ICSA Book Series in Statistics

Series Editors: Jiahua Chen · Ding-Geng (Din) Chen

Karl E. Peace
Ding-Geng Chen
Sandeep Menon *Editors*

Biopharmaceutical Applied Statistics Symposium

Volume 2 Biostatistical Analysis of Clinical Trials

Springer

Contents

Chapter 2
Generalized Tests in Clinical Trials

Stephan Ogenstad

2.1 Introduction

Conventional statistical methods do not provide exact solutions to many statistical problems, such as those arising in ANOVA, mixed models and multivariate analysis of variance (MANOVA), especially when the problem involves a number of nuisance parameters. As a result, users of these methods often resort to approximate or asymptotic statistical methods that are valid only when the sample size is large. With small or ordinary sample sizes, such methods often have poor performance (Weerahandi 1994). The approximate and asymptotic methods may lead to misleading conclusions or may fail to detect truly significance results from clinical studies.

Classical statistical tests may be insensitive to a wide range of situations occurring commonly in practice, particularly when the effect of the factor under study is heterogeneous. All statistical procedures are based on some distributional assumptions. In addition, many statistical procedures (e.g. ANOVA, ANCOVA) use the F-test and are based on the assumption of homoscedasticity (equal variances) and relate to the validity of the often convenient assumption that the structure of any one part of a dataset is the same as any other part. From experience, this assumption is seldom true when responses are different in the separate treatment groups. The assumption of equal variances is usually made for simplicity and mathematical ease rather than anything else. The outcome of using conventional statistical models when the assumptions are not reasonable can lead to serious consequences. In many situations, these procedures can fail to detect significant therapeutic effects even when available data provide sufficient evidence that the effects are present. In other applications, the conventional statistical models sometimes lead to incorrect conclusions, implying that the therapeutic results are significant when they are actually not (Blair and Higgins 1980; Brownie et al. 1990; Graubard and Korn 1987).

S. Ogenstad (✉)
Statogen Consulting LLC, 1600 Woodfield Creek Drive #215, Wake Forest, NC 27587, USA
e-mail: sogenstad@statogen.com

K. E. Peace et al. (eds.), *Biopharmaceutical Applied Statistics Symposium* ,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-981-10-7826-2_2

For instance, in the classical handling of the statistical problem in one-way ANOVA, it is assumed that the population variances are all equal. This is not really a natural assumption. In fact, it is often seen in most applications that the variances tend to be substantially different especially when the mean responses are substantially different. From simulation studies, it has also been observed that the assumption of equal variances is much more serious than the assumption of normally distributed populations, in that the former has the greater chance of leading to wrong conclusions. The classical ANOVA problems that rely on the equal variances assumption can dramatically reduce the power of the tests. Moreover, the magnitude of the lack of power problem of the tests based on the equal variance assumption increases with the number of treatments being compared. We also want to point out that in most applications, despite a common belief, it is not possible to transform data to achieve the approximate normality and equal variances simultaneously. The p-value produced from the classical approach is valid only if the variances are equal, and the test is not appropriate if the variances are significantly different.

In the analysis of repeated measures, it is also, assumed that all treatment groups have equal variances. While there is no serious problem when the assumption is reasonable, the assumption can lead to serious erroneous conclusions when the variances are substantially different. Moreover, in situations of higher-way ANOVA under an incorrect heteroscedasticity assumption, one is more prone to draw misleading conclusions. For instance, one can be misled by the classical F-test to conclude that a certain factor of an ANOVA is significant when in reality a different factor is significant.

Extensions have been made to the classical methods in repeated measures involving mixed models, MANOVA, and growth curves, in particular. Repeated measures and growth curves models are in fact special classes of mixed models. The classical approach to solving these problems provides exact solutions to only a fraction of the problems. Conventional methods alone do not always provide exact solutions to even some simple problems. For instance, in the univariate analysis of variance, the classical approach fails to provide exact tests when the underlying population variances are unequal. In some widely used growth curve models, there are no exact classical tests even in the case of equal variances. As a result, users of these methods often resort to asymptotic results in search of approximate solutions even when such approximations are known to perform rather poorly with moderate sample sizes.

Solutions to the statistical problems are addressed as extensions, as opposed to alternatives, to conventional methods of statistical inference. In Weerahandi (1994), each class of problems is started with a simple model under special assumptions that are necessary for the classical approach to work. After discussing solutions available for such special cases, these assumptions are relaxed when they are considered to be too restrictive or unreasonable in some applications, especially when they are known to have poor size (Type I error) or power performance. For instance, in fixed effects ANOVA, the problem is first considered under the homoscedastic variance/covariance assumption and then later the assumption is dropped.

The generalized methods are exact in the sense that the tests and the confidence intervals are based on exact probability statements rather than on asymptotic approximations. This means that inferences based on them can be made with any preferred accuracy, provided that assumed parametric model or other assumptions are correct. To make this possible, solutions to problems of testing various hypotheses are presented in terms of p-values. There is readily available computer software to implement these exact statistical methods. Exact p-values and confidence intervals obtained with extended definitions also serve to provide excellent approximate solutions in the classical sense. From simulation studies reported in the literature, type I error and power performance of these approximations are usually much better than the performance of more complicated approximate tests obtained by other means.

By exact generalized inference, we mean various procedures of hypothesis testing and confidence intervals that are based on exact probability statements. Weerahandi (1994) uses the term '*exact*' rather than '*generalized*' methods because these methods are not approximations to the problems but exact solutions. Here we confine our attention to the problems of making inferences concerning parametric linear models with normally distributed error terms. In particular, we do not address exact nonparametric methods that are discussed, for instance in Good (1994) and Weerahandi (1994). The purpose of this chapter is to provide a brief introduction to the notions and methods in the generalized inference that enable one to obtain parametric analytical methods that are based on exact probability statements.

