Monte Carlo methods for modified ANOVA

Li, L.

Award date:
2007

Monte Carlo Methods for Modified ANOVA

LI LI

Supervisors:

Prof. Dr. Jurgen Franke

Prof. Dr. R.M.Mattheij

Dipl.-Math. Sascha Feth

MASTER THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS OF THE DEGREE OF
MASTER OF SCIENCE
AT
TU KAISERSLAUTERN

KAISERSLAUTERN, GERMANY
AUGUST 2007

To

My Father

and

Memory of Studentship

Acknowledgements

First I would like to thank the department of 'Mathematical Methods for Dynamics and Durability' at Fraunhofer ITWM for supporting and offering me the good opportunity to do my thesis at the place.

I would like to express my gratitude to Prof. Dr. Franke, TU Kaiserslautern, for his guidance, valuable advices and remarks and to Prof. Dr. Mattheij, TU Eindhoven, for his contribution in correction of the thesis.

I am deeply grateful to Herrn Sascha Feth at ITWM, for his many instructive mini lectures, remarkable suggestions, patience and constant support throughout the work.

At last I would like to thank both faculties of mathematics in TU/e and in TU KL for making me close to a knowledgable person.

Abstract

One of tasks from industrial manufacture is fatigue design of mechanical components, to optimize the loading of the component and its strength or lifetime. Statistical tests are needed to identify effects on the service life of components with respect to the varying of loads. This paper is about testing the euqality of group means when experimental data from fatigue tests are subject to censorship, nonnormal distribution and heteroscedasticity. Commonly used tests, they cannot give a satisfactory result with such data due to restrictive assumptions and it's desirable to develop a more robust and powerful test procedure. In the paper it has proposed a approach to modify ANOVA F test, Kaplan Meier, Satterthwaite and permutation methods are applied for compensating drawbacks of ANOVA F test caused by violation of assumptions. The modified test has been compared with two alternative tests namely likelihood ratio test and Welch test. The power of tests is evaluated by Monte Carlo simulation. The underline study has proved that the modified ANOVA F is the best among the three tests with respect to the power for small sample size.

Contents

Chapter 1

Introduction

This chapter includes a general introduction to analysis of variance (ANOVA) and its widely applications subject to certain assumptions. As a matter of fact, data collected in real situation usually violate assumptions to some degree. A practical example is proposed to illustrate this point and the characteristic of survival data is also present. The power function will be used as a way to measure the test statisitcs. At last general structures of this paper will be present.

1.1 General Idea of ANOVA

The one-way fixed effects ANOVA test[1]is the most often used statistical method today for comparing the effects of several fixed groups. This name ANOVA derives from the fact that in order to test for statistical significance between means, we are actually analyzing variances. The underlying assumptions associated ANOVA include equality of variances, normality of distributions, and independent uncensored of data set[2]. The restrictive assumptions that may be considered unrealistic in some situations. We give one of the practical tasks in mechanical reliability from Fraunhofer Institute Techno- und Wirtschaftsmathematik (ITWM) as a example to illustrate the real situation when the experiment data are not met the assumptions.

1.1.1 Experiment for Fatigue Test

Vehicle manufacturer concerns the quality and reliability of products. Important decisions are made based on the results of statistical tests from data obtained by the experimental design. One test for assessment of life time against fatigue needs to point out if there is significant influence from loads on the life time for certain parts of components.

[1]Refer chapter 2
[2]Refer section 1.1.2

Regarding material fatigue, some parts of vehicle components are designed to be resistant to fatigue. Fatigue life assessment is calculated for spectrum loadings to checkout whether the components has fatigue limit subject to the magnitude of loads. A test is constructed to make a statement with regard to the expected service life under certain loads. In the course of the fatigue test, certain level loadings for the components are determined. A so-called Woehler curve is then ascertained and it shows that lifetime or alternating load cycles is represented as a function of load. The general relation between load S and load cycles N until failure is given by:

$$N = \frac{c}{S^k} \tag{1.1}$$

c is capacity and k Woehler exponent. One can categorize three types of loads for compo-

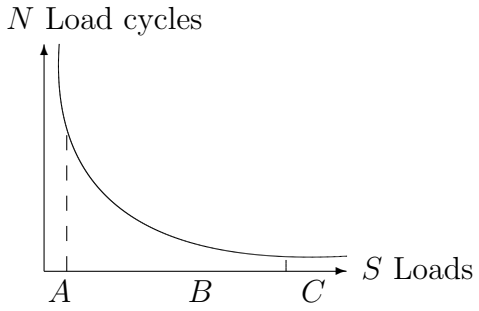

Figure 1.1: Woehler curve

nents as marked in the graph (1.1.1). Area A denotes the spectrum of the load almost gives influnce change in performance in confined time, the line is nearly vertical, B indicates the area life time shows the moderately change due to the loads where the relation (1.1) is held, and area C means after certain level the loads are so large for the components that life time is terminated immediately where the line is nearly horizontal. We put our emphasis on the loads in the area B which are chosen reasonably when constructing the experiment, that means, neither unbearably large nor negligibly small loads.

Taking the logarithms of equation (1.1)leads a form:

$$\log N = -k \log S + \log c \tag{1.2}$$

Rewriting equation (1.2), If one puts $y = \log N$ and $x = \log S$, once again we have a straight line through the origin which is more straightforward for observation,

$$y = -kx + \log c \tag{1.3}$$

During the experiment for the fatigue strength on the components, the failed service time is recorded. One example is showing the scattered failure points in the figure (1.1.1), and the main task is to examine whether there is a remarkable effect from the loads on the lifetime, or the parameter k of the regression line equal to zero.

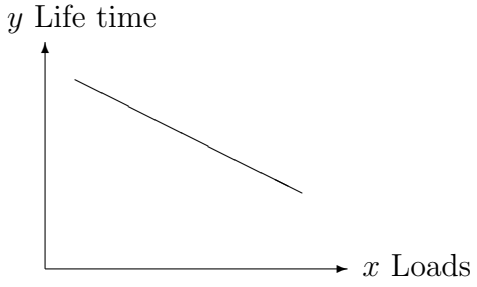

Figure 1.2: Woehler curve after logarithms transformation

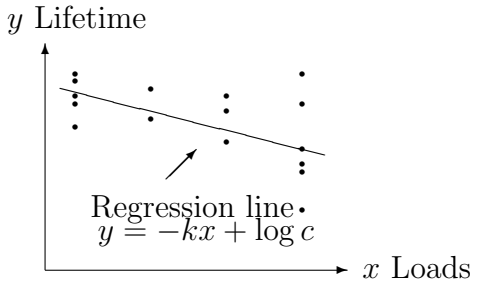

Figure 1.3: Scattergram

1.1.2 Survival Data Analysis

Before conducting a statistical test for identifying whether there exits significant influence from loads on the lifetime as auto manufacturers are interested in, at first we are going to describe experimental data situation. Typically there are several features for this survival data:

- Censoring
- Nonnormal distribution
- Heterogeneity

Censoring

A special source of difficulty in the analysis of survival data is the possibility that some individuals may be not observed for the full time to failure. At the close of life-testing experiment, not all components may have failed especially at the level of the lower loads in our fatigue test. Such incomplete observation of the failure time is called censoring and

the period of observation for censored individuals also be recorded. Generally speaking there are three types of censoring

1. Type I Censoring or time censored. Observations ceases after a fixed time c_i, after which the individuals that could not fail are observed to be censored.

2. Type II Censoring or failure censored. Observations ceases after a predetermined number d of failures, after which the left ones are observed to be censored.

3. Type III Censoring or progressively censoring. The entry times of observations are not simultaneous and the censored times are also different.

In the following tests we consider type I censoring, in the absence of censoring, the ith individual in a sample of n has failure time T_i. We suppose also that there is a period of observation c_i such that observation on that individual ceases at c_i if failure has not occurred by then. Then the observations consist of $X_i = \min(T_i, c_i)$, together with the indicator variable $w_i = 1$ if $T_i \leq c_i$(uncensored), $w_i = 0$ if $T_i > c_i$(censored). In our experiments, c_i will be known and equal under the control as the only cause of censoring is the planned ending at a predetermined time.

Distribution of Failure Time

Checking the survival data at each load point, one can find out the survival data is no long normally distributed. Skewness and heavy tail are often observed in the survival data. Here we give the probability density function(pdf) for two very useful parametric models commonly used in reliability analysis.

Lognormal distribution. For the lognormal distribution the pdf is:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\log x - \mu)^2\right) \tag{1.4}$$

The logarithm of a lognormal random variable $Y = \log x$ follows a normal distribution with mean μ and standard deviation σ. This relationship between the lognormal and normal distribution is often used to simplify the process of using the lognormal distribution.

Weibull distribution. Weibull distribution is a generalization of the exponential distribution, however it has broader application. For the Weibull distribution the pdf is:

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} \tag{1.5}$$

For $x > 0$ and $f(x) = 0$ for $x < 0$, where the value of k determines the shape of the distribution and λ determines its scaling. The expected value and standard deviation of a Weibull random variable can be expressed as:

$$E(X) = \lambda\Gamma(1 + 1/k) \tag{1.6}$$

$$\mathrm{Var}(X) = \lambda^2[\Gamma(1 + 2/k) - \Gamma^2(1 + 1/k)] \tag{1.7}$$

Where Γ is the Gamma function which is defined by $\Gamma(x) = \int_0^\infty t^{x-1}e^{-x}dt$.

We give Weibull distribution to represent the loss of normality in data set in our test.

Heteroscedasticity

The survival data in each group are not regarded to have homogenous variance due to some material properties. In the fatigue test each failure data obtained, is time consuming and costly. The sample sizes due to such reason are small, so finite sample size makes the variance in each group more different. Usually sample sizes were chosen to compensate for the fact that fewer failures are expected at the lower load test.

1.2 Measure of the Test

The example shows that the assumptions for the ANOVA are not realistic. Test statistics has to be examined if it is still robust or could be modified to apply under violated situations. Power function is defined to measure if a test is 'powerful' or 'robust'.

1.2.1 Power Function

To judge the quality of the test statistics, power function is introduced. Power is broadly defined as "the probability that a statistical significance test will reject the null hypothesis for a specified value of an alternative hypothesis". Another way to define it is "the ability of a test to detect an effect, given that the effect actually exists."

When performing a power analysis, we'll need to consider the following important information:

- Significance level or the probability of a Type I error.
- Power to detect an effect. This is expressed as power $= 1 - \beta$, where β is the probability of a Type II error.
- Sample size, or the number of units accessible to the study.

A type I error occurs when one rejects the null hypothesis when it is true. The probability of a type I error is the level of significance of the test of hypothesis, and is denoted by α. The most commonly used criteria α are 0.05, 0.01, and 0.001. Tests in this papar, all the significance criterion is set 0.05. A type II error occurs when one rejects the alternative hypothesis (fails to reject the null hypothesis) when the alternative hypothesis is true.

In our test power means the ability of a test to detect an effect, the size of a test denotes the level when all means are equal.

1.2.2 Simulation of Power Function

In the simulation study, the power function which we are interested in under certain conditions is obtained via so called Monte Carlo method. The generic term 'Monte Carlo method' is used for all numerical methods using sampling of probability distribution. A typical Monte Carlo simulation procedure of the test involves the following:

- Generate M independent data sets under conditions of interest
- Compute the numerical value of the test statistic T for each data set $\Rightarrow T_1, \ldots, T_M$
- To evaluate the test statistics of each data set and calculate proportion of rejections of H_0

1.3 Outline

In the chapter 2 theoretical background of the classical ANOVA F test will be briefly introduced. How the test is affected by violation of underlying assumptions will be presented. ANOVA F is known to be robust under certain violated conditions but the problem can be very serious under the combined violations of assumptions. In such case, classical ANOVA F either should not be used or modifications are needed to overcome drawbacks.

In the chapter 3 linear model is estabilished for the test and the likelihood ratio(LR) test is introduced as one of the alternative test to ANOVA F test. Comparisons of power function with ANOVA test are discussed.

In the chapter 4 modifications for the classical ANOVA F are proposed. Kaplan-Meier and Satterswaite methods are suggested to compensate the censoring and heterogeneity of variance respectively. Box-Cox transformation is also applied to pre-process data. Resampling methods are suggested for obtaining a reliable critical value.

In the chapter 5 Welch test is proposed, which is suggested as an alternative to ANOVA F test when under heterogeneous variance. This approach will be examined with respect to the power function whether it will be alternative to ANOVA F test when more violations are imposed.

In each chapter, simulation studies of corresponding power function are carried out, then performances are analyzed and conclusions are drawn. Some figures of power functions are chosen to illustrate some points made in the paper.

Chapter 2

Classical ANOVA

In this chapter firstly general principles of classical ANOVA F is presented and widely used one-way classification is described. Then simulation studies of power for ill-conditions show that how these violations are detrimental to test.

2.1 Analysis of Variance

Classical ANOVA F test is the most often used statistical method today for comparing the effects of fixed groups. During our study fixed-effects model is chose which assumes that the data come from normal populations which may differ only in their means. The assumptions made by the model are:

1. Normality, the distributions in each group are normal
2. Independence and no censorship
3. Homogeneity of variance

In practice, there are several types of ANOVA depending on the number of treatments and the way they are applied to the subjects in the experiment: One-way classification and Two or more-way classification. One-way classification will be present in the paper.

2.1.1 One Way Classification

In the experiment, there is only one factor, and the analysis of variance that we will be using to analyze the effect of this factor is called a one-way or one-factor ANOVA. Let Y_{ij} be the jth sample observation ($j = 1, 2, ..., n_p$)on ith group ($i = 1, 2, \ldots, p$), so we have following array:

Population 1:	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$	$\bar{Y}_{1.}$
Population 2:	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$	$\bar{Y}_{2.}$
	.....	
Population p:	$Y_{p1}, Y_{p2}, \ldots, Y_{pn_p}$	$\bar{Y}_{p.}$

The goal of analysis of variance is to test the hypothesis that the means of the I-groups are equal. The p groups means may be designated μ_1, μ_2,..., μ_p.The hypothesis will be

$$H_0 : \qquad \mu_1 = \mu_2 = \cdots = \mu_p$$

It's easy to model all of this with an equation of the form:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{2.1}$$

where μ is grand mean, α_i the level effect from ith group (the deviation of each level mean from the grand mean), $\sum n_i \alpha_i = 0$. ε_{ij} is residuals.

