applied. We will, in what follows, address these remarks, make alternative proposals, and try to analyze advantages and drawbacks of the alternative proposals.

POINT AND INTERVAL HYPOTHESES

Point Hypotheses. Hypothesis tests are carried out to decide which of two hypotheses, the null or the alternative hypothesis, is true. A point hypothesis is a hypothesis in which the significance of a result is tested against a single fixed point, e.g., zero. In our example of a method validation, the (point) null hypothesis (H0) is that the two means to be compared are estimates of the same population, i.e., that there is no bias between the outcomes of the new method and the reference method. The alternative hypothesis (H1) is that one must accept when the test indicates that the null hypothesis is not true, i.e., when H0 is rejected. H1, therefore, is the hypothesis that there is a bias.

The statistical hypotheses are

$$\text{H0:}\quad \mu_1 = \mu_2 \qquad \text{i.e., } \delta = \mu_2 - \mu_1 = 0 \text{ (there is no bias)} \tag{1}$$

$$\text{H1:}\quad \mu_1 \neq \mu_2 \qquad \text{i.e., } \delta = \mu_2 - \mu_1 \neq 0 \text{ (there is a bias)}$$

with δ the true difference or bias between the population means μ_1 and μ_2.

Generally, a t-test is then applied to test the hypothesis that the two methods yield identical results. H0 is accepted if the calculated t-value is smaller than the tabulated t-value. However, by applying the t-test, no attention is drawn to situations where, due to the sample size, the imprecision of the measurements is large and the probability of committing a β-error is therefore increased. The calculated t-value and the associated probability are simply accepted as they are. A more illustrative way of testing is to consider the $(1 - \alpha)$ confidence interval. Usually, these intervals are interpreted to include the true difference δ with the probability $(1 - \alpha)$. We will follow this interpretation, which can be considered Bayesian.[1] Since hypothesis tests are related to confidence intervals, one can also use the latter to test the significance of a hypothesis. H0, as specified higher, is accepted here if zero is included in the confidence interval.

Due to a bad measurement precision and a small sample size, an important difference can be declared to be nonsignificant by applying a point hypothesis test. On the other hand, measurements can be so precise that even a very small, nonrelevant difference is considered statistically significant. Notice that in both cases, the confidence interval provides more information than a simple hypothesis test. When as a result of a small sample size a wide confidence interval leads to acceptance of the null hypothesis, the reliability of the conclusion might be questioned. A very short confidence interval around an observed difference close to zero can be taken as a hint that the effect, although it is statistically significant, is not of practical relevance, e.g., for a bias evaluation. Although in both cases the confidence interval approach to point hypothesis testing gives an indication, it does not provide a convenient test. Interval hypothesis testing avoids the drawbacks of point hypothesis testing. Before discussing this, we first need to discuss errors attached to statistical hypothesis tests.

Errors Related to a Hypothesis. When testing a hypothesis, two kinds of errors can be made, namely α- and β-errors. Before a test is applied, one decides on the significance level, α, usually $\alpha = 5\%$. This is the risk one is willing to take of rejecting a true H0, i.e., of concluding that there is a bias when in fact there is none. One should, in principle, also take into account the β-error, the error of wrongly accepting H0. In our case, β is the risk to conclude that there is no bias between two methods when in fact there is one. The power of the test is given by $(1 - \beta)$. This is the probability of finding a difference when this difference is real. The β-error is related to the true precision, the true bias, the sample size, the chosen α-level, and the test statistic applied.

The α-error of a t-test can easily be controlled through the selection of the significance level. Since the true bias is unknown, the β-error cannot be obtained, but if an acceptable bias is specified, it is possible to calculate the probability that this bias will not be detected when it exists. One can also compute the number of experiments needed to make a decision with a specified β-error when it is known what bias is acceptable and one knows or has an estimate of the precision. However, the latter often is not known before the experiment has been carried out, because in practice, one often designs the experiment to estimate the precision parameters and the bias in the same experimental setup. In such a case, the result is that one can control α but not β. In other words, one controls the risk of deciding that there is a bias when this is not true but does not have an a priori idea of the probability that an unacceptable bias will be accepted. However, one may consider that it is more important not to accept a method that is biased to a too-large extent.

Considerations about α and β depend strongly on the consequences of making a wrong decision. In pharmacological studies, for instance, one will try to avoid wrongly concluding that there is a certain effect of a drug, since this would lead to the use of drugs that have no real therapeutic effect. The pharmacologist needs a small α-error and will focus on that error. This is the prevalent attitude through most statistics, but this is not always the best attitude. For instance, the person who studies the bioequivalence of drugs (i.e., Does a new formulation of a drug have the same concentration–time profile in the blood as an existing and accepted one?) is primarily concerned with not wrongly accepting bioequivalence, i.e., not wrongly accepting that there is no difference. This means that bioequivalence studies focus on the β-error. For bioequivalence studies, it has been decided that the consumer risk (the β-error) must not exceed 5%.[2]

In our opinion, analytical chemists are in the situation of bioequivalence specialists: we should avoid considering that methods are unbiased when that is not true. A candidate method should only be accepted when the risk of wrongly deciding that there is no bias is acceptably low (e.g., 5%). One should note that the point hypotheses, as applied higher, do not allow one to do this easily. In the following, it will be explained how the β-error can be fixed by using interval hypotheses.

Interval Hypotheses. In method validation, two methods to be compared will rarely yield exactly the same results. Whatever the difference between the methods, it will be detected as statistically significant by the point hypothesis if the sample sizes are large enough. Actually, there is, however, no need to demonstrate that the bias between two methods or two laboratories is zero. One only wants to demonstrate that the results obtained with the new method are equivalent to the ones of the

(2) Steinijans, V. W.; Hauschke, D. *Clin. Res. Reg. Affairs* **1993**, *10*, 203–220.

2505414279