There is a wide class of problems for which classical fixed-level tests based on sufficient statistics do not exist, and there are simple problems in which conventional fixed-level tests do not exist. For instance, consider the mean μ and variance σ^2 in a normal distribution $N(\mu, \sigma^2)$ and let us assume that the parameter of interest is the second moment of the normal random variable X about a point other than the mean, say k, then the parameter of interest is

$$E(X-k)^2 = \mu^2 + \sigma^2 - 2k\mu + k^2.$$

Classical tests are not available for this parameter unless $k = \mu$ (Weerahandi 1994). If instead, the parameter of interest is $\theta = \mu + k\sigma^2$, then it is possible but not easy to find a test statistic whose value and distribution depends on the parameters only through the parameter of interest, since either μ or σ^2 can be considered as the nuisance parameter.

Actually, these kinds of problems are prevalent even with widely used linear models. For instance, in the problem of comparing the means of two or more normal populations, exact fixed-level tests and conventional confidence intervals based on sufficient statistics are available only when the population variances are equal or when some additional information is available about the variances. The situation only gets worse in more complicated problems such as the two-way ANOVA, the MANOVA, mixed models, and in repeated measures models including crossover designs and growth curves.

In the application of comparing two regression models, Weerahandi (1987) gave the first introduction to the notion of generalized p-value and showed that it is an exact probability of an unbiased extreme region, a well-defined subset of the sample space formed by sufficient statistics. Motivated by that application, Tsui and Weerahandi

(1989) provided formal definitions and methods of deriving generalized *p*-values. In a Bayesian treatment, Meng (1994) introduced a Bayesian *p*-value, as a posterior predictive *p*-value, which is, under the noninformative prior, numerically equivalent to the generalized *p*-value. Weerahandi and Tsui (1996) showed how Bayesian *p*-values could be obtained for ANOVA-type problems that are numerically equivalent to the generalized *p*-values.

As discussed in detail in Weerahandi (1994), exact probability statements are not necessarily related to the classical repeated sampling properties. In special cases, the former may have such implications on the latter, but this is not something that one should take for granted. For instance, in applications involving discrete distributions, often we can compute exact *p*-values, but not exact fixed-level tests. Rejecting a hypothesis based on such *p*-values, say at the 5% level if $\mathrm{p}<0.05$, does not imply that the false positive rate in repeated sampling is 5%. Simply, such a *p*-value is a measure of false positive error and hence we can, in fact, reject the null hypothesis when it is less than a certain threshold. However, in most applications, fixed-level tests based on *p*-values, including the generalized *p*-values, do provide excellent approximate fixed-level tests that are better than asymptotic tests. Indeed, consistent with simulation studies reported in the literature (Gamage and Weerahandi 1998; Burdick et al. 2005), generalized tests based on exact probability statements tend to outperform, in terms of type I error or power, the more complicated approximate tests. Moreover, in many situations, type I error of generalized tests do not exceed the intended level. Therefore, procedures based on probability statements, that are exact for any sample size, are always useful, regardless of if we insist on repeated sampling properties or not. Also to those who insist on classical procedures, and anyone who has difficulties with the meaning of exactness, we can consider the generalized approach as a way of finding good approximate tests and confidence intervals, which are expected to perform better than asymptotic methods. We can benefit from the generalized approach to statistical inference, since it is an extension of the classical approach to inference as opposed to an alternative, providing solutions to a wider class of problems.

2.2 Test Variables and Generalized *p*-Values

Classical *p*-values as well as testing at a fixed nominal level, are based on what is known as test statistics. Basically, a test statistic is a function of some special properties of some observable dataset, that will distinguish the null from the alternative hypothesis. The function should not depend on any unknown parameters to qualify to be a test statistic. In the classical methodology of testing of hypotheses, this is an important requirement since, given a dataset, we should be able to compute such a statistic and compare against a critical value. Test statistics provide a convenient way of constructing extreme regions, on which *p*-values and tests can be based. But, this methodology only works in a very limited set of conditions (Weerahandi 1994). For instance, in the problem of sampling from a normal population, it is not clear how a test statistic could be constructed if the parameter of interest were a function such as, $\theta = \mu + \sigma^2$. The Behrens-Fisher problem is a well-known example of a circumstance where a test statistic based on sufficient statistics does not exist when the variances are not assumed to be equal. This limitation extends well into all types of linear models including ANOVA, regression models, and all types of repeated measures problems.

Tsui and Weerahandi (1989) introduced the notion of *test variables* in the context of generalized inference. Test variables provide a convenient way of defining extreme regions as they play the role of test statistics in the generalized setting since test variables are extensions of test statistics.

A generalized p-value is an extension of the classical p-value, which except in a limited number of applications, provides only approximate solutions. Tests based on generalized p-values are exact statistical methods in that they are based on exact probability statements. While conventional statistical methods do not provide exact solutions to such problems as testing variance components or ANOVA under unequal variances, exact tests for such problems can be obtained based on generalized p-values (Gamage et al. 2013; Hamada and Weerahandi 2000; Krishnamoorthy et al. 2006). In order to overcome the shortcomings of the classical p-values, Tsui and Weerahandi (1989) extended the classical definition so that one can obtain exact solutions for such problems as the Behrens–Fisher problem and testing variance components. This is accomplished by allowing test variables to depend on observable random vectors as well as their observed values, as in the Bayesian treatment of the problem, but without having to treat constant parameters as random variables.