By the assumptions made for model, Y_{ij} is an independent observation from a normally distributed population whose mean is $\mu + \alpha_i$ and variance is denoted by σ^2. This can be written:

$$Y_{ij} \sim \mathcal{N}(\mu + \alpha_i, \sigma^2) \qquad i = 1, \ldots, p; \qquad j = 1, \ldots, n_i,$$

Then implied by the normal distribution of Y_{ij}, the residual will also following normal distribution:

$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2) \qquad i = 1, \ldots, p; \qquad j = 1, \ldots, n,$$

Estimation of Parameters

One can estimate the values of μ, α_i and ε from the given data set. Parameters are estimated by unbiased estimation (expected values equal the parameters being estimated). The estimates are given by

$$\begin{aligned} \bar{Y}_{i.} &= \sum_{j=1}^{n_i} Y_{ij} \\ \bar{Y}_{..} &= \frac{\sum_{i,j} Y_{ij}}{\sum_{i=1}^{p} n_i} \\ s^2 &= \frac{\sum_{i=1}^{p} (\bar{Y}_{i.} - \bar{Y}_{..})^2}{p-1} \end{aligned}$$

The list of the parameters of the problem and their point estimates finding in table(2.1).

Parameter	Estimate
μ	$\bar{Y}_{..}$
α_i	$\bar{Y}_{i.} - \bar{Y}_{..}$
$\mu + \alpha_i$	$\bar{Y}_{i.}$
σ^2	s^2

Table 2.1: The list of Parameters

The Partition of Sums of Squares

In general, the total variance is partitioned into the component that is due to true error (i.e.,within treatment) and the components that are due to differences between treatments.

$$\mathrm{SS}_{\mathrm{Tot}} = \mathrm{SS}_{\mathrm{Betw}} + \mathrm{SS}_{\mathrm{Wi}}$$

When the unknown parameters in $Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ are replaced by their estimates from the data, we have

$$Y_{ij} - \bar{Y}_{\cdot\cdot} = (\bar{Y}_i - \bar{Y}_{\cdot\cdot}) + (Y_{ij} - \bar{Y}_{i\cdot})$$

Square the deviations and sum them over all scores. It can be proved algebraically that the cross-product terms always add to zero when summed over an entire set of data.

$$\sum_{i=1}^{p}\sum_{j}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 + \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$$

Thus we have

$$\mathrm{SS}_{\mathrm{Tot}} = \sum_{i=1}^{p}\sum_{j}^{n_i}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$$

$$\mathrm{SS}_{\mathrm{Betw}} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

$$\mathrm{SS}_{\mathrm{Wi}} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\cdot})^2$$

F Test

we need to test on a comparison of the variance due to difference between treatments variability (called mean Square due to treatment, or $\mathrm{MS}_{\mathrm{Betw}}$) with the within-treatment variability (called Mean Square Error, or $\mathrm{MS}_{\mathrm{Wi}}$). In one-way ANOVA, statistical significance is tested for by comparing ANOVA F test statistic.

$$F = \frac{\mathrm{MS}_{\mathrm{Betw}}/\sigma^2}{\mathrm{MS}_{\mathrm{Wi}}/\sigma^2} = \frac{\mathrm{MS}_{\mathrm{Betw}}}{\mathrm{MS}_{\mathrm{Wi}}} \tag{2.2}$$

Where:

$\mathrm{MS}_{\mathrm{Betw}} = \frac{\mathrm{SS}_{\mathrm{Betw}}}{p-1}$, p is number of treatments
$\mathrm{MS}_{\mathrm{Wi}} = \frac{\mathrm{SS}_{\mathrm{Wi}}}{p(n_i-1)}$, n_i is the number of each group

F distribution with $p-1, p(n_i-1)$ degrees of freedom. F distribution is the test statistic which is the quotient of two mean sum of squares which have a chi square distribution. When H_0 is true, the quantity $(p-1)\mathrm{MS}_{\mathrm{Betw}}/\sigma^2$ has a χ^2 distribution with $p-1$ degree

of freedom, for under the null hypothesis, MS_{Wi} is simply a variance calculated from a sample of size p to estimate σ^2. And for the quantity $p(n_i-1)s^2/\sigma^2$, we know already that it has a χ^2 distribution with $p(n_i-1)$ degree of freedom.

F test is usually made as two-side test. Since under the null hypothesis, the variance estimated based on within-treatment variability should be about the same as the variance due to between-treatments variability. Otherwise in the case where the groups are clearly different, the variance estimated based on within-treatment variability should be larger than the variance due to between-treatments variability. Therefore F test in ANOVA is one-side test since the values of F larger than $F[1-\alpha; p-1, p(n_i-1)]$ as the critical region.

In the simulation study, one way ANOVA F test will be written as ANOVA(F) test. The former ANOVA indicates the test statistics and the F in the bracket denotes a critical value coming from F distribution.

2.2 Simulation Study: ANOVA with Violated Conditions

First, in the simulation study two sample sets with equal sample size 10 is compared. The odd group with $\mu_1 = 20$, the mean of even group is changing $\mu_2 = \mu_1 + v$, $v \in [-10, 10]$. The large standard deviation $\sigma = 5$ is chosen for both group when under the homogenous case and one group is changing to $\sigma = 7.5$ under the heterogeneous case. When considering about the censoring case, in our design is set Type I censoring, that is, every censored data has the same censoring value. If we take the censoring point 23 and 25, which is approximately 27% and 16% data set censored all null distribution for both normal and Weibull distribution situations. When v is becoming large, that is, one group mean is moving to the right side, the percentage of censoring of this group is increasing. When under the Weibull distribution, parameters k and λ are needed for generating data set. From the equations (1.6) and (1.7), one can get parameters k and λ from known mean and variance.

The simulation of power function is obtained by considering the hypothesis $H_0 : \mu_1 = 20$ against $H_1 : \mu_2 \neq 20$. Power simulations[1] are run in different situations which are listed in Table (2.2)

The output of the simulation study has put into graphical form which can make conclusions fairly clear. Not all the figures in the simulation studies are showed up and discussed here, some figures included are chosen to illustrate some points made in the paper. Each figure represents a power function curve under different situation with reference of the power curve without any violation of assumptions (ANOVA standard), in order to have a clear view how violations affect the power of the test.

[1] M = 2.000 in each simulation

For design No.1 only with censoring, one can see the size of the hypothesis (the power

Table 2.2: Designs used in simulation

Design No.	Distribution	Censoring Percentage at null point	Difference of Deviation
1	N	16%	0
2	N	0	2.5
3	N	16%	2.5
4	W	0	0
5	W	16%	0
6	W	0	2.5
7	W	16%	2.5

when the means are equal) is kept at significant level and the performance of power function is not so influnced when the percentage of censoring is small (at null hypothesis, 16% censored). But for the high percent censoring, showed from far right side the power function deviates from the original. It can be stated that ANOVA(F) is robust against censoring when the percentage of censored data in the sample size is small, and the power of test is affected by high percentage censoring. The figure for violation of normality in design No.4 shows the curve is well fit as the standard one (no violated condition) except a little diviation around far left side. This happens due to the reason that the skewness and heavy tail of Weibull distribution. The good performance of power in this situation clearly indicates that the ANOVA(F) is insensitive to the non-normality if the skewness is not serious. In design No.2 under the heterogeneity of variance, i.e., the type I error is not inflated but the power is notably lower than the standard one. It's widely accepted that the size of ANOVA(F) does not dramatically change due to moderate departures from the assumption of homogeneity, while the lost power for rejecting the alternative hypothesis can not be ignored. This indicates ANOVA(F) is sensitive to the heterogeneity of variance. In design No.3 and No.7 show the combination of heterogeneity and censoring not only accompanied the loss of power but also the shifting minimal point to the left side, the type I error is inflated. It happens because the sample set cannot keep balanced with these two violations, much information is missing from the right side and in such case ANOVA(F) fails to recognize data situation. Curve in No.7 shows with the absence of non-normality, the shifting is more serious, for the skewness of non-normality itself makes the unbalanced data set even worse.

One purpose of examining the performance of power of the classical ANOVA(F) under different violated assumptions shows that these violations are detrimental to tests. At the same time it also verifies that ANOVA(F) is robust against small percent censoring data, insensitive to moderate violation of normality. These facts make data analyzing technique attractive because information in survival data is often lacking concerning these two assumptions. The drawbacks of ANOVA(F) with hight percent censoring, it loses its power

and under the case of heterogeneity of variance, the whole power is notably low. It has been studied that the size of test will be more libera if variances are negatively correlated with the sample sizes (Krutchkoff 1988). The shifting minimal point appears when under the combination of violations. In such situations the test statistics should be replaced or modified to overcome drawbacks if we still want to use ANOVA(F). Next chapter we are going to discuss the test in the linear model and likelihood ratio test is proposed as one of the alternative test.

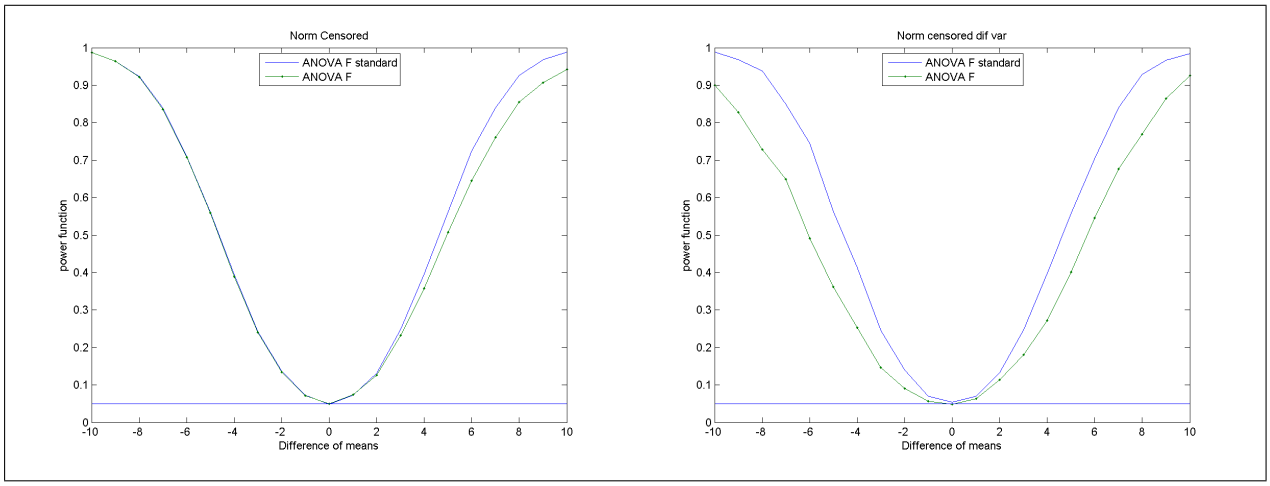

Figure 2.1: Design No.1 and Design No.2

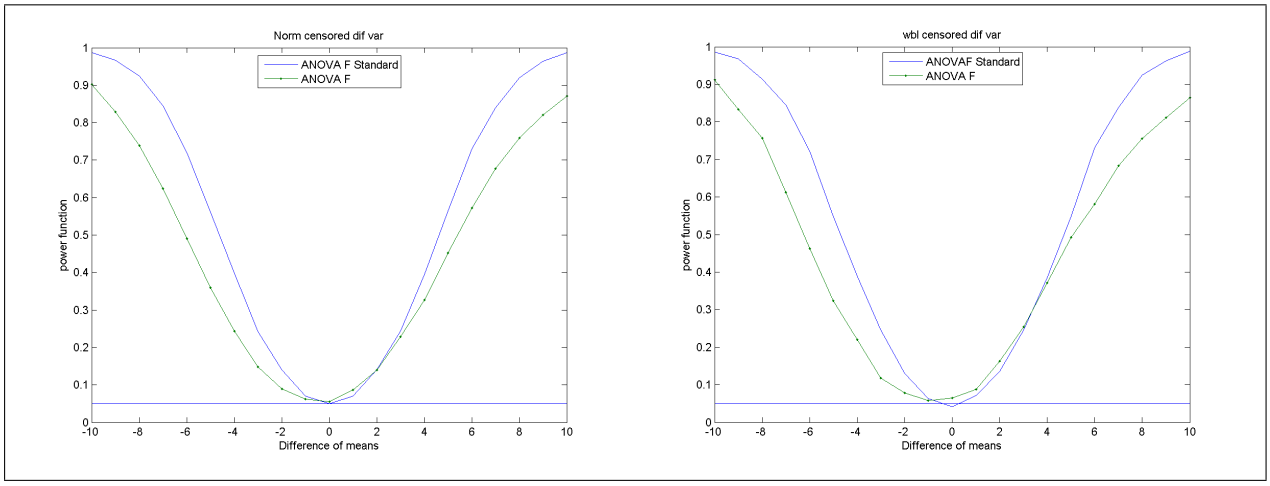

Figure 2.2: Design No.3 and Design No.7

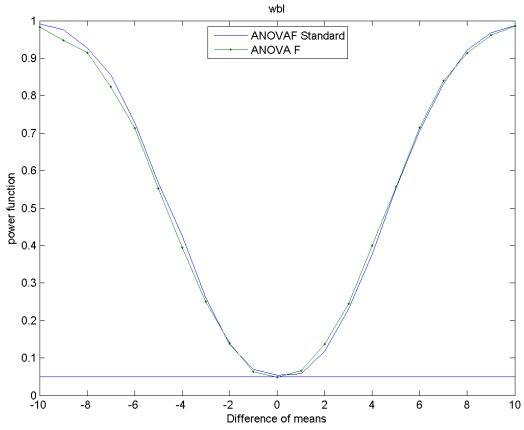

Figure 2.3: Design No.4

Chapter 3

Linear Models for ANOVA

The primitive idea to propose ANOVA F test in linear model since its convenience in computation. When maximum likelihood estimate is introduced and it provides extreme convenience for likelihood ratio test(LR). Therefore LR test then is introduced as one of the alternative test. One merit of LR test is for censored data, since the estimate can be easily corrected by changing likelihood functions. The simulation studies, comparisons with the ANOVA F are presented for giving intuitively justification.

3.1 The Linear Model

The classical linear regression model has the form:

$$y = \mathrm{X}\beta + \varepsilon \tag{3.1}$$

where Y is an $n \times 1$ column vector of observed values of random variables, the $n \times p$ matrix X is called the *regression matrix* of rank p. In the experimental design situations the elements of X are chosen to be 0 or 1, in this case X is commonly called *design matrix*. β is a $p \times 1$ vector of unknown parameters, and ε is an $n \times 1$ vector of "errors",assumed to be normally distributed with mean zero and variance is σ^2, that is $\varepsilon \sim N(0, \sigma^2 I)$.

