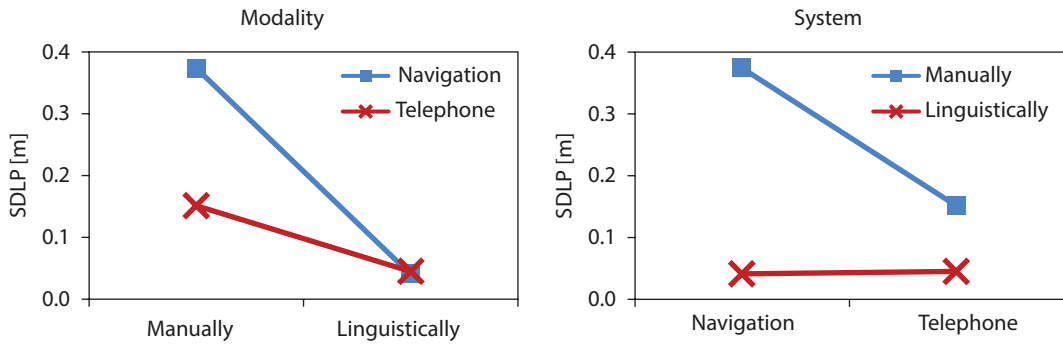
For the F-value, the degrees of freedom are given in brackets. SDLP is the Standard Deviation of Lane Position

12.3.3 Statistics: Presentation of Results

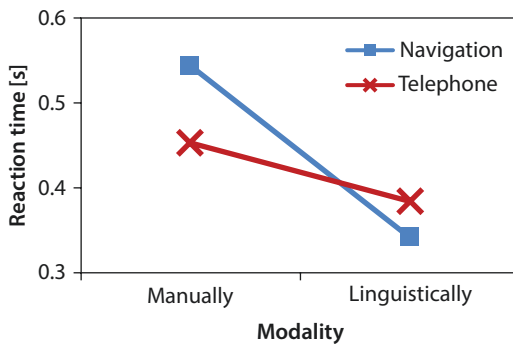
The significance test can only indicate whether the independent variable has had an effect. This is the necessary prerequisite for differences to be interpreted at all. If the test is not significant, the differences found may be random and should therefore not be described as effects. From this point on, the first step in interpreting the effects is to indicate where effects occurred at all. For this purpose, the results of the statistical tests are listed according to **Table 12.5**. In the case of a large number of dependent variables or complex experimental designs, a tabular display can be useful here. **Table 12.6** shows an example of a two-factor variance analysis. In the comparison of navigation system and the selection of a telephone number (IV 2: system) the difference between verbal and manual operation (IV 1: modality) was examined. The standard deviation of lane position (SDLP) and a reaction time to road signs requiring steering were measured. You can find the F- and p-values of the analysis of variance. Since the degrees of freedom for each test were the same, they are given in brackets after the F-value. The significant results at $\alpha = 5\%$ are shown in bold. It can be seen that in SDLP both main effects and the interaction are significant, in reaction time the main effect of the modality and the interaction.

Starting from such a result, the significant effects are then displayed and described graphically or in a table. The best type of representation, especially for two-factor experi-

mental designs, is the line graph, since this makes the different interpretation possibilities most clearly visible. Two types of representation are possible for two-factor experimental designs, as shown in **Fig. 12.8**. On the left side the blue line represents the navigation system, the red line the telephone. You can see that both systems have the lines running down, i.e. the SDLP is smaller for linguistic operation. With speech the lane keeping performance is better. This description corresponds to the main effect of the IV 1 “modality”. It can also be seen that there is no main effect for the IV 2 “system”, as the navigation system leads to a larger SDLP when operated manually, while there are no differences when operated verbally. Thus, the main effect of this IV cannot be interpreted, although it is significant. This different effect of the system depending on the modality corresponds to the significant interaction: The effect of one IV can only be interpreted as depending on the other IV. Each interaction can be interpreted in two directions: As an effect of IV 1 depending on the levels of IV 2 and vice versa. For example: The improvement of the SDLP through verbal operation is significantly stronger for the navigation system than for the telephone (Interpretation 1). When used manually, the SDLP is significantly worse with the navigation system than on the phone. This difference cannot be found in verbal operation (Interpretation 2). Both interpretations can be clearly seen in the left side of **Fig. 12.8**. Looking at the right side of the illustration, the second interpretation is more important here. The red line is parallel to the x-axis, while the blue line falls. Depending on which effects are



■ Fig. 12.8 Presentation of the effects for the example in ■ Table 12.6. The mean values are shown. For explanation, see text



■ Fig. 12.9 Results of the reaction time for the example from ■ Table 12.6. The mean values are shown. For explanation, see text

significant and significant for the reader, the right or left illustration may be more meaningful.