To provide formal definitions, consider a random vector $\mathbf{Y}$ with the cumulative distribution function $F(\mathbf{y}; \boldsymbol{\xi})$, where $\boldsymbol{\xi} = (\theta; \boldsymbol{\delta})$ is a vector of unknown parameters. θ is the parameter of interest and $\boldsymbol{\delta}$ is a vector of nuisance parameters. Let $\mathbf{y}$ be the observed value of the random vector $\mathbf{Y}$. An extreme region with the observed sample point on its boundary can be denoted as $\mathrm{C}(\mathbf{y}; \theta, \boldsymbol{\delta})$. The boundary of extreme regions could be allowed to be any function of the quantities $\mathbf{y}, \theta$, and $\boldsymbol{\delta}$, and therefore, we need to allow test variables to depend on all these quantities. However, an extreme region is of practical use only if its probability does not depend on $\boldsymbol{\xi}$. Furthermore, a subset of the sample space obtained by more general methods should truly be an extreme region in that its probability should be greater under the alternative hypothesis than under the null hypothesis, as defined more formerly below.

Definition. A *generalized test variable* is a random variable of the form $T = T(\mathbf{Y}; \mathbf{y}, \boldsymbol{\xi})$ having the following three conditions:

1. The observed value $t = T(\mathbf{y}; \mathbf{y}, \boldsymbol{\xi})$ of T does not depend on unknown parameters.
2. The probability distribution of T does not depend on nuisance parameters.
3. Given t, $\mathbf{y}$ and $\boldsymbol{\delta}$, $P(T \leq t; \theta)$ is a monotonic function of θ.

2.3 Generalized Confidence Intervals

The classical approach to interval estimation suffers from more difficulties than that of hypothesis testing. Even when the problem does not involve nuisance parameters and there are exact confidence intervals, in some applications, they lead to results that contradict the very meaning of confidence. Both Ghosh (1961) and Pratt (1961) independently provided a very simple example of a uniformly most accurate confidence interval having highly undesirable properties, and connects two fundamental performance measures in confidence set estimation. Weerahandi (1994) showed how such undesirable confidence intervals can be avoided by expanding the class of intervals available to choose from. Just as in the case of testing of hypotheses, here we extend the class of available procedures for any given problem by insisting on exact probability statements rather than on sampling properties. This will enable us to solve such problems as the Behrens-Fisher problem for which exact classical confidence intervals do not exist. As in the Bayesian approach, the idea is to do the best with the observed data at hand instead of discussing other samples that could have been observed, was the process to be repeated. The generalized confidence intervals are nothing but the enhanced class of interval estimates obtained from exact probability statements with no special regard to repeated sampling properties that are of little practical use (Weerahandi 1994, 2004).

The definition of a confidence interval is generalized so that problems such as constructing exact confidence regions for the difference in two normal means can be undertaken without the supposition of equal variances. Under certain conditions, the extended definition is shown to preserve a repeated sampling property that a practitioner expects from exact confidence intervals. The proposed procedure can also be applied to the problem of constructing confidence intervals for the difference in two exponential means and for variance components in mixed models. With this description, we can carry out fixed level tests of parameters of continuous distributions on the basis of generalized p-values.

Thus, Weerahandi (1993) extended the conventional definition of a confidence interval in such a way that an applicably useful repeated sampling property is preserved. The research into this field was prompted by the need of exact confidence intervals in statistical problems involving nuisance parameters. For instance, even for a simple problem such as constructing confidence intervals for the difference in means of two exponential distributions, exact confidence intervals based on sufficient statistics are not available. The possibility of extending the definition of confidence intervals was suggested by the existence of p-values in this type of problem. Weerahandi (1987) used an extended p-value to compare two regressions with unequal error variances. The usefulness of generalized p-values explicitly defined by Tsui and Weerahandi (1989) is evident from a number of studies and applications, including those by Thursby (1992), Zhou and Mathew (1994), and Koschat and Weerahandi (1992).

To generalize the definition of confidence intervals, we first examine the properties of interval estimates obtained by the conventional definition. Consider a population represented by an observable random variable Y. Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ be a random sample of n observations from the population. Suppose the distribution of the random variable Y is known except for a vector of parameters $\boldsymbol{\xi} = (\theta, \boldsymbol{\delta})$, where θ is a parameter of interest and $\boldsymbol{\delta}$ is a vector of nuisance parameters. We are interested in finding an interval estimate of θ based on observed values of $\mathbf{Y}$. The problem is to construct generalized confidence intervals of the form $[A(\mathbf{y}), B(\mathbf{y})] \subset \Theta$, where Θ is the parameter space and $A(\mathbf{y})$ and $B(\mathbf{y})$ are functions of $\boldsymbol{y}$, the observed data.

In the classical approach to interval estimation we find two functions of the observable random vector, say $A(\mathbf{Y})$ and $B(\mathbf{Y})$ such that the probability statement

$$P[A(\mathbf{Y}) \leq \theta \leq B(\mathbf{Y})] = \gamma, \tag{2.1}$$

is satisfied, where γ is specified by the desired confidence level.

If the observed values of the two statistics are $a = A(\mathbf{y})$ and $b = B(\mathbf{y})$, then $[a, b]$ is a confidence interval for θ with the confidence coefficient γ. For instance, if $\gamma = 0.95$, then the interval $[a, b]$ obtained in this manner is called a 95% confidence interval. If in the situation of interval estimation of the parameter θ, the interval could be constructed a large number of times to obtain new sets of observation vectors $\boldsymbol{y}$, then the confidence intervals obtained using the formula (2.1) will correctly include the true value of the parameter θ 95% of the times. After a large number of independent situations of setting 95% confidence intervals for certain parameters of interest, we will have correctly included the true value of the parameter in the corresponding intervals 95% of the times. It, of course, has no implication about the coverage based on the sample that we have actually observed. Indeed, Pratt (1961), Ghosh (1961), and Kiefer (1977) provide examples where the current intervals violating the very meaning of *confidence*. In particular, they showed that in those applications the so-called exact confidence intervals do not contain the parameters at all. The only thing truly exact about a confidence interval is the probability statement on which the interval is based. If indeed repeated samples can be obtained from the same experiment, then the claimed confidence level will no longer be valid and in the limit, the value of the parameter will be known exactly, so that statistical inference on the parameter is no longer an issue. In view of this, Weerahandi (1993) searched for intervals that would enhance the class of solutions and extended the class of candidates eligible to be interval estimators by insisting on the probability statement only. This will allow us to find interval estimates for situations where it is difficult or impossible to find A($\mathbf{Y}$) and B($\mathbf{Y}$) satisfying (1) for all possible values of the nuisance parameters. He further showed how this can be accomplished by making probability statements relative to the observed sample, as done in the Bayesian approach, but without having to treat unknown parameters as random variables. More precisely, we can allow $A()$ and $B()$ to depend on the observable random vector $\mathbf{Y}$ and the observed data $\mathbf{y}$ both. When there are a number of parameters of interest, in general, we could allow subsets of the sample space possibly depending on the current sample point $\mathbf{y}$ of $\mathbf{Y}$.