3.1.1 Linear Model for One-way Classification

We are going to specify our linear model for dealing with one way classification for simplicity which means we investigate the effect of a single factor. Let Y_{ij} be the jth sample observation$(j = 1, 2, ..., p)$on ith group, so we have following array:

Population 1:	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$
Population 2:	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$
.....	
Population p:	$Y_{p1}, Y_{p2}, \ldots, Y_{pn_p}$

In order to apply the general linear theory, we combine above information into the single model:

$$y_{ij} = \mu_i + \varepsilon_{ij}$$

where, μ_i is mean of each group, ε_{ij} is residuals due to each observations. Then we introduce the design matrix X which only contains ones and zeros as presenting in the general linear model, see the following form:

$$\begin{pmatrix} y_{11} \\ y_{12} \\ \cdots \\ y_{1n_1} \\ \hline y_{21} \\ y_{22} \\ \cdots \\ y_{2n_2} \\ \hline \cdots \\ \hline y_{p1} \\ y_{p2} \\ \cdots \\ y_{pn_p} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & 0 & 0 & \cdots & 0 \\ \hline 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & \cdots & 0 \\ \hline \cdots & \cdots & \cdots & \cdots & \cdots \\ \hline 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_p \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \cdots \\ \varepsilon_{1n_1} \\ \hline \varepsilon_{21} \\ \varepsilon_{22} \\ \cdots \\ \varepsilon_{2n_2} \\ \hline \cdots \\ \hline \varepsilon_{p1} \\ \varepsilon_{p2} \\ \cdots \\ \varepsilon_{pn_p} \end{pmatrix} \tag{3.2}$$

Or in general form as

$$Y = \mathrm{X}\beta + \varepsilon \tag{3.3}$$

Where $\beta = [(\mu_i)]$ is a $p \times 1$ matrix, and n is the total numbers of observations, just simply adding all the group members, X is a $n \times p$ design matrix. The null hypothesis of interest is H: $\mu_1 = \mu_2 = \cdots = \mu_p$, or $\mu_1 - \mu_p = \mu_2 - \mu_p = \cdots = \mu_{p-1} - \mu_p = 0$ now can be written in the form of linear hypotheses as well in the matrix:

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & 0 & \cdots & 0 & -1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_p \end{pmatrix} = 0 \tag{3.4}$$

i.e $\mathrm{K}\beta = 0$. Here the rows of K are linearly independent so that K is $r \times p$ of rank r, where $r = p - 1$.

After constructing the linear model and setting the hypothesis, the next problem is proposed, how to estimate the vector of unknown parameter β. One of the popular ways is called Least Square estimate.

Least Square Estimate Method

In linear model the most efficient method of obtaining an estimate of β is called least square method. This approach is known to minimize the sum of squares of the residuals with respect to β.

$$\hat{\beta} = \arg\min \varepsilon^{\mathrm{T}}\varepsilon = \arg\min (y - \mathrm{X}\beta)^{T}(y - \mathrm{X}\beta) \tag{3.5}$$

Here $\hat{\beta}$ is called least squares estimate of β. We note that $\hat{\beta}$ can be obtained by writing:

$$\begin{aligned} \varepsilon^{\mathrm{T}}\varepsilon &= (Y - \mathrm{X}\beta)^{T}(Y - \mathrm{X}\beta) \\ &= \mathrm{Y}^{\mathrm{T}}\mathrm{Y} - 2\beta^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{Y} + \beta^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{X}\beta \end{aligned}$$

then differentiating $\varepsilon^{\mathrm{T}}\varepsilon$ with respect to β:

$$\frac{\partial \varepsilon^{\mathrm{T}}\varepsilon}{\partial \beta} = 0 \tag{3.6}$$

From the solution of equation (3.6), we get the least square estimate(LS):

$$\hat{\beta} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y} \tag{3.7}$$

3.2 Maximum Likelihood Estimate

LS shows its unbiased properties and works properly under normality assumption of error terms. If the assumptions are violated, we need to find a more robust method to get estimates. Actually people prefers using maximum likelihood estimate(MLE) which owns some nice properties and can also provide good estimation even for censored data which least square method could not achieve.

3.2.1 Theoretical Background

The basic idea underlying MLE is quite simple. Usually, when specifying a probability density function(pdf), we treat the pdf as a function of the value of the random variable with the distribution parameters θ assumed to be known. The MLE of the unknown parameters θ, is the values to maximize the likelihood function $L(\theta|X)$ given the observed vector of data set. It is usually easier to find the maximum of a likelihood function by first taking its log and working with the resulting log-likelihood

$$\ell(\theta|X) = \ln L(\theta|X)$$

Since the natural log is a monotonic function, likelihood function ℓ has the same maxima as log likelihood function, so that the maximum of likelihood function also corresponds to the maximum of the likelihood function. The score S of a likelihood function is the first derivative of log likelihood function with respect to the parameters

$$S(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \begin{pmatrix} \partial \ell(\theta)/\theta_1 \\ \vdots \\ \partial \ell(\theta)/\theta_n \end{pmatrix} \tag{3.8}$$

for a vector of n parameters. From elementary calculus it follows that the score evaluated at the MLE is zero, $S(\theta) = 0$. This provides one approach for obtaining MLEs.

3.2.2 MLE for Uncensored Case

For n independent uncensored samples, $X = (x_1 \cdots, x_i)$ from a continuous random variable, we can likewise define the likelihood function as the product of the individually density function:

$$L(X; \theta) = \prod_{i=1}^{n} f(x_i; \theta) \tag{3.9}$$

Then take the log likelihood function will become the summation form

$$\ell(\theta; X) = \sum_{i=1}^{n} \ln[L(\theta; x_i)] \tag{3.10}$$

which has the score function to be set zeros

$$S(\theta) = \frac{\partial \ell(\theta; X)}{\partial \theta} = 0 \tag{3.11}$$

Example. MLE for Normal Uncensored Case

Since the components of the vector of Y are normally distributed, the likelihood function from our linear model then is:

$$L = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2} S\right) \tag{3.12}$$

where:

$$S = (Y - \mathrm{X}\beta)^T (Y - \mathrm{X}\beta) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} \mathrm{x_{ij}} \beta_\mathrm{j})^2 \tag{3.13}$$

Then the log-likelihood function will be:

$$l(\beta, \sigma^2; Y) = \log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\mathrm{x_{ij}}\beta_\mathrm{j})^2 \tag{3.14}$$

Differentiating with respect to β_j (j=1,. . . ,p) and σ^2 can one we get the ML-estimates of $\hat{\beta_j}$ and $\hat{\sigma^2}$.

$$\frac{\partial l}{\partial \beta_j} = 0 \tag{3.15}$$

$$\frac{\partial l}{\partial \sigma^2} = 0 \tag{3.16}$$

From equations (3.15) and (3.16) so setting this set of p equations to zero and solving for β and σ gives:

$$\hat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \tag{3.17}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\hat{\beta}_j)^2 \tag{3.18}$$

MLE of Uncensored Normal = Least Square Estimate

Compare the results of the MLE and LSE from equation (3.7) and (3.17), under the assumption of the normal error, method of maximum likelihood method is equivalent to the method of least square method. In such case LSE are usually use for its computational convenience. When the commas underlying distribution of the ϵ_i is not normal, the LSE of β_j is not the same as ML estimate, and we prefer to use MLE for its accuracy and good properties for the estimation.

3.2.3 MLE for Censored Case

In our experiments we are dealing with the survival data, some of which are always censored. So we are discussing about how to get the MLE censored case. The maximum likelihood method is, at this point, by far is one of the most appropriate analysis method for censored data. When performing maximum likelihood analysis, the likelihood function needs to be expanded to take into account the suspended items. We have a Likelihood for each type:

$$L(x_i) = f(x_i) \qquad \text{if not censored} \tag{3.19}$$

$$L(x_i) = Pr(T > x_i) = 1 - F(x_i) \qquad \text{if censored} \tag{3.20}$$

The full Likelihood is composed of the product of these terms. Consider a censor indicator variable w_i defined as:

$$\begin{aligned} w_i &= 1 \qquad \text{if } t_i \text{ is not censored} \\ w_i &= 0 \qquad \text{if } t_i\text{is censored} \end{aligned}$$

Then the likelihood function can be written:

$$L(X;\theta) = \prod_i (f(x_i;\theta))^{w_i}(1 - F(x_i;\theta))^{(1-w_i)} \tag{3.21}$$

where the two products are taken over uncensored and censored subjects respectively. The log likelihood is:

$$\ell(\theta) = \sum_u \log f(x_i;\theta) + \sum_c \log(1 - F(x_i;\theta)) \tag{3.22}$$

Where u and c denote uncensored and censored situation respectively. Then take the derivative of log likelihood function to get unknown parameters which maximize the the functions.

Example. MLE for Weibull Censored Case

Now consider maximum likelihood estimation for weilbull distribution in censored case. The cumulative probability function of weibull distribution is:

$$F(x;k,\lambda) = 1 - e^{-(x/\lambda)^k} \tag{3.23}$$

The likelihood function for censored data is defined as (3.19), in our case it will be:

$$1 - F(x;k,\lambda) = e^{-(x/\lambda)^k}$$

Following the procedure forming of likelihood function and log likelihood function from equations (3.21) and (3.22). The full log likelihood function of weibull distribution with censored data will be

$$\ell = d\log k - kd\log\lambda + (k-1)\sum_u \log x_i - \left(\frac{1}{\lambda}\right)^k \sum {x_i}^k \tag{3.24}$$

where d is the number of uncensored data, $\sum_u x_i$ denotes the sum of uncensored data. The derivatives of the log likelihood function are:

$$\begin{aligned} \frac{\partial \ell}{\partial \lambda} &= k\lambda d - \frac{k}{\lambda^{k-1}} \sum {x_i}^k \\ \frac{\partial \ell}{\partial k} &= \frac{d}{k} - d\log\lambda + \sum_u \log x_i - \frac{1}{\lambda^k} \sum {x_i}^k \log\frac{x_i}{\lambda} \end{aligned} \tag{3.25}$$

Thus the ML estimators of $\hat{\lambda}$ and $\hat{k}$ can be found explicitly by solving the log likelihoods equal to zero.

$$\hat{\lambda}^k = \frac{\sum {x_i}^k}{d} \tag{3.26}$$

and the parameter k can get by solving the form $\frac{d}{k} + \sum_u \log x_i - d\frac{\sum {x_i}^k \log x_i}{\sum {x_i}^k} = 0$ iteratively.

3.3 Likelihood Ratio Test

Maximum likelihood provides for extremely convenient tests of hypothesis in the form of likelihood Ratio test(LR), or Chi square test. This test statistics that examine whether a reduced model provides the same fit as a full model. The likelihood-ratio test statistic is given by:

$$\mathrm{LR} = -2\ln\lambda = -2\ln(\frac{L(\hat{\theta}_0|X)}{L(\theta_N|X)}) = -2[\ell(\hat{\theta}_0|X) - \ell(\theta_N|X)] \tag{3.27}$$

Where $L(\theta|X)$ denotes the likelihood function, $\ell(\theta|X)$ is log-likelihood function. $L(\theta_0)$ is the likelihood evaluated at the MLE subject to the null hypothesis and θ_N maximize likelihood function under the whole model. For sufficiently large sample size, the LR test statistics is χ^2_{d-c} distributed, a χ^2 with $d-c$ degrees of freedom (Wald 1943), d is the dimension of whole parameter space, and c is the dimension of hypothesis space, d is always larger than c.

3.3.1 LR Tests for Normal Case

Consider the general linear model, $y = \mathrm{X}\beta + \varepsilon$, where we assume that the $n \times 1$ vector of residual errors ε is multivariate normal, with mean vector zero and covariance matrix V, V is a diagonal matrix whose ith elements is the variance of the ith mean. Writing the vector of residual as $\varepsilon = y - \mathrm{X}\beta$ gives the resulting likelihood for V and β, as

$$L(\beta, \sigma^2|y, X) = (2\pi)^{-n/2}(|V|)^{-1/2}\exp(-\frac{1}{2}(y-\mathrm{X}\beta)^T V^{-1}(y-\mathrm{X}\beta))$$

which has log-likelihood

$$\ell = \ln L = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|V| - \frac{1}{2}(y-\mathrm{X}\beta)^T V^{-1}(y-\mathrm{X}\beta)$$

Here β is a vector of fixed effects and matrix V is a function of variance components. Thus, the parameters to be estimated are the vector β of fixed effects and the p variance, σ_i^2.

Suppose we wish to compare the relative fit of two models that assume the same covariance

structure V, but have different vectors of fixed effects, a vector $\hat{\beta}$ for the full model and a vector $\hat{\beta}_H$ subjects to the constraints. The resulting likelihood ratio test statistic is

$$\begin{aligned} -2\ln\lambda &= -2[\ell(\hat{\beta}_H) - \ell(\hat{\beta})] \\ &= [(y - \mathrm{X}\hat{\beta})^T\hat{V}^{-1}(y - \mathrm{X}\hat{\beta}) - (y - \mathrm{X}\hat{\beta}_\mathrm{H})^T\hat{V}^{-1}(y - \mathrm{X}\hat{\beta}_\mathrm{H})] \end{aligned} \tag{3.28}$$

For large sample sizes, this test statistic follows a χ^2 distribution with $d - c$ degrees of freedom, where d and c are the degrees of freedom for full and reduced models respectively.

LR Test for Normal Heteroscedastic

If the means of p different populations with different variances. V is a diagonal matrix whose ith element is the variance of ith mean. Denoting the variance of the ith mean by $\mathrm{Var}(i)$, under the the quadratic product in the LR test reduces to

$$(y - \mathrm{X}\hat{\beta})^T\hat{V}^{-1}(y - \mathrm{X}\hat{\beta}) = \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{\mathrm{Var}(i)}$$

Hence,recalling equation (3.28) the LR test for a normal heteroscedastic is given by:

$$LR = -2\ln\ell = \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{\mathrm{Var}(i)} - \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{\mathrm{Var}(i)} \tag{3.29}$$

Where $\hat{y}_i = \mathrm{X}\hat{\beta}$ is the value estimated under the null hypothesis.

