CHAPTER 31

Statistical approaches to make decisions in clinical experiments

Albert Vexler and Xiwei Chen
Department of Biostatistics, School of Public Health and Health Professions, State University of New York at Buffalo, Buffalo, NY, USA

Q2

THEMATIC SUMMARY BOX

At the end of this chapter, students should be able to:

- Correctly formulate statistical hypotheses with respect to the aims of epidemiological and/or biomedical studies
- Construct and provide statistical decision-making test rules corresponding to practical experiments
- Use parametric and nonparametric likelihood testing techniques in applied researches
- Understand basic properties of likelihood ratio type tests in parametric and nonparametric manners
- Use basic test procedures and their components in practical statistical decision-making mechanisms
- Employ statistical software at a beginning level

Introduction, preliminaries, and basic components of statistical decision-making mechanisms

Often, experiments in biomedicine and other health-related sciences involve mathematically formalized tests, employing appropriate and efficient statistical procedures to analyze data. Mathematical strategies to make decisions via formal rules play important roles in medical and epidemiological discovery, in policy formulation, and in clinical practice. In this context, in order to make conclusions about populations on the basis of samples from those populations, clinical trials commonly require the application of the mathematical statistical discipline.

The aim of the scientific methods in decision theory is to simultaneously maximize quantified gains and minimize losses in reaching a conclusion. For example, statements of clinical experiments can request to maximize factors (gains) such as

Oxidative Stress and Antioxidant Protection: The Science of Free Radical Biology & Disease, First Edition. Edited by Donald Armstrong and Robert D. Stratton.

accuracy of diagnosis of medical conditions, faster healing, and greater patient satisfaction, while minimizing factors (losses) such as efforts, durations of screening for disease, more side effects, and costs of the experiments.

There are many constraints and formalisms to deal with while constructing statistical tests. An essential part of the test-constructing process is that statistical hypotheses should be clearly formulated with respect to the objectives of clinical studies.

Statistical hypotheses

Commonly, statistical hypotheses and the corresponding clinical hypotheses are associated but stated in different forms and orders. In most clinical experiments, we are interested in tests regarding characteristics or distributions of one or more populations. In such cases, the statistical hypotheses must be very carefully formulated, and formally and clearly stated, displaying, for example, the nature of associations between characteristics or distributions of populations. For example, suppose that the clinical hypothesis is that the population mean time to heal with an antibiotic is different from the mean time to heal without the antibiotic. In this case, the statistical hypothesis to be tested should be that the population mean time to heal with an antibiotic is equivalent to the mean time to heal without the antibiotic. Here, we will test for the equivalence of parameters of populations. Note that one can ask to test for distribution difference with/without the antibiotic.

The term *Null Hypothesis*, symbolized H_0, is commonly used to show our primary statistical hypothesis. For example, when the clinical hypothesis is that a biomarker of oxidative stress has different circulating levels with respect to patients with and without atherosclerosis, a null hypothesis can be proposed corresponding to the assumption that levels of the biomarker in individuals with and without atherosclerosis are distributed equally. Note that the clinical hypothesis points out that we want to indicate the discriminating power of the biomarker, whereas H_0 says there are not significant associations between the disease and biomarker's levels. The reason lies in the ability to formulate H_0 clearly and unambiguously, as well as quantify and calculate expected errors in decision-making procedures. If the null hypothesis were formed in a similar manner to the clinical hypothesis, we probably could not unambiguously determine which links between the disease and biomarker's levels we should test.

Common errors related to the statistical testing mechanisms

The null hypothesis is usually a statement to be tested. Commonly, the statistical testing procedure results in a decision to reject or not reject the null hypothesis. In the context of testing statistical hypotheses, in order to provide a formal test procedure, as well as compare mathematical strategies for making decisions (e.g., with respect to statistical powers of tests), algorithms for monitoring test characteristics associated with the probability to reject a correct hypothesis should be considered. Here, we define the statistical power of a test as the probability that H_0 is correctly rejected when H_0 is false. In general, while developing and applying test procedures, the practical statistician faces the task of controlling the probability of the event that a test's outcome requests to reject H_0 when in fact H_0 is correct, a Type I error. For example, assume that L is the test statistic based on the observed data, C is a threshold, and

the decision rule is to reject H_0 for large values of L, that is, when $L > C$; then, the threshold should be defined such that $\Pr(L > C|H_0) = \alpha$, where α is a presumed significance level, that is, the probability of committing a Type I error. Note that when we compare two statistical tests, we mean to compare powers of the tests, given that the rate of Type I error is fixed.

It is clear that in order to construct statistical tests, we must review the corresponding clinical study, formalizing objectives of the experiments and making assumptions in hypothesis testing. A violation of the assumptions can result in incorrect results from the test, as well as a vital malfunction of the Type I error control system. Moreover, should the user verify that the assumptions are satisfied, errors in the verifications can affect the Type I error control.