Such intervals Weerahandi referred to as *generalized confidence intervals*. The construction of such regions can be facilitated by generalizing the classical definition of pivotal quantities. A random variable of the form $R = R(\mathbf{Y};\ \mathbf{y},\ \boldsymbol{\xi})$, a function of $\mathbf{Y}$, $\mathbf{y}$, and $\boldsymbol{\xi}$, is said to be a *generalized pivotal quantity* if it has the following two properties:

Property A: The probability distribution of R does not depend on unknown parameters.
Property B: The observed pivotal, $r_{obs} = R(\mathbf{y};\ \mathbf{y},\ \boldsymbol{\xi})$ does not depend on nuisance parameters $\boldsymbol{\delta}$.

Property A allows us to write probability statements leading to confidence regions that can be evaluated regardless of the values of the unknown parameters. Property B ensures that when we specify the region with the current sample point $\mathbf{y}$, then we can obtain a subset of the parameter space that can be computed without knowing the values of the nuisance parameters.

Suppose we have constructed a generalized pivotal $R = R(\mathbf{Y};\ \mathbf{y},\ \boldsymbol{\xi})$ for a parameter of interest and we wish to construct a confidence region at confidence coefficient γ. Consider a subset C_γ of the sample space chosen such that

$$P(R \in C_\gamma) = \gamma. \tag{2.2}$$

The region defined by (2.2) also specifies a subset C($\mathbf{y}$; θ) of the original sample space satisfying the equation $P(\mathbf{Y} \in C(\mathbf{y};\theta)) = \gamma$. Unlike classical confidence intervals, this region depends not only on γ and θ but also on the current sample point $\mathbf{y}$. With this generalization, we can obtain interval estimates on θ relative to the observed sample with no special regard to samples that could have been observed but were not. Although the generalized approach shares the same philosophy of the Bayesian approach that the inferences should be made with special regard to the data at hand, here we do not treat parameters as random variables and hence the probability statements are made with respect to the random vector $\mathbf{Y}$. Having specified a subset of the sample space relative to the current sample point, we can evaluate the region at the observed sample point and proceed to solve (2.2) for θ and obtain a region Θ_c of the parameter space that is said to be a $100\gamma\%$ generalized confidence interval for θ if it satisfies the equation

$$\Theta_c(r) = \{\theta \in \Theta | R(\mathbf{y};\mathbf{y},\xi) \in C_\gamma\},$$

where the subset C_γ of the sample space of R satisfies Eq. (2.2).

It should be reemphasized that generalized confidence intervals are not alternatives, but rather extensions of classical confidence intervals. In fact, for a given problem there is usually a class of confidence intervals satisfying the probability statement (2), a feature of classical intervals as well. Weerahandi (1994) discussed how the choice of appropriate generalized pivotals could be facilitated by invoking the principals of sufficiency and invariance. Even after we have obtained a particular pivotal quantity we could construct a variety of confidence regions. Depending on the application, a left-sided interval, a right-sided interval, a two-sided interval symmetric around the parameter, the shortest confidence interval, or some other interval might be preferable.

Comparing Two Normal Populations

In order to demonstrate the approach, we will show the case of comparing two normal populations. In the analysis of two-sample data, it is common to choose the t-test statistic to evaluate equality of the distributions. The test statistic is derived under the assumption of equal variances and independent normally distributed observations. We start by deriving the test statistic under this assumption, and later we derive a test variable when equality of variances is no longer assumed.

Let $X_1, \ldots, X_m$ be independent observations from a normal distribution $N(\mu_x, \sigma_x^2)$, and let $Y_1, \ldots, Y_n$, be independent observations from a normal distribution $N(\mu_y, \sigma_y^2)$. Then $\bar{X}$, $\bar{Y}$, S_x^2, and S_y^2 are the maximum likelihood estimators of μ_x, μ_y, σ_x^2, and σ_y^2, respectively. Since $\bar{X}$, $\bar{Y}$, S_x^2, and S_y^2 are complete sufficient statistics for the parameters of the two distributions, all inferences about the parameters can be based on them. The four statistics are independent, and their distributions are given by

$$\bar{X} \sim N(\mu_x, \frac{\sigma_x^2}{m}), \quad \bar{Y} \sim N(\mu_y, \frac{\sigma_y^2}{n}),$$

$$\frac{mS_x^2}{\sigma_x^2} \sim \chi^2_{m-1}, \quad \frac{nS_y^2}{\sigma_y^2} \sim \chi^2_{n-1}.$$