3.3.2 LR Test for Nonnormal Distribution

We also build LR test for nonnormal distribution which is represented as weibull distribution in our design. Unlike the normal distribution parametrized with mean and variance, Weibull distribution is equipped with two parameters, one scale parameter λ and the shape parameter k. The likelihood function for weibull function is

$$L(k,\lambda) = \prod_{i=1}^{n}\frac{k_i}{\lambda_i}\left(\frac{x_i}{\lambda_i}\right)^{k_i-1}e^{-(x_i/\lambda_i)_i^k}$$

The likelihood functions under the null hypothesis and the whole model are:

$$\begin{aligned} L(\hat{\theta}_H) &= \max_{\theta\in\Theta_H}L(\lambda,k) \\ L(\hat{\theta}) &= \max_{\theta\in\Theta}L(\lambda,k) \end{aligned}$$

Where $\hat{\theta}_H$ is the estimated parameter space under the null hypothesis and θ is under the whole model. For the test one of the assumption is homogeneity of variance, this condition is satisfied in weibull cse via the relation of parameters and variance $\mathrm{Var}(X) = \lambda^2[\Gamma(1+\frac{2}{k}) - \Gamma^2(1+\frac{1}{k})]$, which we impose equivalent from each group when calculating parameters. However under the heterogeneous case it is not necessary to keep this constraint.

Then the likelihood-ratio is given by

$$LR = -2\ln\left(\frac{L(\theta_H)}{\max_{\theta\in\Theta} L(\theta|X)}\right) = -2\ln\left(\frac{L(\hat{\theta}_H)}{L(\hat{\theta})}\right) \tag{3.30}$$

3.4 Simulation Study

While new models have been successfully established then the power of LR(chi) test need to be examined by the simulation study. As usual two sample groups are compared in the test. The procedure is exactly the same as described in the last chapter. The designs for the simulation[1] is following. In each figure, ANOVA(F) is also presented to give the direct comparison for these two methods.

Table 3.1: Designs used in simulation

Design No.	Distribution	Censoring Percentage at null point	Difference of Deviation
1	N	16%	0
2	N	0	2.5
3	N	16%	2.5
4	W	0	0
5	W	16%	0
6	W	0	2.5
7	W	16%	5

From power curve from design No.1, LR(chi) test gives better result for the censoring case comparing to ANOVA(F) one the right side, but one can notice that the whole power is lifted including the type first error based on the observation that there is notable space between these two curves. It's not a surprising result since the chi square distribution of LR test needs to be guaranteed by large sample number which can not be satisfied by our small sample size ($n = 10$). But under the heterogeneity the LR(chi) test loses its power comparatively to ANOVA(F) test. One explanation is, LR(chi) test approaches chi square distribution for large sample size, when we have limited samples in each group(10 observations). The violation to chi square distribution is more severe accompanied with heterogeneity. The critical value for chi square distribution is also not adequate to apply. Design

[1] M=1.000 in each simulation

No.4 figure indicates that the LR(chi) test is sensitive to non normality than ANOVA(F) test. Tests for cases that combined with violated normality could not be trusted anymore. For Design No.3 which combined the censorship and heterogeneity of variance, LR(chi) is more suitable with censoring but not powerful for the heterogeneity. The effect of combined violations shows clearly in the graph, that even though the power of both side are not as good as ANOVA(F) test for the shifted minimal point, LR(chi) test can almost keep the minimal point at null distribution. In this case the type I error is not inflated at this points.

For LR(chi) test in the linear model, the test can be beneficial when censored data is presented. But for most violated situations testing in linear model are not as good as ANOVA(F) test, especially for the violation of non-normality assumption. With the asymptotical chi square distribution property, both LR(chi) test and its critical values are suspicious to trust. Comparing with the easy-handling ANOVA(F) test the procedure of LR(chi) test is too costly, and crucially it does not have an improved performance. Therefore LR is not an ideal test under such violations of assumptions.

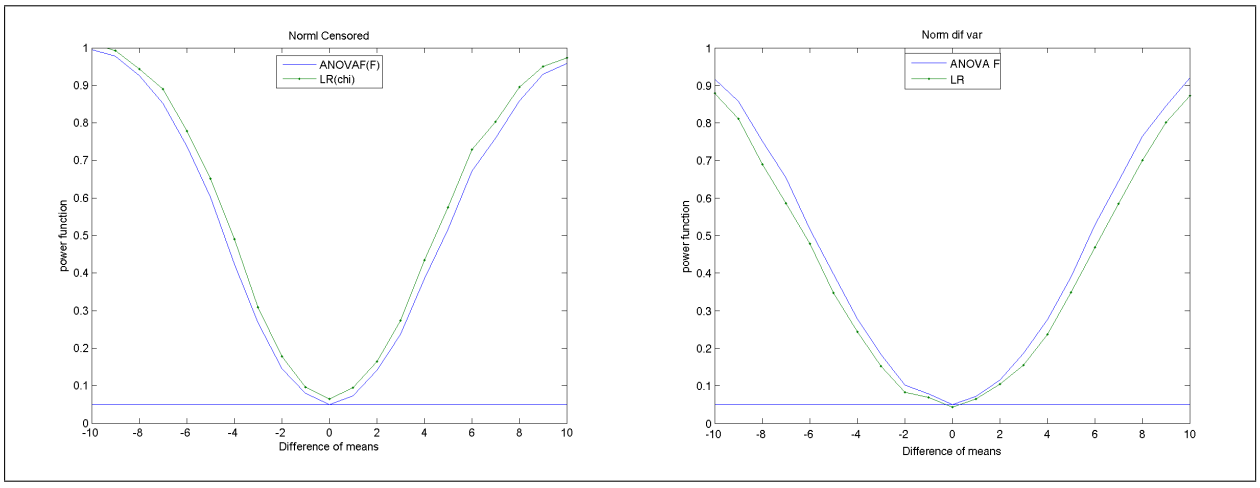

Figure 3.1: Design No.1 and Design No.2

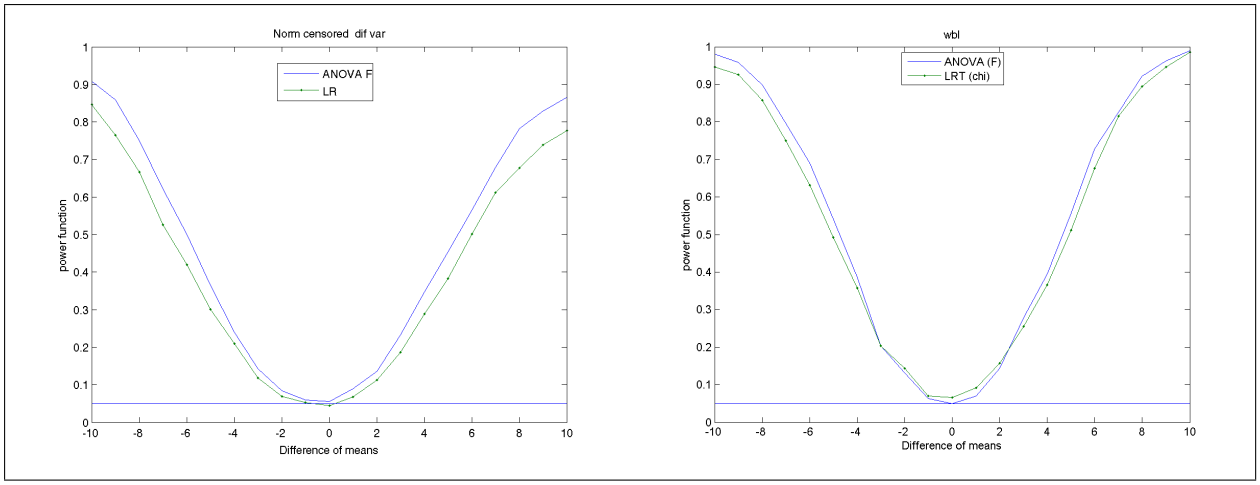

Figure 3.2: Design No.3 and Design No.4

Chapter 4

Modifications of ANOVA

Since we have compared the power functions of the likelihood ratio test from the linear model with ANOVA F test, it turns out that ANOVA F performs more robustly than likelihood ratio test from the linear model under the violations of assumptions in most cases. Therefore we are concentrating on the modifications of ANOVA F to get a more robust and powerful testing procedure based on it. Kaplan Meier and Satterthwaite methods are proposed for compensating the case of censored data and the case of heteroscedasticity respectively. Box-Cox transformation is used to pre-process the data for filtering purpose. At last Resampling methods will be investigated for getting empirical critical value for the test.

4.1 Kaplan Meier Method

Since we are dealing the survival data which are always subject to the censoring. Kaplan Meier method actually is a method of estimating survival functions for censored data. It's a nonparametric method which is more efficient when no suitable theoretical distribution is known. Now we are going to use nonparametric methods to analyze survival data instead of attempting to fit a theoretical distribution.

4.1.1 Basic Definition of Survival Functions

Let T denote the survival time. The distribution of T can be characterized by the following functions.

1. Survival function. This function denoted by $S(t)$, is defined as the probability that an individual survives longer than t:

$$\begin{aligned} S(t) &= P(\text{an individual survives longer than t}) \\ &= P(T > t) \end{aligned} \tag{4.1}$$

From the definition of the cumulative distribution function $F(t)$ of T, survival function can be written equivalently:

$$\begin{aligned} S(t) &= 1 - P(\text{an individual fails before time t}) \\ &= 1 - F(t) \end{aligned} \tag{4.2}$$

Here $S(t)$ is a nonincreasing function of time t with the properties

$$\begin{aligned} S(t) = 1 \quad &\text{for} \quad t = 0 \\ S(t) = 0 \quad &\text{for} \quad t = \infty \end{aligned}$$

2. Hazard Function. The hazard function $h(t)$ of survival time T gives the conditional failure rate. This is defined as the probability of failure during a very small time interval, assuming that the individual has survived to the beginning of the interval, or as the limit of the probability that an individual fails in a very short interval, t to $\triangle t$ per unit time, defined by

$$h(t) = \lim_{\triangle t \to 0} \frac{pr(t < T < t + \triangle t | t < T)}{\triangle t} \tag{4.3}$$

The hazard function can also be defined in terms of the cumulative distribution function $F(t)$ and the probability density $f(t)$

$$h(t) = \frac{f(t)}{1 - F(t)} \tag{4.4}$$

4.1.2 Relationship of The Survival Function

From the definition above, one can get the relationships of the survival functions, from (4.2) and (4.4), the hazard function can be rewritten:

$$h(t) = \frac{f(t)}{S(t)} \tag{4.5}$$

Since the probability density function is the derivative of the cumulative distribution function,

$$f(t) = \frac{d}{dt}[1 - S(t)] = -S'(t) \tag{4.6}$$

Substituting (4.6) into (4.5) yields

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{\mathrm{dt}} \log_e S(t) \tag{4.7}$$

Therefore one can get survival function from integrating (4.6) from zero to t. Hence for the continuous distribution we have

$$S(t) = \exp[\int_0^t h(t)dx] \tag{4.8}$$

For discrete distributions, the survival function can be expressed by a direct application of the product law of probabilities, that

$$S(t) = \prod_j (1 - h_j) \tag{4.9}$$

Where h_j is the discrete hazard function at each point. If $f(t)$ is known, the survival function $S(t)$ can be obtained from the basic relationship between $f(t)$, $F(t)$. Thus given any one of $f(t)$, $F(t)$ or $S(t)$, the other two can be derived.

4.1.3 Kaplan Meier Estimator

Kaplan Meier estimator (also named product-limit) is a nonparametric estimation which does not require specification of the functional form of distribution. It has used to estimate the survivorship function. This method is applied if some of the data are censored. Now suppose the survival distribution is discrete with atoms $f_j(\theta)$ at preassigned points $a_j (a_1 < a_2 < ...)$. We assume that an individual censored at c could have been observed to fail at c. With this convention, the contribution to the likelihood from a subject observed to fail at a_j is f_j, and from a subject censored at c is

$$\text{pr}(T > c) = S(c+;\theta) = 1 - \Sigma_{j:a_j \leq c} f_j(\theta) \tag{4.10}$$

In term of the discrete hazard function $h_j(\theta)$ we have as in (4.9)

$$\begin{aligned} f_j(\theta) &= h_j(\theta) \prod_{k<j} [1 - h_k(\theta)] \\ S(c+;\theta) &= \prod_{j:a_j \leq c} [1 - h_j(\theta)] \end{aligned}$$

Each term is a product over the atoms a_j of the survival distribution. To derive the full likelihood from a sample of n observations, we first collect all the terms corresponding to the atom a_j. If there are d_j failures among the $r_j = r(a_j)$ individuals in view at a_j, the contribution to the total likelihood is

$$L = [h_j(\theta)]_j^d [1 - h_j(\theta)]^{r_j - d_j} \tag{4.11}$$

Where the two products are taken over uncensored and censored subjects respectively. The log likelihood is

$$\ell = \sum_j d_j \log h_j(\theta) + (r_j - d_j) \log[1 - h_j(\theta)] \tag{4.12}$$

The log likelihood (4.12) is exactly that for g independent binomials, with respectively r_j trials, d_j failures, and probability of failure h_j. It is particularly easy to maximize there as the parameter vector is h_j itself. Thus, $\hat{h}_j$ is the solution of

$$\frac{\partial l}{\partial h_j} = \frac{d_j}{h_j} - \frac{r_j - d_j}{1 - h_j} = 0$$

i.e $h_j = d_j/r_j$. We know a nonparametric estimator of the survivor function $S(t)$ can be expressed in terms of the discrete hazard function from (4.9)

$$\hat{S}(t) = \prod^{(t)}(1 - \frac{d_j}{r_j}) \tag{4.13}$$

The mean survival time μ can be shown to equal the area under the survival curve. The estimator is based upon the entire range of data. Note that some methods suggest using only the data up to the last observed event; Hosmer and Lemeshow (1999) point out that this biases the estimate of the mean downwards, and they recommend that the entire range of data should be used. To estimate μ, we have the formula:

$$\hat{\mu} = \int_0^\infty \hat{S}(t)dt \tag{4.14}$$

Thus, if the times to death are ordered as $t^{(1)} \leq t^{(2)} \leq \cdots \leq t^{(m)}$ (if there are m uncensored observations) and $t^{(m)}$ is the largest observation of n observations (i,e., $t^{(m)} = t_{(n)}$) when $t_{(n)}$ is an uncensored observation), then according to the equation (4.14), the mean survival time $\hat{\mu}$ can be expressed

$$\begin{aligned} \hat{\mu} &= 1.000t^{(1)} + \hat{S}(t^{(1)})(t^{(2)} - t^{(2)}) + \hat{S}(t^{(2)})(t^{(3)} - t^{(2)}) + \cdots \\ &+ \hat{S}(t^{(m-1)})(t^{(m)} - t^{(m-1)}) \end{aligned} \tag{4.15}$$

which is the sum of the areas of the rectangles under the survival curve formed by the uncensored observations.
The variance of μ is estimated by

$$\widehat{\mathrm{Var}(\mu)} = \sum_r \frac{A_r^2 d_j}{r_j(r_j - d_j)} \tag{4.16}$$

where A_r is the area under the curve $S(t)$ to the right of $t_{(r)}$. The kth A_r in terms of the m uncensored observations is

$$\begin{aligned} A_r &= \hat{S}(t^{(k)})(t^{(k+1)} - t^{(k)}) + \hat{S}(t^{(t+1)})(t^{(k+2)} - t^{(k+1)}) \\ &+ \cdots + \hat{S}(t^{(m-1)})(t^{(m)} - t^{(m-1)}) \end{aligned} \tag{4.17}$$

In order to get the unbiased estimate, Kaplan and Meier suggests that equation should (4.17) be multiplied by $m/(m-1)$, to correct the bias.