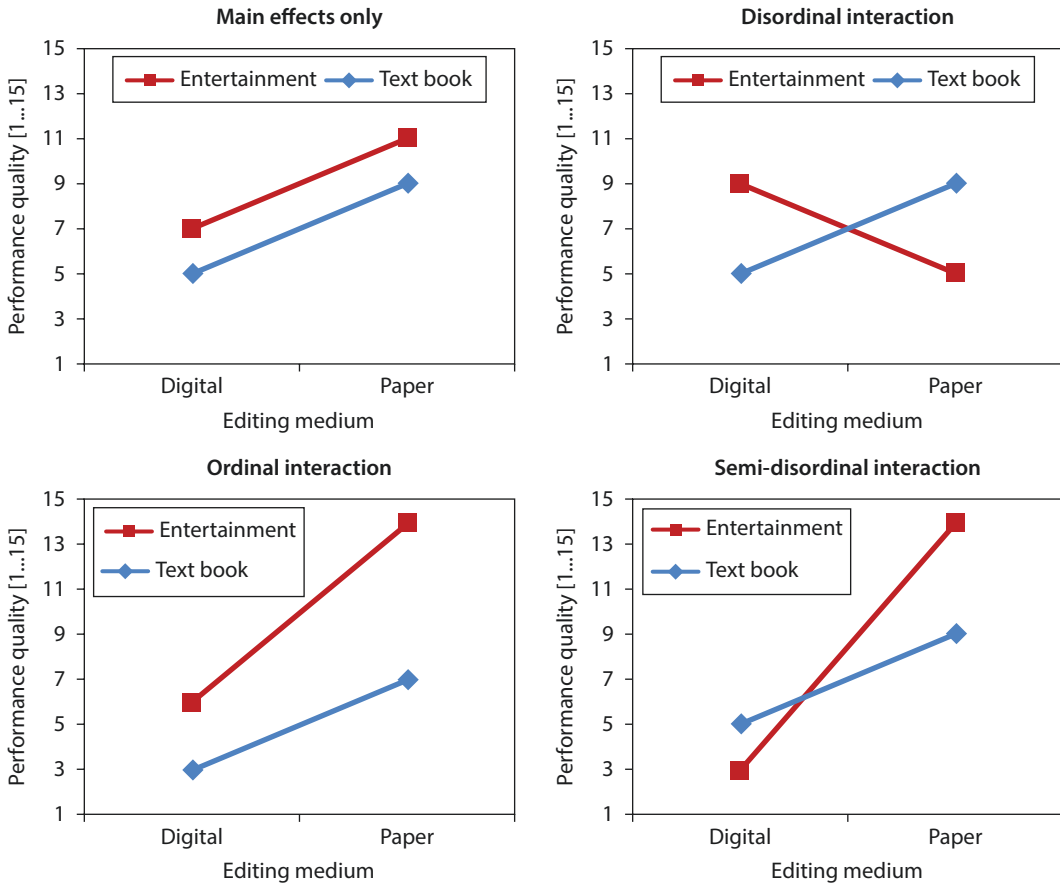
■ Figure 12.9 shows the results of the reaction time. Here the main effect of the modality and the interaction were significant. It can be seen that the reaction time is shorter with verbal operation than with manual operation. The interaction shows that the interaction with the navigation system benefits more from the verbal operation than the operation of the telephone. Or: With manual operation, the response time suffers more from the navigation system. The telephone is a bit worse when using the verbal operation.

This somewhat detailed example illustrates the different roles of the statistical test results and the description of the results. Not every statistically significant result can be interpreted, as the SDLP example shows. For the interpretation

a graphical description of the data is necessary. However, only those effects that were actually significant may be interpreted in the graphics. From this point of view it is very important to choose the right type of presentation.

In principle, different types of effect patterns can occur in two-factor experimental designs, which are often misinterpreted. ■ Figure 12.10 provides an overview of the most important types of effects. The example is a fictitious experiment to read fiction (entertainment) vs. a textbook either on the screen or on paper. The speed of reading was evaluated as a performance quality. The following four cases are important:

- Main effects only: At the top left you can see that the performance on paper is better for both types of text. Further one recognizes that the performance for the entertainment reading is better. Thus, here one finds the two main effects, but no interaction.
- Disordinal interaction: At the top right it becomes clear that no main effect can be interpreted, even if it were significant. The effect of the type of reading depends on the medium. With the digital medium, entertainment is better, on paper, textbook reading. The effect of the medium thus depends on the type of reading. Digital is better for entertainment reading, paper better for text books. Therefore, only the interaction is to be interpreted here.
- Ordinal interaction: At the bottom left you can see that both main effects and the interaction may be interpreted. On paper, reading is better than with the digital ver-



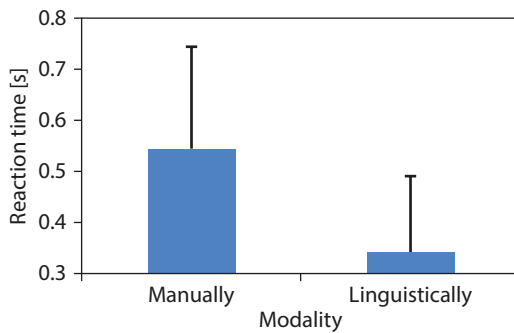
■ Fig. 12.10 Overview of types of interactions. For interpretation, see text

sion. Reading fiction (entertainment) is faster than reading a textbook (main effects). Furthermore, the advantage of entertainment reading on paper is stronger than with a digital medium. Or: the effect of paper is stronger in entertainment reading than in a text book (interaction).