Under the assumption of equal variances ($\sigma^2 = \sigma_x^2 = \sigma_y^2$), inferences about the parameters can now be made on the basis of the complete sufficient statistics, $\bar{X}$, $\bar{Y}$, and

$$S^2 = \frac{\sum_{i=1}^m (X_i - \bar{X})^2 + \sum_{i=1}^n (Y_i - \bar{Y})^2}{m+n} = \frac{mS_x^2 + nS_y^2}{m+n}$$

and

$$\frac{(m+n)S^2}{\sigma^2} \sim \chi^2_{m+n-2}.$$

The parameter of primary interest is $\Delta = \mu_x - \mu_y$, and the hypotheses can be written as

$$H_0 : \Delta \leq 0 \,\text{versus}\, H_a : \Delta > 0$$

or

$$H_0 : \mu_x \leq \mu_y \text{ versus } H_a : \mu_x > \mu_y.$$

The family of joint distributions of $\bar{X}$, $\bar{Y}$, and S^2 is both location- and scale-invariant, so we can reduce the problem to tests based on the statistic $T = (\bar{X} - \bar{Y})/S$. Because the distribution of $\bar{X} - \bar{Y}$ can be standardized as

$$\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim N(\phi, 1)$$

the distribution of T is given by

$$\frac{T\sqrt{mn(m + n - 2)}}{\sqrt{m + n}} \sim t_{m+n-2\phi},$$

that is, the noncentral t-distribution with $m + n - 2$ degrees of freedom and the noncentrality parameter $\phi = \Delta/[\sigma\sqrt{1/m + 1/n}]$. The p-value is

$$P(T \geq (\bar{X} - \bar{Y})s^{-1} | \Delta = 0) = 1 - G_{m+n-2}((\bar{X} - \bar{Y})s^{-1}\sqrt{mn(m + n - 2)/(m + n)}),$$

where s is the observed pooled standard deviation, and G_{m+n-2} is the cumulative distribution function of Student's t-distribution with $m + n - 2$ degrees of freedom.

It is well known that the t-test is the uniformly most powerful unbiased test for the situation above. The Wilcoxon rank-sum test is almost as efficient under these conditions (Lehmann 1975; Hodges and Lehmann 1956). If the distributions are heavy-tailed, the Wilcoxon rank-sum test is a more efficient test. When the alternative involves a change in scale as well as in location $F_x(t) = F_y((t - \Delta)/\sigma)$, then both these tests may be inefficient.

When the variances are not equal we are still interested in the inference about the difference $\Delta = \mu_x - \mu_y$. This problem has no exact fixed-level conventional test based on the complete sufficient statistics (Linnik 1968; Weerahandi 1994).

For instance, consider constructing interval estimates based on functions of the observed data. The difference in sample means is location-invariant, and its distribution is $\bar{X} - \bar{Y} \sim N(\Delta, \sigma_x^2/m + \sigma_y^2/n)$. The generalized pivotal quantity

$$R = (\bar{X} - \bar{Y} - \Delta)\sqrt{\frac{\sigma_x^2 s_x^2/(mS_x^2) + \sigma_y^2 s_y^2/(nS_y^2)}{\sigma_x^2/m + \sigma_y^2/n}}$$

can generate all invariant interval estimates '*similar*' in σ_x^2 and σ_y^2. Furthermore, let

$$Z = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\sigma_x^2/m + \sigma_y^2/n}}, \quad Y_x = mS_x^2/\sigma_x^2, \quad Y_y = nS_y^2/\sigma_y^2$$

where $Z \sim N(0, 1)$, $Y_x \sim \chi^2_{m-1}$, and $Y_y \sim \chi^2_{n-1}$ are all independent random variables. Moreover, the random variables $Y_x + Y_y \sim \chi^2_{m+n-2}$ and $B = Y_x/(Y_x + Y_y) \sim$

$Beta[(m-1)/2, (n-1)/2]$, and Z are also independently distributed. The pivotal quantity now becomes

$$R = Z\sqrt{s_x^2/Y_x + s_y^2/Y_y} = Z(Y_x + Y_y)\sqrt{s_x^2/B + s_y^2/(1-B)}.$$

Interval estimates of Δ based on R can be obtained from probability statements about R. The cumulative distribution function of R can be expressed as

$$P\{R \le r\} = P\left\{T \le r\sqrt{\frac{m+n-2}{s_x^2/B + s_y^2/(1-B)}}\right\} = EG_{m+n-2}\left\{r\sqrt{\frac{m+n-2}{s_x^2/B + s_y^2/(1-B)}}\right\}$$

where G_{m+n-2} is the cumulative distribution function of T and the expectation, E, is taken with respect to the beta random variable B.

The constant $c_\gamma = c_\gamma(s_x^2, s_y^2)$ needs to be found to satisfy

$$EG_{m+n-2}\left\{c_\gamma\sqrt{\frac{m+n-2}{s_x^2/B + s_y^2/(1-B)}}\right\} = \gamma.$$

A $100\gamma\%$ one-sided generalized confidence interval of Δ is $\left[(\bar{\mathrm{X}} - \bar{\mathrm{Y}}) - c_\gamma(s_x^2, s_y^2), \infty\right]$. A symmetric confidence interval about the point estimate $(\bar{X} - \bar{Y})$ of Δ is

$$(\bar{X} - \bar{Y}) - c_{(1+\gamma)/2}(s_x^2, s_y^2) \le \Delta \le (\bar{X} - \bar{Y}) + c_{(1+\gamma)/2}(s_x^2, s_y^2)$$

(Ogenstad 1998; Weerahandi 1994).

2.4 Illustrations

One-Way ANOVA Comparing Three Groups

Suppose that we have a dataset such that for comparing the mean effects of two active treatments (B and C) and a placebo (A). As can be experienced from analyzing a number of datasets, it is common that the variability in responses will increase with increasing mean levels. Let us say that after a preliminary review of the data and the figure we produced below (Fig. 2.1), based on equal sample sizes in the treatment groups, our '*intuition*' tells us that the treatment means are significantly different.