The Kaplan Meier method provides very useful estimates of mean and variance. It's the most widely used method in survival data analysis. Breslow and Crowley (1974) and Meier (1975b) have shown that under certain conditions, the estimate is consistent and asymptomatically normal.

Example. Kaplan Meier Method of Survival function S(t), Mean and Variance

Suppose the following remission duration are observed from 10 patients ($n = 10$) with solid tumors. Six patients relapse at 3.0, 6.5, 6.5, 10, 12, and 15 months; 1 patient is lost to follow-up at 8.4 months; and 3 patients arestill in remission at the end of the study after 4.0, 5.7 and 10 months. The calculation of $S(t)$ is shion in Table . Acording to the

Table 4.1: Calculation of the Kaplan Meier of S(t) for Data

a_j	r_j	d_j	$1 - \frac{d_j}{r_j}$	$\hat{S}(t)$
3.0	10	1	9/10	$9/10 = 0.900$
4.0+	-	-	-	-
5.7+	-	-	-	-
6.5	7	2	5/7	$9/10 \times 5/7 = 0.643$
8.4+	-	-	-	-
10.0	4	1	3/4	$9/10 \times 5/7 \times 3/4 = 0.482$
10.0+	-	-	-	-
12.0	2	1	1/2	$9/10 \times 5/7 \times 3/4 \times 1/2 = 0.241$
15	1	1	0	0

equations (4.14), the data in this example $m = 5$, $t^{(1)} = 3.0$, $t^{(2)} = 6.5$, $t^{(3)} = 10$, $t^{(4)} = 12$, $t^{(5)} = 15$, the mean survival time $\hat{\mu}$ can be expressed using (4.15) as

$$\begin{aligned} \hat{\mu} &= 1.000 \times 3.0 + 0.900(6.5 - 3.0) + 0.643(10 - 6.5) + 0.482(12 - 10) \\ &+ 0.241(15 - 12) \\ &= 10.088 \end{aligned}$$

From equations (4.17), one first computes the five A_r. The first A_r is

$$\begin{aligned} A_1 &= \hat{S}(t^{(1)})(t^{(2)} - t^{(1)}) + \hat{S}(t^{(2)})(t^{(3)} - t^{(2)}) + \cdots + \hat{S}(t^{(4)})(t^{(5)} - t^{(4)}) \\ &= 3.150 + 2.251 + 0.964 + 0.732 \\ &= 7.088 \end{aligned}$$

The second A_r is

$$\begin{aligned} A_4 &= \hat{S}(t^{(2)})(t^{(3)} - t^{(2)}) + \cdots + \hat{S}(t^{(5)})(t^{(6)} - t^{(5)}) \\ &= 2.251 + 0.964 + 0.732 \\ &= 3.938 \end{aligned}$$

The third, fourth, and fifth A_r's are, respectively,

$$\begin{aligned} A_7 &= 0.964 + 0.732 + 1.687 = 1.687 \\ A_9 &= 0.723 \end{aligned}$$

Thus according to the equation (4.16)

$$\widehat{\mathrm{Var}(\mu)} = \frac{(7.088)^2}{9 \times 10} + \frac{(3.938)^2 \times 2}{7 \times 5} + \frac{(1.687)^2}{3 \times 4} + \frac{(0.732)^2}{1 \times 2} = 1.942$$

The estimated $\widehat{\mathrm{Var}(\mu)}$ multiplies the factor $m/(m-1) = 5/4$ to correct the bias, these result is 2.330.

4.1.4 ANOVA F Test based on Kaplan Meier Method

In the classical F test of one way ANOVA, X_{ij}, $i = 1, \ldots, a$, $j = 1, \ldots, n_i$ denotes a double sequence of independent random variables, rewrite the formula of ANOVA F test in the following way:

$$\begin{aligned} \mathrm{MSE} &= \frac{1}{N-a} \sum_{i=1}^{a} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2 \\ &= \frac{1}{N-a} \sum_{i=1}^{a} (n_i - 1) S_i^2 \qquad (4.18) \\ \mathrm{MST} &= \frac{1}{a-1} \sum_{i=1}^{a} n_i (\bar{X}_{i.} - \bar{X}_{..})^2 \qquad (4.19) \end{aligned}$$

Where

$$\begin{aligned} \bar{X}_{i.} &= \frac{\Sigma_j X_{ij}}{n_i} \\ \bar{X}_{..} &= \frac{\Sigma_j \Sigma_i X_{ij}}{N} \\ S_i^2 &= (n-i)^{-1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2 \end{aligned}$$

$\bar{X}_{i.}$, $\bar{X}_{..}$ and S_i^2 are mean and sample variances. $N = n_1 + n_2 + \cdots + n_a$ is the total number of the samples.

In ANOVA F test in stead of taking sample means $\bar{X}_{i.}$, $\bar{X}_{..}$ and sample variances S_i as always used before, substituting $\bar{\mu}_i$, $\bar{\mu}$ and Var_i which come from Kaplan Meier method. The ANOVA F test based on the Kapan-Meier method will be

$$\begin{aligned} F &= \frac{\mathrm{MST}}{\mathrm{MSE}} \\ &= \frac{\frac{1}{a-1} \sum_{i=1}^{a} n_i (\bar{\mu}_i - \bar{\mu})^2}{\frac{1}{N-a} \sum_{i=1}^{a} (n_i - 1) \mathrm{Var}_i^2} \qquad (4.20) \end{aligned}$$

The grand mean, group means and sample variances are estimated from KM method which can give the accurate estimate for the censoring case. Another merit is that this method is nonparametric method which does not care about the distribution function of data set. The comparative results of identifying the test based on KM method estimator will be present in the simulation study.

4.2 Satterthwaite Method for Heteroscedasticity

When we check the classical F test of the one-way ANOVA under heteroscedasticity, it shows that the size of ANOVA F test does not dramatically change due to departures from this assumption for the equal sample size[1]. But when the difference of unequal variance is significant or with unequal sample size, the type I error can be highly inflated. The Satterthwaite method is used to make the adjustment for the degree of freedom for F test to overcome the drawback of inflated type I error.

4.2.1 Two Groups under Heteroscedasticity

If the test involves only two sample groups X and Y, we know ANOVA F test is the same as t test statistics which is:

$$T(X,Y) = \frac{\bar{X}_N - \bar{Y}_M}{\sqrt{S_e^2(1/N + 1/M)}} \tag{4.21}$$

S_e^2 is called the pooled estimate of the variance, M and N are the sample sizes of these two groups. The formula is

$$\begin{aligned} S_e^2 &= \frac{\sum_{i=1}^{a}(n_i - 1)S_i^2}{N - a} \\ &= \frac{(N-1)S_N^2 + (M-1)S_M^2}{N + M - 2} \end{aligned}$$

Where s_N^2, s_M^2 is the the sample variance of each group.

ANOVA F = t test for two sample groups

---

[1]Refer the Chpater 2 Design No.2

Set equal group size $M = N$. For ANOVA(F) test we know that the formula is

$$\begin{aligned} F &= \frac{\text{MST}}{\text{MSE}} \\ &= \frac{N/2(\bar{X}_N - \bar{Y}_N)^2}{S_e^2} \\ &= \frac{(\bar{X}_N - \bar{Y}_N)^2}{\frac{N}{2}S_e^2} \\ &= (\frac{(\bar{X}_N - \bar{Y}_N)}{\sqrt{(S_e^2(1/N + 1/N))}})^2 \\ &= (t)^2 \end{aligned}$$

Where

$$\begin{aligned} \text{MST} &= \frac{\text{SST}}{a-1} \\ &= \frac{N(\bar{X}_N - (\bar{X}_N + \bar{Y}_N)/2)^2 + N(\bar{Y}_N - (\bar{X}_N + \bar{Y}_N)/2)^2}{2-1} \\ &= N(\bar{X}_N/2 - \bar{Y}_N/2)^2 + N(\bar{(Y)}_N/2 - \bar{X}_N/2)^2 \\ &= \frac{N}{2}(\bar{X}_N - \bar{Y}_N)^2 \end{aligned}$$

If the assumption of equal variances is violated, we have to compute the adjusted t statistic using individual sample standard deviations rather than a pooled standard deviation. The modified t test is:

$$T'(X, Y) = \frac{\bar{X}_N - \bar{Y}_M}{\sqrt{\frac{1}{N}s_N^2 + \frac{1}{M}s_M^2}} \tag{4.22}$$

At the same time, the Satterthwaite method is used for approximating of the degrees of freedom. The formula of Satterthwaite adjustment for the freedom of degree is following instead of $N + M - 2$.

$$\text{df} \approx \frac{(\frac{1}{N}s_N^2 + \frac{1}{M}s_M^2)^2}{\frac{1}{N-1}(\frac{1}{N}s_N^2)^2 + \frac{1}{M-1}(\frac{1}{M}s_M^2)^2} \tag{4.23}$$

In literature it is pointed out that Satterthwaite method provides adjusted t test asymptotically approaches a t distribution, allowing for an approximate t test to be calculated when the population variances are not equal.

4.2.2 More than Two Groups under Heteroscedasticity

Modified t test cannot be applied for more than two groups, but one can still use Satterthwaite method for determining the denominator degrees of freedom in ANOVA F test.

In most references it has only provided the approximation of degree of freedom for two groups. But it can be generalized for more general cases, in stead of the original degrees of freedom $\sum_i n_i - a$, one can use the following

$$\mathrm{df} \approx \frac{\left(\sum_{i=1}^{i=a} \frac{1}{n_i} s_i^2\right)^2}{\sum_{i=1}^{i=a} \frac{1}{n_i - 1} \left(\frac{1}{n_i} s_i^2\right)^2} \tag{4.24}$$

Where n_i is the sample size and s_i^2 is sample variance of the ith group, $i = 1, \ldots, a$. The results from simulation study will be presented later.

4.3 Box-Cox Transformation

Significant violations of the assumptions can seriously increase the chances of committing either a Type I or II error. Thus, one reason utilizes data transformations is that it can reduce skewness or stabilize the variance. However, an appropriate transformation of a data set can often yield a data set that approximately follows a normal distribution.

Box and Cox (1964) have proposed a family of transformations that can be used with non-negative responses and which includes as special cases all the transformations in common use. The basic idea of the power transformation is

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0 \end{cases} \tag{4.25}$$

In our experiments we choose the logarithm for transforming studies samples in the simulation. In this paper transformation is applied in the most seriously violated situations where all together with non-normality, heterogeneous variance and censored data.

4.4 Resampling Method for Critical Value

When the assumptions for ANOVA F are heavily violated, the distribution of test departs from F distribution. But the critical value we still used is based on the fact that the test is supposed to be F distributed. When we modify ANOVA with KM, SW or BC methods, F distribution will also not be held, therefore the critical value which is obtained from F distribution is not so adequate in such case. Resampling methods which are independent of different data distribution are frequently used in practice to adjust critical values of test statistics. Special resampling methods, Bootstrap method and permutation test are investigated in this papar.

4.4.1 Bootstrap Method

Typically Bootstrap method is more often investigated in order to establish more accurate data dependent critical values. The bootstrap method attempts to determine the probability distribution from the data itself, without recourse to large sample size.

Suppose we have i.i.d data $X_1, \ldots, X_n$ from F and a statistic $T_n(X_1, \ldots, X_n) = T_n$. Bootstrap idea is:

1. Substitution principle. Replace F by $\hat{F}_n$, the empirical distribution based on original data $X_1, \ldots, X_n$.

2. Resampling from $\hat{F}_n$. $\hat{F}_n$ is known, which is conditioned on $X_1, \ldots, X_n$, draw B samples from $\hat{F}_n$, $X^*_{1,i}, \ldots, X^*_{n,i}$, $i = 1, \ldots, B$. Now statistic T_n is calculated by B groups samples, $T^*_{n,i} = T_n(X^*_{1,i}, \ldots, X^*_{n,i})$.

3. Repeat B times and look at the distribution of these objects, then T_n can be evaluated by $\frac{1}{B}\sum_{i=1}^{B} T^*_{n,i}$

Nonparametric Bootstrap

For nonparametric Bootstrap, $\hat{F}_n$ is empirical distribution which is obtained by resampling data from original data, $\forall i, j$, $X^*_{i,j} \in (X_1, \ldots, X_n)$, randomly draw a replacement observation from the original sample we put it back before drawing the next observation. As a result, any number can be drawn more than once or not at all.

Parametric Bootstrap

For parametric Bootstrap, instead of drawing with replacement from the original sample and one can draw $\hat{F}_n$ from parameter $\hat{\theta}$, where $\hat{\theta}$ is the estimation obtained from the original sample, one can write $\hat{F}_n = \hat{F}_{\hat{\theta}}$. Usually parameters are estimated by maximum likelihood estimates which is one of unbiased estimators.

4.4.2 Permutation Test

From practical point of view, permutation test is proposed instead of nonparametric Bootstrap method due to the fact that the sample size is small and part of them are censored in our experiment. If we choose to pick up data randomly from original data set as nonparametric Bootstrap does, the chance to get all censored data in each sample set is quite high. More practically, permutation test is based on permutation resamples drawn at random from the original data. Permutation resamples are drawn without replacement, in contrast to nonparametric Bootstrap samples. Thus we'd get the same set of numbers we started with, though in a different order. Each time the information of original data is fully used but only once.