- Semi-disordinate interaction: In the case shown at the bottom right, the main effect of the medium may be interpreted. On paper, performance is better than digital. The type of reading does not have a uniform effect and cannot therefore be interpreted as the main effect. The interaction is in turn interpretable if it is significant. The effect of the medium is stronger for entertainment reading than for text books. On a digital medium the textbook reads better, while on paper the entertainment reading is better to read.

These examples make it clear that the scientific content is not in the significance test, but in the graphical or tabular presentation of the measured values. The significance test is necessary for deciding what may be interpreted. With the corresponding presentation, it becomes clear to the reader what the effects mean. A list of the results of the significance tests is worthless without descriptive statistics and graphs.

Another useful way of displaying the data is via bar graphs. Often mean values and standard deviations are shown here (see ■ Fig. 12.11). Adjacent bars are very well suited for direct comparisons. The standard deviations are helpful in relativizing the size of the differences. As shown above, statistical testing compares the effect with a measure of random error to investigate significance. In a way, this is analogous to the representation of the mean values and the standard deviation.



■ Fig. 12.11 Mean value of the reaction time and standard deviation depending on the modality

Overall, it should be noted that graphics contain the relevant information in a clear and easy to understand manner. Axes must always be meaningful and labelled with units. It is not advisable to use colored backgrounds, 3-D representations, etc., as the data points will be pushed into the background compared to the graphic design elements and will then often be difficult to recognize. Also against this background line graphics are a very effective way of displaying.

12.4 External and Internal Validity

The description of the two central statistical approaches revealed a common ground. In both cases, an essential question is the representativeness of the results. This is also referred to as “external validity”, i.e. transferability. This depends above all on the drawing of an appropriate sample. Whenever not only statements about quantities to be measured directly are to be made, but these are to be interpreted, a second aspect of validity arises here: Are the characteristics valid, do they really measure what they are to measure? What question do you have to ask a driver to predict that he will buy a particular vehicle? This aspect of external validity is described in more detail in ► Sect. 11.1.1. In addition, there is a third aspect of external validity, which results from the examination situation. Are the data obtained e.g. in a driving simulator, representative for driving in one’s own vehicle in normal traffic? On the one hand, this is about the closeness to reality of the

study, on the other hand it is about the attitude of the participants. The better they succeed in conveying an understanding of the meaning and purpose of the investigation, the more they will be able to behave “normally”. Instruction, the clarification of the aims of the experiment, plays a central role here.

Under certain circumstances, however, it may also be necessary not to properly inform the test participants beforehand. If for example the effect of a collision warning system should be investigated, it is important that the drivers are surprised by a critical event similar to the one in real traffic. Therefore, it may be necessary to distract the attention of the drivers via a cover story in order to achieve a surprise effect in the driving simulator as well. For ethical reasons, the participants are to be informed in detail after the attempt. They must also be able to exclude their data from the experiment. In principle, any deception is ethically questionable and its use thoroughly weighed up.

In summary, three aspects of external validity can be distinguished:

- Representative sample
- Valid measurement methods
- Realistic situations with “normal” behaviour

In the second type of statistical question, the search for the effect of influencing factors, “internal validity” is added. It is a question of whether an effect found is actually undoubtedly attributable to the influence of the independent variable. This is only possible from the logic of the experiment, if the investigated groups differ only in the independent variables, otherwise they are treated completely identically. As described above, this is always a problem with repeated measurement designs when there may be fatigue or practice effects. If for example the trip with the visual warning system would always be the second trip, a practice effect could also explain the better reaction time. In order to ensure that the effect is actually due to the IV, the sequence of treatments will therefore always be permuted (see ■ Fig. 12.3). The internal validity thus depends on how well the influence of interference variables can be

eliminated or kept constant in the various conditions. This is the reason to standardize the test procedure including the instructions as much as possible in order to ensure a comparable treatment of all participants. Ultimately, however, internal validity cannot be guaranteed by a flowchart. As an investigator, one should always ask oneself whether a certain result could not also be explained by other factors than the influence of the independent variables. This critical way of thinking is an essential prerequisite for good research.

References

- Bortz, J., Schuster, C.: Statistik für Human- und Sozialwissenschaftler (Lehrbuch mit Online-Materialien). Springer, Berlin (2010)
- Faul, F., Erdfelder, E., Buchner, A., Lang, A.-G.: Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods*. **41**, 1149–1160 (2009)
- Sedlmeier, P., Renkewitz, F.: *Forschungsmethoden und Statistik in der Psychologie*. Pearson, München (2008)
- Wellek, S.: *Testing Statistical Hypotheses of Equivalence and Noninferiority*, 2. Aufl. Taylor & Francis, Boca Raton (2010)