Although these data were indeed generated from normal populations with unequal means and variances, application of the classical F-test will not support our '*intuition*' in this case at all, because the p-value of the usual F-test is as large as 0.16. Using XPro (X-Technologies, Inc.), a software that calculates exact p-values, we compute the p-value for testing the equality of treatment means under the more reasonable assumption of unequal variances. The XPro Software produces a p-value that is 0.043, which is in line with the impression we get from the figure that we constructed. The discrepancy in p-values in this example is quite dramatic. It clearly demonstrates

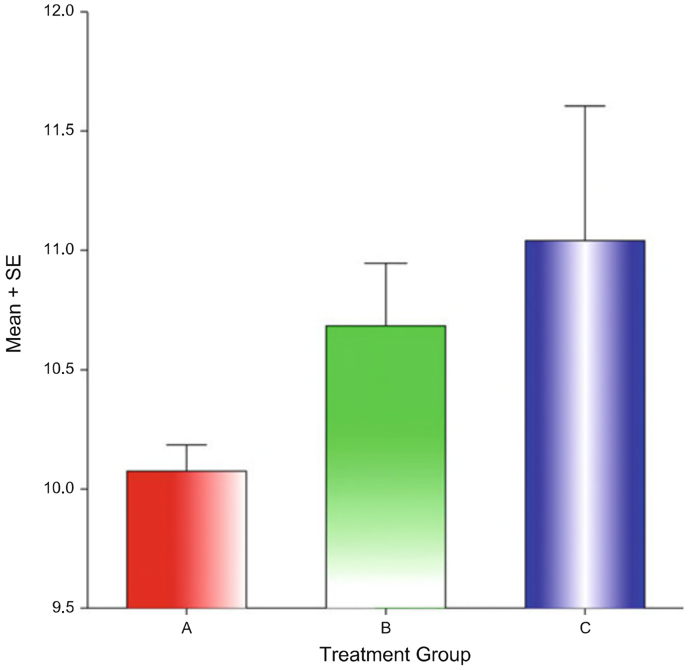


Fig. 2.1 Treatment group means + standard errors, based on equal sample sizes in the treatment groups

the serious weakness of the classical F-test in the presence of heteroscedasticity. Because the test ignores the problem of heteroscedasticity, the classical F-test fails to detect significant differences in treatments, despite the fact that the data provides sufficient information to do so. The complete ANOVA table to this illustration can be found in Appendix. As a note, the F-test is even more unreliable if the sample sizes in the treatment groups are different.

One-Way ANOVA Comparing Seven Groups

Although, based on equal sample sizes in the treatment groups, the treatment effects to the naked eye are quite different (Fig. 2.2), the p-value when applying the classical ANOVA to test the null hypothesis of equal means against the alternative hypothesis that not all means are equal is 0.11, which is not statistically significant at the 5% significance level. With the generalized F-test, the p-value without the equal variances assumption is 0.0098, which shows a very different outcome.

Repeated Measures Under Heteroscedasticity

We will now show an example of hemodynamic monitoring, which has long formed the cornerstone of heart failure (HF) and pulmonary hypertension diagnosis and management. There is a long history of invasive hemodynamic monitoring initially using pulmonary artery (PA) pressure catheters in the hospital setting, to evaluate the

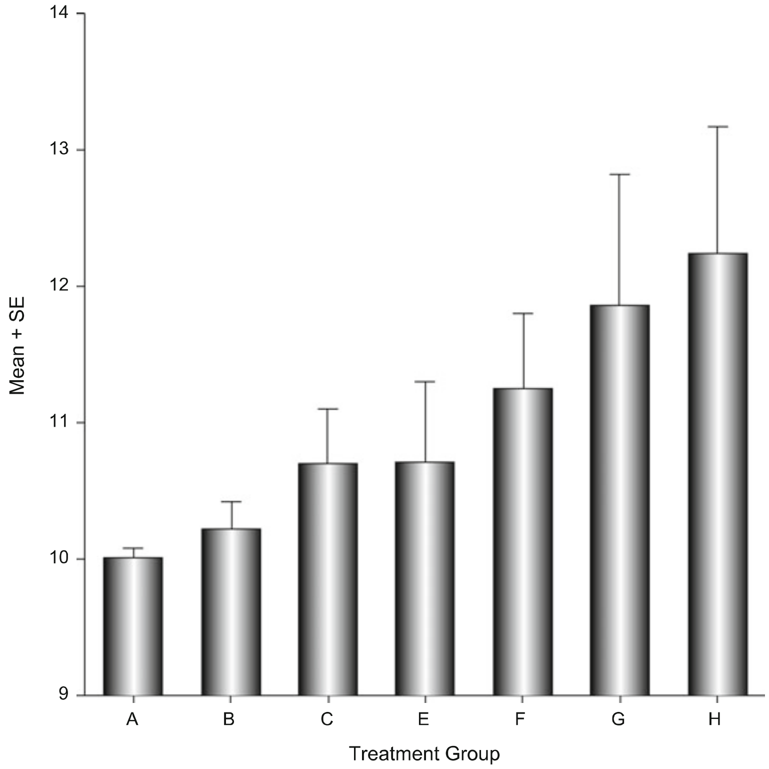
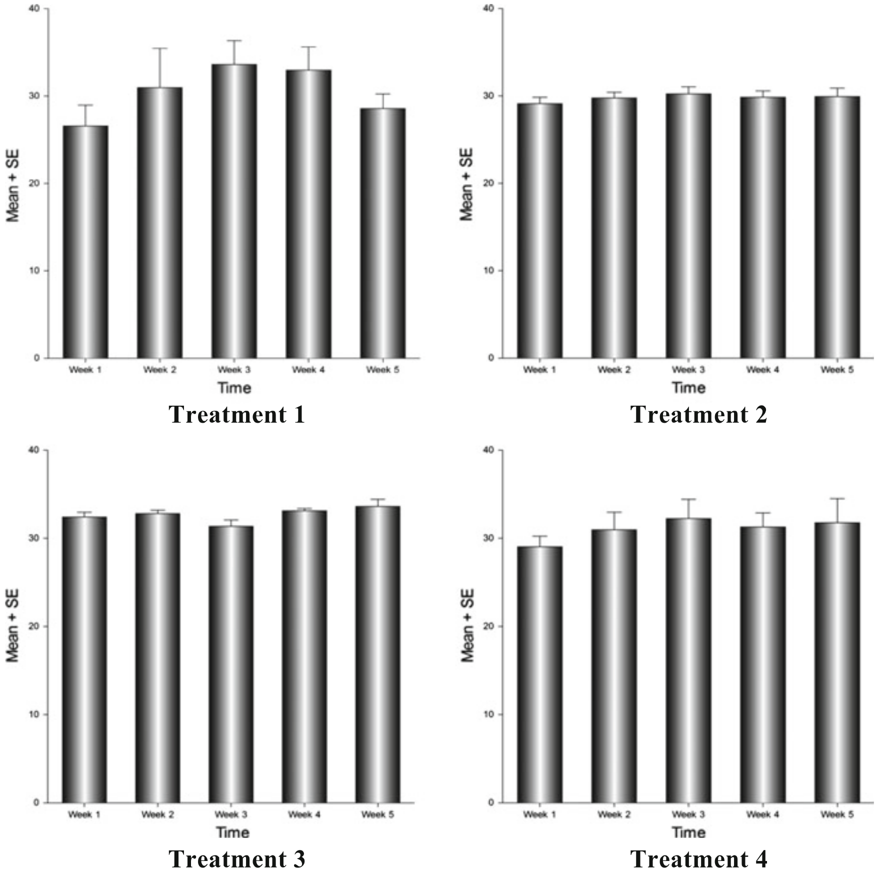