Example. Permutation test for comparing two populations
Given independent sizes n and m from two populations

1. Permute orginial data $m+n$, get a resample of size n and a separate resample of size m. All the orginal data is used but once.
2. Compute statistic that compares the two groups
3. Repeat this resampling process hundreds of times.

In simulation studies parametric Bootstrap and permulation test will be carried out. Even though the problems addressed in this paper are in a parametric setting, but with severe violations of assumptions, skewness, censoring, and heterogenous variance, it's hard to tell if the data are still following any parametric distribution.

4.5 Simulation study

To investigate the performances of the power function based on proposed ideas of Satterthwaite (SW), Kaplan Meier (KM) and Box-cox (BC) transformation, for test statistics and Bootstrap method for approximating critical value, Monte Carlo simulation study is carried out. Two sample groups are compared in the test. The procedure is exactly the same as described before[2].

At first we investigate modifications of the test statistics, then according to the simulation results, methods which can offer the best for compensating drawbacks of test statistics, will take to the next step. Resampling method provides more accurate critical value.

4.5.1 Simulation Study of modified ANOVA

First of all, we want to look at the case where modifications are working individually, assuming there is no other disturbance from other violated assumptions. At each stage, it has checked whether single modification works properly for its own mission to remedy the drawback of the test as we've discussed in the theoretical part. If they offer some helps to the test as expected, we will combine these modifications together and examine them in the worst situation which contains the superposition of all violations of assumptions. But if these modifications do not work properly for individual assumption, we will not expect them to work for combined cases.

[2]Refer Chapter 2

The designs for the simulation are listed in the following tables[3]. In each figure, the curve with modifications are presented together with the one without modifications. In this way, one can get the direct view how the modified ANOVA performs. About the notations, the name for each modification is written with the test distribution in bracket, for example, classical ANOVA with Kaplan Meier method we simply denote as KM(F).

Table 4.2: Designs used in simulation

Design No.	Distribution	Censoring Percentage at null point	Difference of Deviation
1	N	0%	5
2	W	0%	5
3	N	16%	0
4	N	27%	0
5	W	16%	0
6	W	27%	0
7	N	16%	2.5
8	W	16%	2.5
9	N	27%	2.5
10	W	16%	2.5

SW Modification Under Heteroscedasticity

From design No.1 and No.2, under the heterogeneity of variance, small difference can be seen between ANOVA(F) test and ANOVA(F) with SW modification (SW(F)) with the equal sample size. It's known that the power of ANOVA (F) is robust to the moderate deviation from equal variance especially for equal sample size. For uequal sample size, one can find more obvious effect. Even though with small effect, the modification does help for keeping type I error close to the significant level when it has intended to inflate for both normal and Weibull distributions. But at the same time the whole power of test for rejecting false hypothesis is also correspondingly lower.

Kaplan Meier Method for Censoring Case

In the case of censoring, for classica ANOVA test with Kaplan Meier correction (KM(F)), the power function significantly improved for both normal case (design No.3, No.4) and Weibull case (design No.5 and No.6), supported by the evidence that notable space between two power curves, especially on the right side where the percentage of censoring is quite high. KM(F) increases the power for testing high percent censoring data situation. One can also notice that KM(F) lifts the whole power which is not good for the size of

[3]M=2.000 in each simulation study

test because the type I error is a little bit inflated. Comparing two cases of censoring percentage at null hypothesis, 16% and 27% respectively, the inlfated type I error is positively realted to the percentage of censoring data set. However, the inflated size of test problem can be fixed by Satterthwaite method which showed its ability to keep significant level.

For obtaining curve of power function, since KM method does not work in the situation when all group members are censored (this happens more often when the moving group shifting to right side), we choose to reject the hypothesis in such case to make the continuity of the power function (the probability of such case is less than %1 at each testing point).

Combined SW and KM Modifications

Even though with some "side effect", as a whole SW and KM modifications perform satisfactorily under the heterogeneity and censoring, now one can combine these two methods together and examine whether they can cooperate for the combination of violated conditions. From design No.7 and No.8, it can be observed that these two methods cooperate together to lift the power and keep the size of test closed to the significant level. Comparing to figures only with KM or SW modification, KM method always intends to make type I error inflate but SW method helps to drag it back to the significance level. When there two methods work together, SW method can counteract such effect from KM to some degree, and the whole power still higher than ANOVA(F). But as noticed before, the position of minimal point is shifting under the superposition of violations. This can be explained for the severely disturbed assumptions (heterogeneity and right censored together) which make data set heavily imbalanced and ANOVA(F) test cannot handle such situation properly. The shifted distance of minimal point depends on the difference of two group variances when the censoring percentage is fixed. KMSW(F) cannot totally fix shifting problem when the violation is serious.

Box-Cox Transformation

To examine if pre-pocess data is good idea in our study, we transform data by taking logarithm, applying pre-pocess data together with KM and SW method in the simulation study. Design No.9 and No.10 show that contrary to popular belief, transformation does not have a notable effect on the power curves under heterogeneity and censoring. We do not mean to downplay the importance of transformations for other purposes, but in such combined violated assumptions case it is worthless to use it as a mean of modification.

4.5.2 Resampling Methods for Critical Value

Resampling methods are the time consuming method, we only apply these methods for modifications which offer the best performance in the test statistics. Obviously, KMSW(F)

works satisfactorily among all kinds of modifications. Parametric Bootstrap and permutation test[4] are applied for obtaining the more accurate critical value for modified ANOVA under the most severe violation situations.

From design No.11, for the parametric bootstrap case, the power of modified test is

Table 4.3: Parametric Bootstrap and permutation test for KMSW(F)

Design No.	Distribution	Censoring Percentage at null point	Difference of Deviation	Method
11	W	16%	2.5	Parametric Bootstrap
12	W	16%	2.5	Permuation

lifted, including the type I error, it is more liberal compared to the KMSW(F). The power of right side is slightly more powerful than KMSW(F). Compared to the design No.12, in the permutation case, it can successfully keep the type I error at significance level, while the power of both sides is a little lower than KMSW(F) but still above ANOVA(F). The explanation for the better performance of permutation test is with severe violations of assumptions, the data is hardly following any parametric distribution. Therefore in such case permutation is a better choice to give a more reliable critical value.

4.5.3 Further Simulation

After comparing the power under two groups with equal sample sizes, we are interested in whether modified ANOVA test will work for a more general situations, for more groups or unequal sample sizes. Such general situations are very broad tasks to be achieved in our paper. However here we give a representative general look, three groups and unequal sample sizes are investigated if KMSW(F) will work for those specified situations.

Table 4.4: Designs used in simulation.$p_i(n_i)$ means p_i groups with k_i observations each

Design No.	Distribution	Censoring Percentage at null point	Difference of Deviation σ	$p_i(n_i)$
13	N	0	2.5	2(6,11)
14	W	0	2.5	2(6,11)
15	W	16%	2.5	2(6,11)
16	W	16%	2.5	3(10,10,10)
17	W	16%	2.5	3(6,8,10)

Design No.13 and No.15 show that with unequal sample size, the liberal type I error

[4]M=500 in resampling simulation

is under control more obviously by Satterthwaite method compared to Design No 1. and Design No. 2 with equal sample size. From design No.15 two groups with unequal sample size, from the right side of the curve, it seems that the power of ANOVA(F) is higher than KMSW(F), that's because the whole power curve of ANOVA(F) shifts to the left side. But the curve of KMSW(F) is approximately symmetrical with respect to the middle point. The size of KMSW(F) is also better than ANOVA(F) but with slightly inflated. One knows with the help of permutation method the inflated size of KMSW(F) will be more closed to significant level. Thus KMSW(F) is more powerful than ANOVA(F) for unequal sample size in two groups.

For more than two groups, form design No.16 and No.17, when the sample sizes are equal, KMSW(F) is also more powerful compared with ANOVA(F) directly observed from the curve. When with different sample size, in design No.16, the same as the performance of the design No.15, it shows ANOVA(F) is more powerful than KMSW(F) due to the shifting, but its not. Based on the result its yields conclusion, the modified ANOVA is not limited to the only two groups or equal sample size.

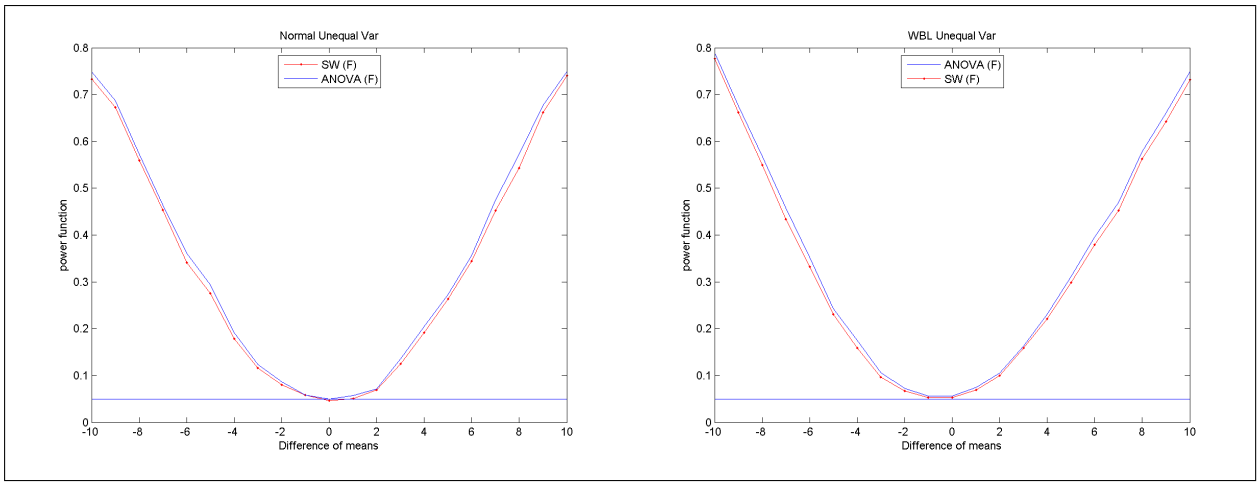

Figure 4.1: Design No.1 and Design No.2

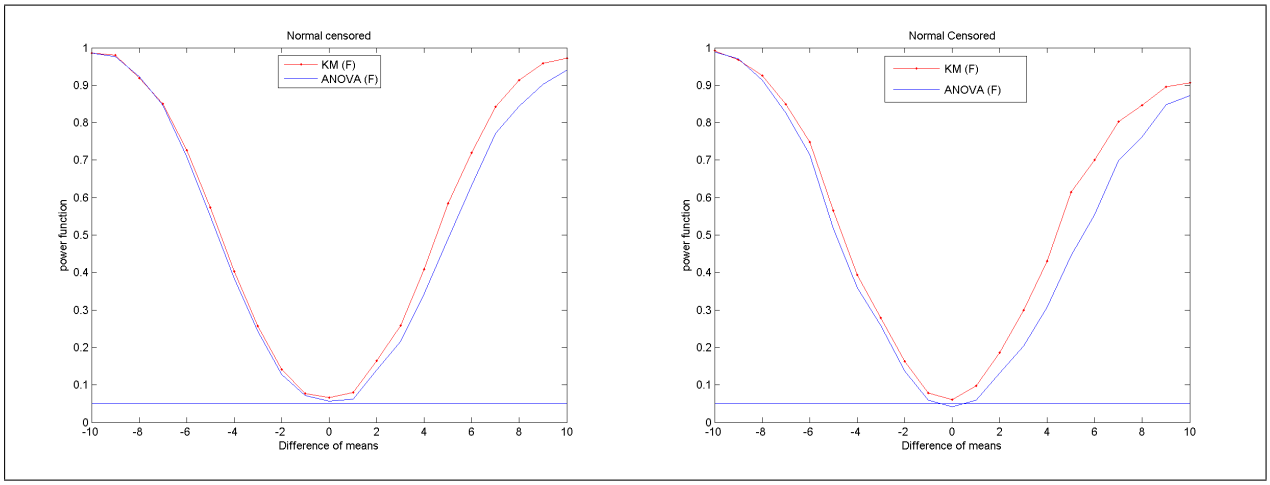

Figure 4.2: Design No.3 and Design No.4

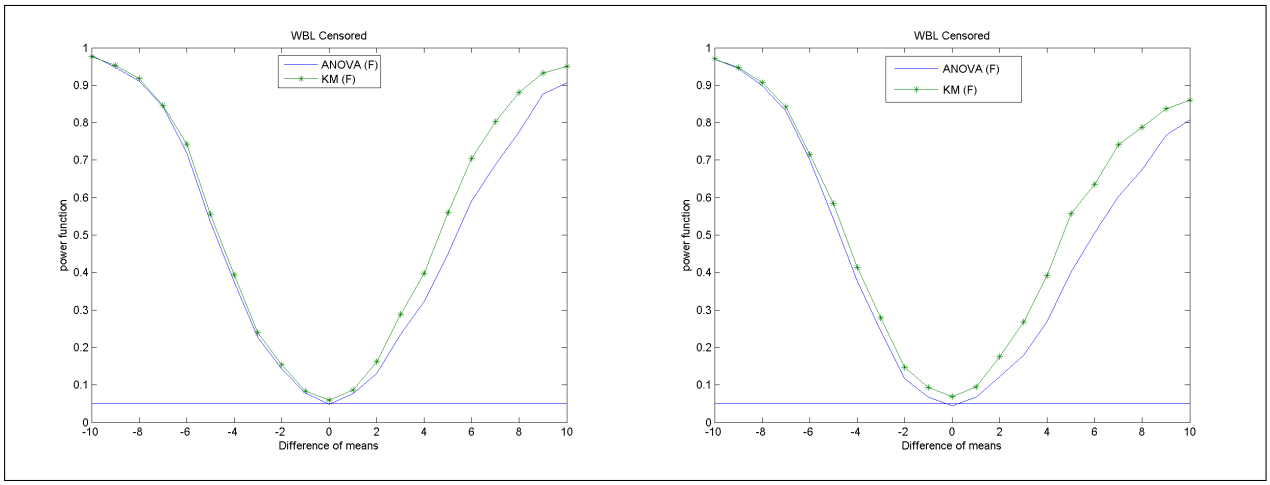

Figure 4.3: Design No.5 and Design No.6

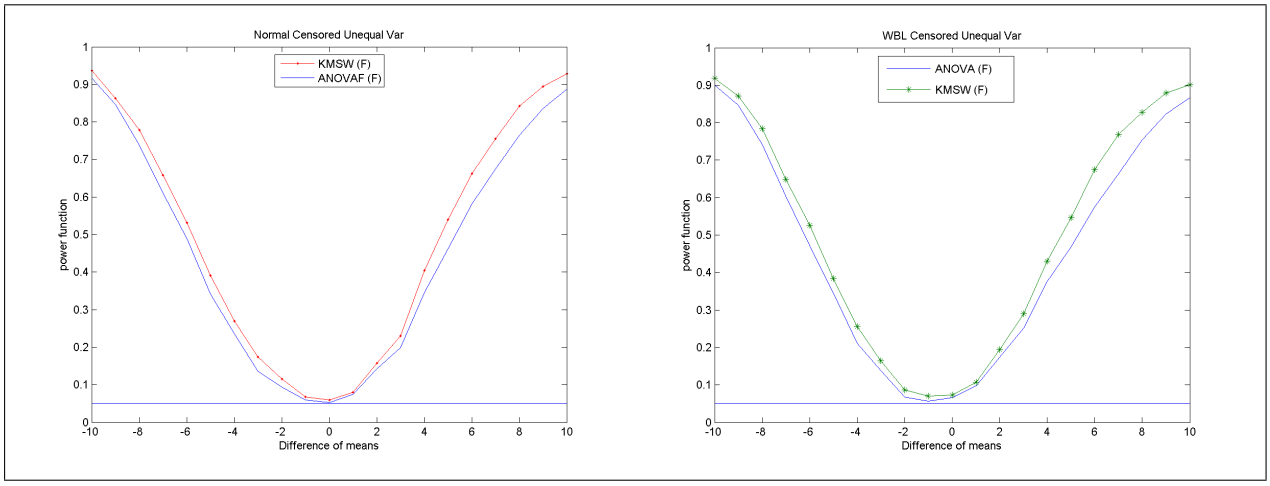

Figure 4.4: Design No.7 and Design No.8

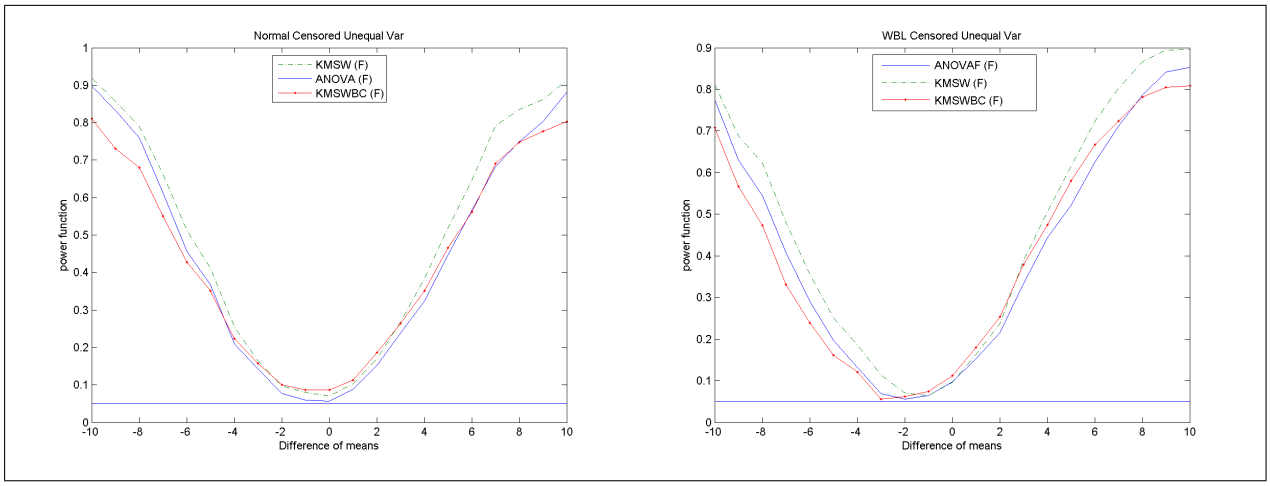

Figure 4.5: Design No.9 and Design No.10

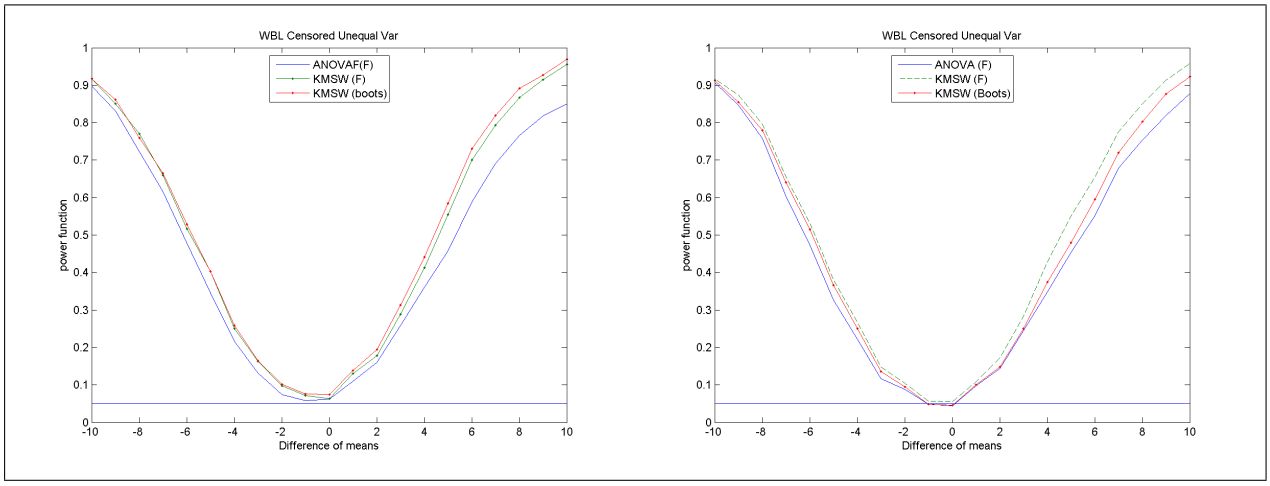

Figure 4.6: Design No.11 and Design No.12

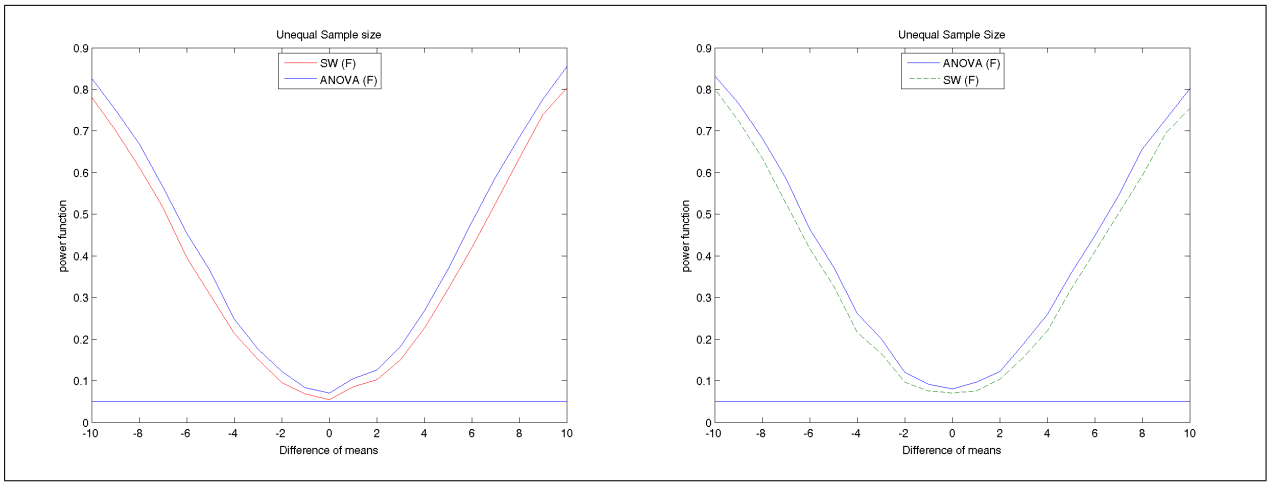

Figure 4.7: Design No.13 and Design No.14

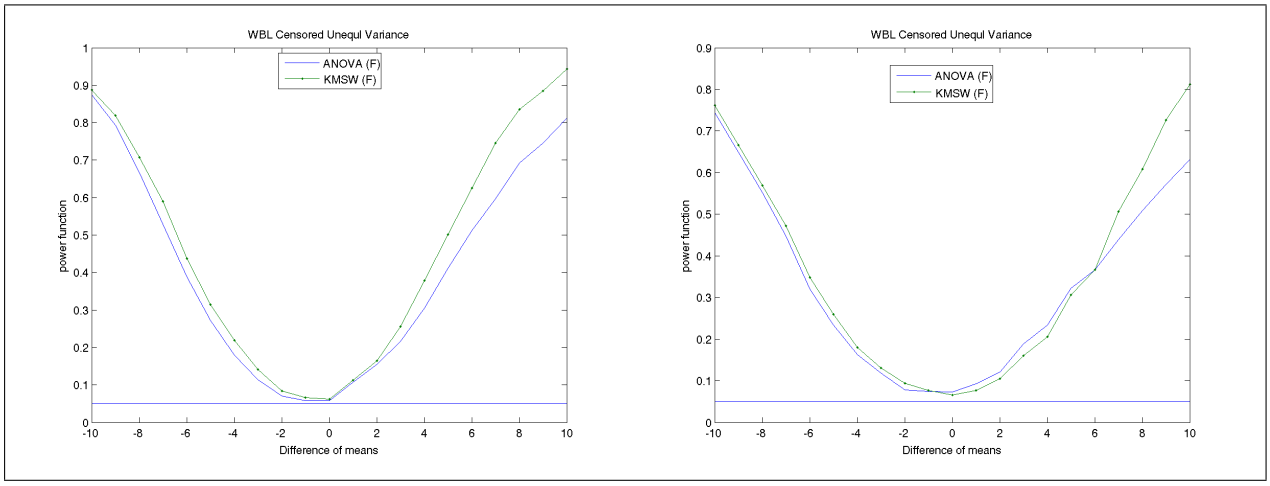

Figure 4.8: Design No.16 and Design No.17

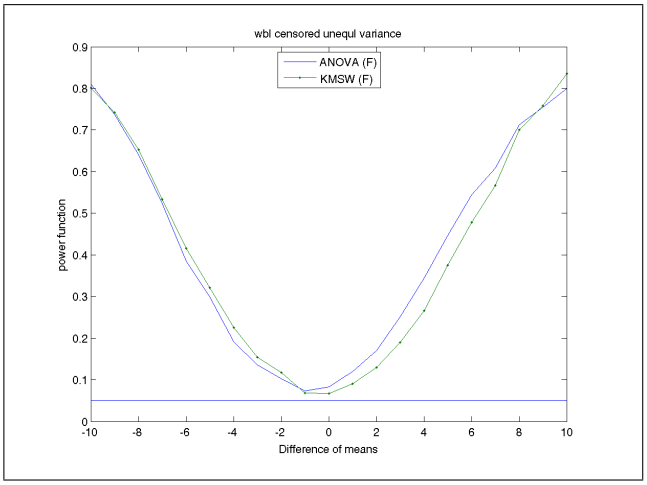

Figure 4.9: Design No.15

Chapter 5

Welch's test

From previous chapters, power comparisons were performed under violations of assumptions, nonnormality, heterogeneity of variance and censoring. Modified ANOVA with KM, SW and permutation method for the accurate critical value shows its robustness to the violation situations among all tests proposed till now. Another alternative test which is often recommended in situation where is under heterogeneous and nonnormality is proposed.

The proposed test is called Welch test, which is a p sample extension of Welch's (1938). The procedure is designed under heterogeneous variance. Kohr and Games's (1974) results indicate that the Welch procedure is more robust against heterogeneity of variance than classical ANOVA F test, particularly when the sample size is not equal. Levy (1971) shows welch procedure is also robust against nonnormality. The purpose of this chapter is to examine alternative Welch test to modified ANOVA F test, to determine whether the alternative is worth of utilizing when adding one more violation of censoring.

5.1 Welch's Test

Considering the case of p groups, $X_{ij}(i = 1, \ldots, p, j = 1, \ldots, n_i)$. Let T_i and s_i^2 be the usual unbiased estimates[1] of the mean μ and σ^2 respectively. If the σ_j^2 are not equal, but $\mu_1 = \mu_2 = \cdots = \mu_k$, Welch (1951) demonstrated that the distribution of the test statistic

$$V = \frac{\sum_j w_j(\bar{T}_j - \bar{G}')^2/(p-1)}{1 + \frac{2(p-2)\Lambda}{p^2-1}} \sim F_{p-1,f} \tag{5.1}$$

[1]One tries KM and MLE respectively

$F_{p-1,f}$ denotes the test statistics having an F distribution with $(p-1)$ and f degrees of freedom, where

$$w_i = n_i/s_j^2 \tag{5.2}$$

$$\bar{G}' = \sum^{p} w_i T_i / \sum^{p} w_j \tag{5.3}$$

$$\Lambda = \sum^{p} \frac{1}{n_i - 1}(1 - \frac{w_i}{\Sigma w_i})^2 \tag{5.4}$$

$$f = (p^2 - 1)/3\Lambda \tag{5.5}$$

The key principle of Welch's test is that it depends not only on the sample size but also the sample variance. When taking averages one should weigh each component with its information (one over its estimated variance), it makes Welch test less sensitive to the inequality of variances and allows to approximate the distribution of the function by an F distribution. Welch proposed the approximate degrees of freedom solution which is also well accepted and commonly used in practical applications because of its simplicity and accuracy.

The results of a Monte Carlo study will be presented comparing with modified ANOVA test and Welch test.

5.2 Simulation Study

The aim of simulation is power comparison for modified ANOVA and new proposed Welch test under the most severe violations of assumptions, heterogenous variance, nonnormality and censoring. The power results will show us whether the Welch test will be alternative test to modified ANOVA. With Monte Carlo study we construct exact approximate procedure, an intensive simulation[2]study was made in order to determine the power of several procedures studied.

Firstly two same sample size groups are examined, the test procedure is the same before[3], then three groups comparisons are investigated. Unequal sample sizes of both cases will be also examined. Designs for test are shown in the following table.