Fig. 2.2 Treatment group means + standard errors, based on equal sample sizes in the treatment groups

utility of a number of implantable devices that can allow for ambulatory determination of intracardiac pressures. Although the use of indwelling PA catheters has fallen out of favor in a number of settings, implantable devices have afforded clinicians an opportunity for objective determination of a patient's volume status and pulmonary pressures. Some devices, such as CardioMEMS' and thoracic impedance monitors present as part of implantable cardiac defibrillators, are supported by a body of evidence that show the potential to reduce HF-related morbidity and have received regulatory approval, whereas other devices have failed to show benefit and, in some cases, harm (Davey and Raina 2016).

We will consider potential data on pulmonary artery pressure where patients have been placed on one of four treatments ($G = 4$) to bring down the PA pressure. The patients have five scheduled visits at weeks 1, 2, 3, 4, and 5 with their investigator. Shown in Fig. 2.3 are bar graphs reflecting the arithmetic means, based on equal sample sizes in each group, with standard errors of a hypothetical dataset of normally distributed observations that was generated by simulating the following model



Treatment 1

Treatment 2

Treatment 3

Treatment 4

Fig. 2.3 Treatment group means + standard errors, based on equal sample sizes in the treatment groups and weeks

$$Y_{i(g)t} = \theta_g + \beta_t + \gamma_{gt} + \alpha_{i(g)} + \varepsilon_{i(g)t},$$

where $t = 1, \ldots, 5$, $i(g) = 1, \ldots, n_g$, $g = 1, \ldots, 4$. $\alpha_{i(g)}$ is the random effect due to among-subject variation, θ_g, $g = 1, \ldots, 4$ are the treatment effects, β_t, $t = 1, \ldots, 5$ are effects due to visits, γ_{gt} are their interactions, and ε_{it} are the residual terms.

Extending the usual assumption about variance components to possibly unequal group variances, we now have

$$\alpha_{i(g)} \sim N(0, \sigma_\alpha^2), \quad \varepsilon_{i(g)t} \sim N(0, \sigma_g^2),$$

where $t = 1, \ldots, 5$, $i(g) = 1, \ldots, n_g$, $g = 1, \ldots, 4$.

Although the data seems typical in a repeated measures design, a closer look at the data reveals that the treatment group variances, in this case, are substantially

Table 2.1 Classical analysis of variance results

ANOVA table					
Source	DF	SS	MS	F-value	P-value
Weeks	4	86.8247	21.7062	1.289	0.284
Treatments	3	110.992	36.9972	2.137	0.136
Within Treatment	16	276.997	17.3123		
Treatment × Weeks	12	135.048	11.254	0.668	0.775
Error	64	1078.08	16.845		
Total	99	1687.94			

different, which is evident in Fig. 2.3. Obviously, in this application, it is not reasonable to assume that the variances are equal. But should it make any difference to our conclusions whether or not the assumption is reasonable? To examine this, let us first ignore the fact that variances are different and apply the classical ANOVA as usually done by most people. The ANOVA table (Table 2.1) obtained by applying formulas for classical repeated measures analysis for the case of homoscedastic variances is shown below.

According to the p-values appearing in the ANOVA table, none of the effects including the treatment effect are significant. Now we will drop the equal variances assumption and retest the hypothesis that there is no difference in the mean PA pressures between the different treatments. The p-value for testing the difference between the treatments then becomes 0.0009. This means that the difference between the treatments is highly significant despite what the classical ANOVA suggested. Usually milder assumptions make the p-value of a test larger and power of a test smaller. But here the assumption of equal variances is so unreasonable that the p-value under the assumption of equal variances is substantially larger. This illustration clearly displays the reduction of the power of classical F-tests under heteroscedasticity.

2.5 Statistical Software

XPro computes exact p-values for testing hypotheses and computes confidence intervals based on exact probability statements. This becomes particularly important when one is using small or unbalanced data. The assumptions upon which standard methods are based are then typically biased, resulting in unrealistic p-values and confidence intervals. The software supports the exact inference in various linear models. It has been proven to be able to detect significant and nonsignificant experimental results early, even with small sample sizes. XPro procedures are complimentary to such program as StatXact which specialize in exact non-parametric methods, such as those dealing with contingency tables and categorical data. Most software programs provide exact parametric methods only under the assumption of homoscedasticity in the ANOVA. In addition to such classical procedures, XPro provides procedures based on milder assumptions. To make this possible XPro performs high dimensional numerical integrations and solves highly nonlinear equations. The complexities of the underlying formulas make the problem of computing exact *p*-values and confidence limits very tedious. XPro makes use of efficient algorithms tailor made for exact inferences in linear models and provides an easy to use interface that facilitates all necessary analyses without passing the burden of any such numerical methods to the user. The methods used are based on Weerahandi (1994). *P*-values and confidence intervals, based on exact statistical calculations, are provided for a large number of following statistical procedures, models, and relationships.