In the Welch test, unbiased estimates are required. Both maximum likelihood estimate and Kaplan Meier estimate are investigated in the simulation study. The power comparison will show, which one fits such situation better.

[2]M=2.000 in each simulation

[3]Refer chapter 2

Table 5.1: Designs used in simulation

Design No.	Estimate	Groups	Sample sizes
1	MLE	2	10,10
2	KM	2	10,10
3	MLE	2	6,11
4	KM	2	6,11
5	MLE	3	10,10,10
6	KM	3	10,10,10
7	MLE	3	6,8,10
8	KM	3	6,8,10

5.2.1 Simulation for Two Sample Size

The output of the simulations study has been seen in the figures from which made conclusions fairly clear. From the superposition of the violations of assumption, exaggerated sample variance difference, high percentage censoring data, mostly influences the right side. Welch test and modified ANOVA do not work for all censored data set, therefore to make the continuity of power function, the test will reject the hypothesis when all samples are censored.

For the equal sample size, from design No.1 and No.2 one can see that Welch test performs better for keeping type I error around significant level compared to modifeded ANOVA. Welch test with KM method for estimates gives better performance for keeping type I error while with MLE it gives better power on the left side where is not much disturbance from censoring. Nevertheless, modified ANOVA is the most powerful test among these three tests by comparing the power of both side, and for inflated type I error, it can be improved by applying Bootstrap method for critical value.

For unequal sample size, design No.3 and No.4 show that compared to modified ANOVA, Welch test has good performance for the size of the test. But the power of both side is low, especially on the right side of the curve when with KM estimate.

5.2.2 Simulation for further Study

Three groups and unequal sample sizes are compared in the simulation studies. From the design No.5 and No.6, when with three groups, the power of Welch test with KM estimate on the left side is lifted above than that of ANOVA(F), but not as good as modified ANOVA. For the type I error, both Welch test and modified ANOVA almost have no difference.

Design No.7 and No.8 show the unequal sample size case, Welch test has nealy the same

performace as modified ANOVA, but the type I error is a slightly more inflated. Modifed ANOVA is till most powerful test.

5.3 Conclusion

Welch test is proposed as a alternative to ANOVA under the heterogeneity and nonnormality. Power comparisons show us with censoring case in addition, the only benefit one can get from this proposed test is that it can keep type I error closed to the significant level. Compared to modifed ANOVA, this advantage is not much notable and the power of left side is much lower for the unequal case. Welch test is very sensitive to the censoring condition. Based on those simulation results, Welch test could not beat modified ANOVA to be alternative.

5.4 Recommendations

The ANOVA F test is so sensitive to the combined violations of assumptions that it should not be used when the assumptions are suspected. Two proposed alternative tests, likelihood ration test and Welch test, are not the alternative to use even though they give good performance under certain violation cases - likelihood ratio test is robust against the censoring case and Welch test is not so sensitive to unequal variance.

The modified ANOVA can be used in place of the ANOVA F test under the superposition of violations. Modifications includes Kaplan Meier method for estimates, Satterthwaite method for approximating the degree of freedom and permutation method for obtaining the desired critical value. The modified ANOVA gives the most satisfactory power performance due to our simulation studies which yield to small group number study. The simulation results indicate that modified ANOVA can be extended to more general cases. Nethertheless, the extension simulation studies which are for such general cases need to be examined in the future.

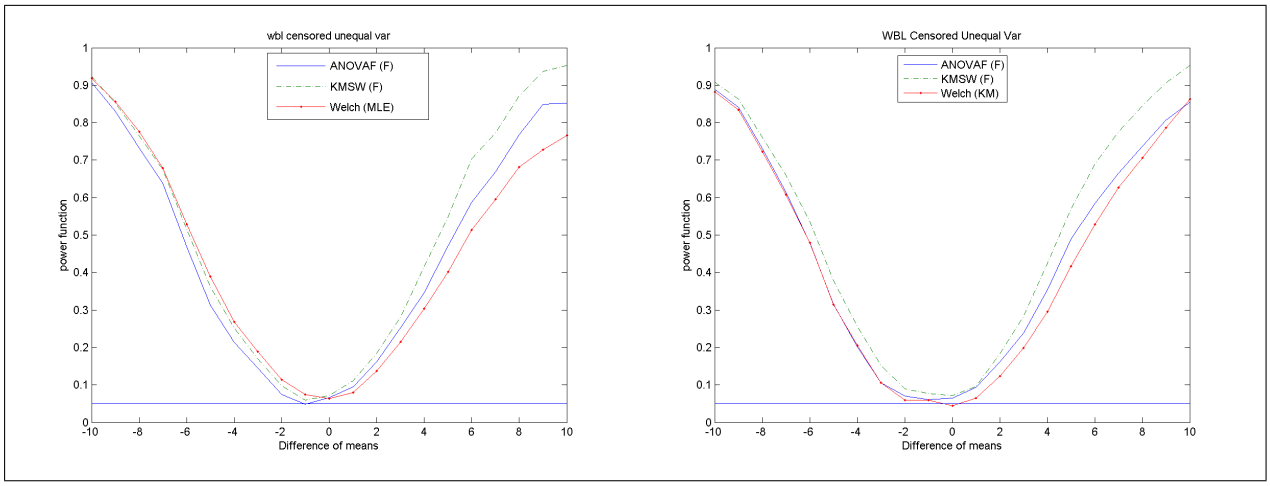

Figure 5.1: Design No.1 and Design.No.2

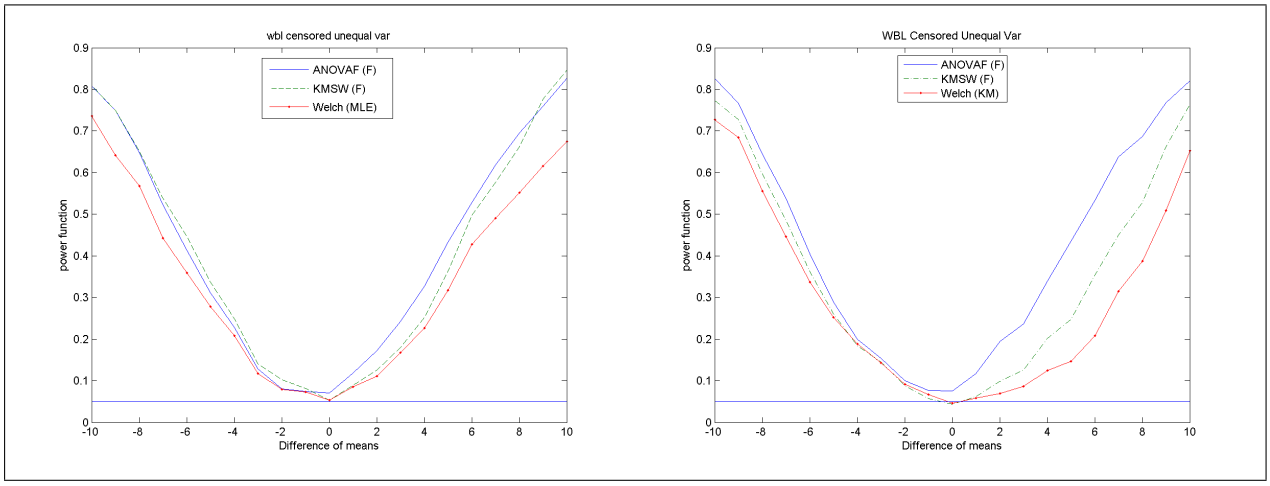

Figure 5.2: Design No.3 and Design No.4

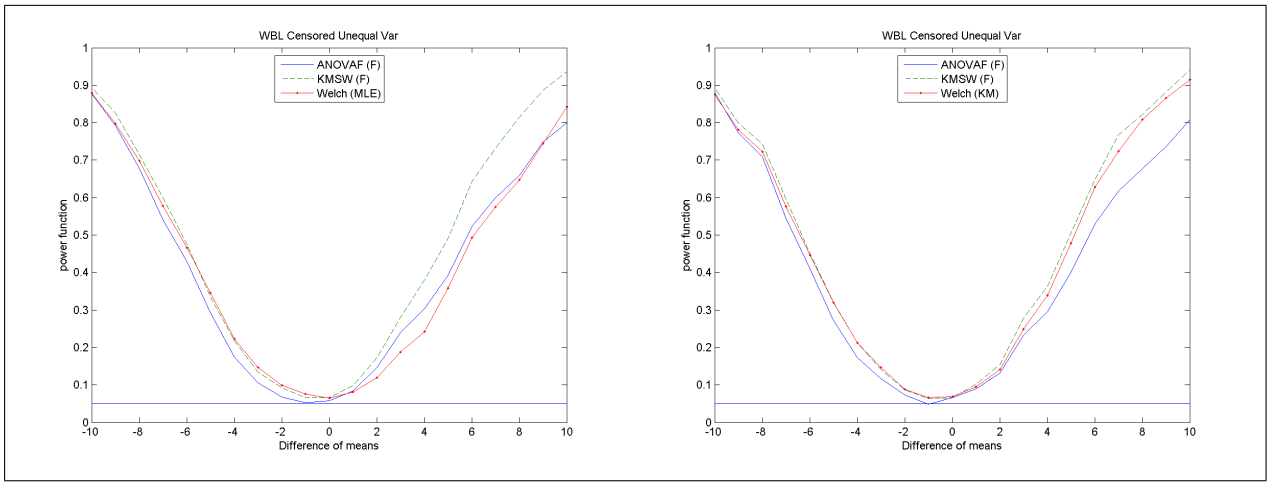

Figure 5.3: Design No.5 and Design No.6

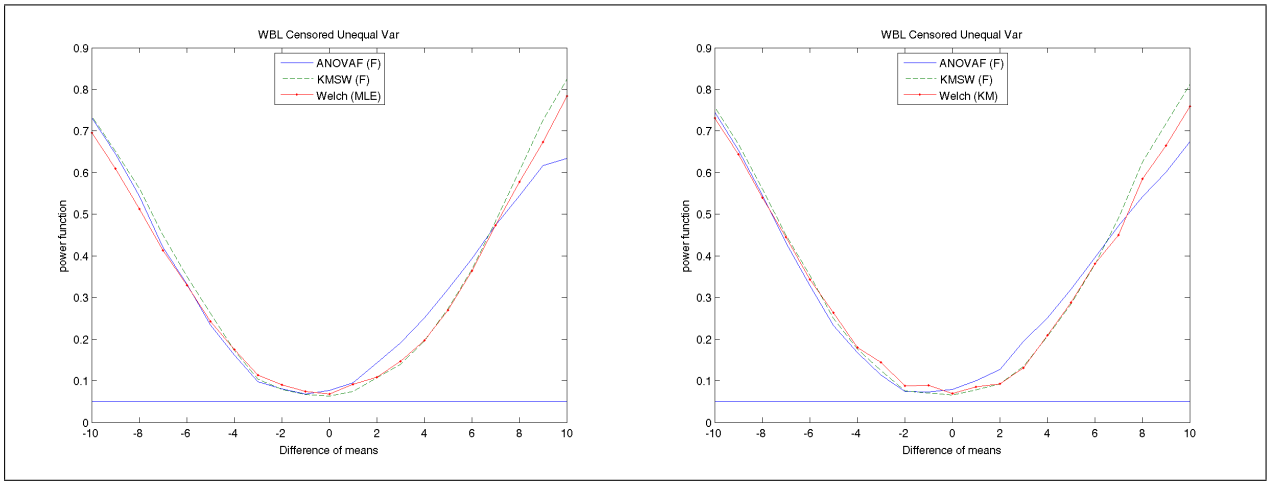

Figure 5.4: Design No.7 and Design No.8

Bibliography

[1] D. R. Cox and D. Oakes, 1984. *Analysis of Survival Data*, Chapman and Hall.

[2] E. T. Lee, 1992. *Statistical Methods For Survival Data Analysis*, Wiley series in probability and mathematical statistics.

[3] O. J. Dunn and V.A. Clark, 1974. *Applied Satistics: Analysis of Variance and Regression*, John Wiley and Sons.

[4] B. Efron, 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for industrial and applied mathematics.

[5] I. R. Miller, J. E. Freund and R. Johnson, 1990. *Probability and Statistics for Engineers*, Prentice-Hall, Inc.

[6] A. Wald, 1947. *Sequential Analysis*. Wiley, New York.

[7] P. Meier, 1975b. *Estimation of a Distribution Function From Incomplete Observations*. In prespectives in probability and statistics, edited by J. Gaui. Applied Probability Trust, Sheffield, England.

[8] N. E. Breslow and J. Crowley, 1974. *A Large Sample Study of the Life Table and Product Limit Estimates under Random Censoring*, Ann. Statist. 2, pp. 437-453.

[9] S. Weerahandi, 1995. *ANOVA under Unequal Error Variances*, Biometrics 51, 589-599.

[10] R. G. Krutchkoff, 1988. *One-way Fixed Effects Analysis of Variance when the Error Variances May be Unequal*, J. Statist. Comput. Siml., Vlo. 30, pp. 259-271.

[11] K. J. Levy, 1978. *An Empirical Comparison of the ANOVA F-Test with Alternatives which are More Robust against Heterogeneity of Variance*, J. Statist. Comput. Simul., Vol. 8, pp. 49-57.

[12] R. L. Kohr and P. A. Games, 1947. *Robustness of the Analysis of Variance, the Welch Procedure and a Box Procedure and to Heterogeneous Variances*, Journal of Experimental Education, 43, No.1.

[13] K. J. Levy, 1978. *Some Empirical Power Results Associated with Welch's Robust Analysis of Variance Technique*, J. Statist. Comput. Simul., Vol. 8, pp. 43-48.

[14] Y. Chi, 2005. *Multiple Testing Procedures Based on Weighted Kaplan-Meier Statistics for Right-censored Survival Data*, Statist. Med. Vol. 24, pp. 23-35.

[15] A. Janssen and T. Pauls, 2003. *How Do Bootstrap And Permutation Tests Work?*, The Ann. Statist. Vol. 31, No. 3, pp. 768-806.

[16] K. Krishnamoorthy, F. Lu and T. Mathew. *A Parametric Bootstrap Approach for ANOVA with Unequal Variances: Fixed and Random Models.*

[17] B. L. Welch, 1938. *The Significance of the Difference Between Means when the Population Variances are Unequal*, Biometrika 29, pp. 350-62.

[18] S. Feth, 2006. *Biasreduzierung fuer Quantilschaetzer in der Betriebsfestigkeit*, Diplomarbeit, TU Kaiserslautern.