As mentioned, StatXact (Cytel Corporation), is used for a host of nonparametric statistical procedures and sample size determination, and LogXact (Cytel Corporation), for the construction of logistic and Poisson regression models. Both StatXact and LogXact allow the user to select exact, Monte Carlo, or regular asymptotic methods of calculating *p*-values and confidence intervals. If exact methods take too long or are unavailable because of computer memory limitations, the user may select Monte Carlo techniques. Monte Carlo results are often very close to those produced by exact methods. XPro likewise provides the user with a Monte Carlo option for the majority of its procedures. It is generally used under the same conditions mentioned above.

Appendix

One-way ANOVA table

```
               Sample sizes and MLEs of parameters
Column        Sample size         Sample mean        Sample variance
A               20                   10.0752           0.237953
B               17                   10.6838           1.08007
C               19                   11.0419           5.66803

                       ANOVA Table
Source       DF           SS            MS             F-value
Treatment     2          9.3291        4.66455         1.88988
Error        53         130.813        2.46817
Total        55         140.142


           Testing the Equality of All Means

              Classical F-Test
P-value under the equal variances assumption:        0.161

              Generalized F-Test
P-value without the equal variances assumption:      0.043
```

References

Blair, R. C., & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's t statistic under various non-normal distributions. *Journal of Educational and Statistics, 5,* 309–335.

Brownie, C., Boos, D. D., & Hughes-Oliver, J. (1990). Modifying the t and ANOVA F-tests when treatment is expected to increase variability relative to controls. *Biometrics, 46,* 259–266.

Burdick, R. K., Park, Y.-J., Montgomery, D. C., & Borror, C. M. (2005). Confidence intervals for misclassification rates in a gauge R&R study, *Journal of Quality Technology*.

Davey, R., & Raina, A. (2016). Hemodynamic monitoring in heart failure and pulmonary hypertension: From analog tracings to the digital age. *World Journal of Transplantation, 6*(3), 542–547.

Gamage, J., Mathew, T., & Weerahandi, S. (2013). Generalized prediction intervals for BLUPs in mixed models. *Journal of Multivariate Analysis*, *220*, 226–233.

Gamage, J., & Weerahandi, S. (1998). Size performance of some tests in one-way ANOVA. *Communications in Statistics Simulation and Computation, 27,* 625–640.

Ghosh, J. K. (1961). On the relation among shortest confidence intervals of different types. *Calcutta Statistical Association Bulletin* 147–152.

Good, P. (1994). *Permutation tests—a practical guide to resampling methods for testing hypotheses*. Springer.

Graubard, B. I., & Korn, E. L. (1987). Choice of column scores for testing independence in ordered 2 × K contingency tables. *Biometrics, 43,* 471–476.

Hamada, M., & Weerahandi, S. (2000). Measurement system assessment via generalized inference. *Journal of Quality Technology, 32,* 241–253.

Hodges, I. L., & Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics, 27,* 324–335.

Kiefer, J. (1977). Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association, 72,* 789–808.

Kim, H. (2008). Moments of truncated Student-t distribution. *Journal of the Korean Statistical Society, 37,* 81–87.

Koschat, M. A., & Weerahandi, S. (1992). Chow-type tests under heteroscedasticity. *Journal of Business & Economic Statistics, 10*(22), 1–228.

Krishnamoorthy, K., Mathew, T., & Ramachandran, G. (2006). Generalized P-values and confidence intervals: A novel approach for analyzing lognormally distributed exposure data. *Journal of Occupational and Environmental Hygiene, 3,* 642–650.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks*. Oakland, CA: Holden-Day.

Linnik, Y. (1968). Statistical Problems with Nuisance Parameters. Translation of Mathematical mono-graph No. 20, American Mathematical Society, New York.

Meng, X. L. (1994). Posterior Predictive p-values. *The Annals of Statistics, 22*(3), 1142–1160.

Ogenstad, S. (1998). The Use of Generalized Tests in Medical Research. *Journal of Biopharmaceutical Statistics, 8*(4), 497–508.

Pratt, J. W. (1961). Length of confidence intervals. *Journal of the American Statistical Association, 56,* 541–567.

Thursby, J. G. (1992). A comparison of several exact and approximate tests for structural shift under heteroscedasticity. *Journal of Econometrics, 53,* 363–386.

Tsui, K., & Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association, 84,* 602–607.

Zhou, L., & Mathew, T. (1994). Some tests for variance components using generalized p-values. *Technometrics, 36,* 394–421.

Weerahandi, S. (1987). Testing regression equality with unequal variances. *Econometrica, 55,* 1211–1215.

Weerahandi, S. (1993). Generalized confidence intervals. *Journal of the American Statistical Association, 88,* 899–905.

Weerahandi, S. (1994). *Exact statistical methods for data analysis*. New York: Springer.

Weerahandi, S. (2004). *Generalized inference in repeated measure* (p. 44) New York: Wiley.

Weerahandi, S., & Tsui, K. (1996). Solving ANOVA problems by Bayesian approach, comment on posterior predictive assessment of model fitness via realized discrepancies by Gelman, Meng, and Stern, Statistica Sinica 6, 792–796.

X-Techniques, Inc (1994). XPro: Exact Procedures for Parametric Inference, X-Techniques, Inc, Millington, NJ.