The practitioner may also be interested to consider another related type of error in statistical testing procedures. If H_0 is false but fails to be rejected, the incorrect decision of not rejecting H_0 is called a Type II error. The Type II error rate can be defined as $\Pr(L < C|H_1)$, when we assume that L is the test statistic based on the observed data, C is a threshold, and the decision rule is to reject H_0 when $L > C$. Type II errors may occur when the effect size, biases in testing procedures, and random variability combine to lead to results insufficiently inconsistent with H_0 to reject it.[1] Essentially, it is the dichotomization of the study results into the categories "significant" or "not significant" that leads to Type I and Type II errors. Although errors resulting from an incorrect classification of the study results would seem to be unnecessary and avoidable, the Neyman–Pearson (dichotomous) hypothesis testing is ingrained in scientific research due to the apparent objectivity and definitiveness of the pronouncement of significance.[1,2]

Q3

p-Values

The traditional testing procedure assumes to define a test threshold and reject or not reject H_0 based on comparisons between values of tests statistics and the threshold. An alternative approach to hypothesis testing is to obtain the *p*-value.

As a continuous measure of the compatibility between a hypothesis and data, a *p*-value is defined as the probability of obtaining a test statistic (a corresponding quantity computed from the data, such as a *t*-statistic) at least as extreme or close to the one that was actually observed, assuming that H_0 is true.[3] *p*-Values can be divided into two major types: one-sided (upper and lower) and two-sided. Assuming there are no biases in the data collection or the data analysis procedure, an upper one-sided *p*-value is the probability under the test hypothesis that the test statistic will be no less than the observed value. Similarly, a lower one-sided *p*-value is the probability under the test hypothesis that the test statistic will be no greater than the observed value. The two-sided *p*-value is defined as twice the smaller of the upper and lower *p*-values.[2,4]

If the *p*-value is small, it can be interpreted that the sample produced a very rare result under H_0, that is, the sample result is inconsistent with the null hypothesis statement. On the other hand, a large *p*-value indicates the consistency of the sample result with the null hypothesis. At the pre-specified α significance level, the decision is to reject H_0 when the *p*-value is less than or equal to α; otherwise, the decision is to not reject H_0. Therefore, the *p*-value is the smallest level of significance at which

H_0 would be rejected. In addition to providing a decision-making mechanism, the p-value also sheds some light on the strength of the evidence against H_0.[5]

Misinterpretations of p-values are common in clinical trials and epidemiology. In one of the most common misinterpretations, p-values are erroneously defined as the probabilities of test hypotheses. In many situations, the probability of the test hypothesis can be computed, but it will almost always be far from the two-sided p-value.[3] Note that the p-values can be viewed as a random variable, uniformly distributed between 0 and 1 if the null hypothesis is true. For example, suppose that the test statistic L has a cumulative distribution function (CDF) F under H_0, and a CDF G under a one-sided upper-tailed alternative H_1. Then, the p-value is the random variable $1 - F(L)$, which is uniformly distributed under H_0.[6]

Sorts of information applicable to construct test procedures

The interests of clinical investigators usually lead to the problem of mathematically expressing procedures, using statistical decision rules based on sample data to test statistical hypotheses. In this case, when the users construct the decision rules, two additional information resources can be incorporated. The first is a defined function that consists of the explicit, quantified gains and losses in reaching a conclusion and their relative weights. Frequently, this function determines the loss that can be expected corresponding to each possible decision. This type of information can incorporate a loss function into the statistical decision-making process.

The second source reflects prior information. Commonly, in order to derive prior information, researchers should consider past experiences in similar situations. The Bayesian methodology formally provides clear technique manuals on how to construct efficient statistical decision rules for various complex problems related to clinical experiments, employing prior information.[7,8]

Parametric approach

A clinical statistician may use a sort of technical statements related to the observed data, while constructing the corresponding decision rules. The above-mentioned types of information used for test constructing can induce the technical statements, which oftentimes are called assumptions regarding the distribution of data. The assumptions often define a fit of the data distribution to a functional form that is completely known, or known up to parameters, since a complete knowledge of the distribution of data can provide all the information investigators need for efficient applications of statistical techniques. However, in many scenarios, the assumptions are presumed and very difficult to prove, or to test for being proper. The simple, but widely used, assumptions in biostatistics are that data derived via a clinical study follow one of the commonly used distribution functions: the normal, lognormal, t, χ^2, gamma, F, binominal, uniform, Wishart, and Poisson. The data distribution function can be defined up to parameters.[9] For example, the normal distribution $N(\mu, \sigma^2)$ is the famous bell curve, where the parameters μ and σ^2 represent the mean and variance of the population from which the data were sampled. The values of the parameters μ and σ^2 may be assumed to be unknown. Mostly, in such cases, assumed functional forms of the data distributions are involved in making statistical

decision rules via the use of statistics, which we name Parametric Statistics. If certain key assumptions are met, parametric methods can yield very simple, efficient, and powerful inferences.

Nonparametric approach

The statistical literature has widely addressed the issue that parametric methods are often very sensitive to moderate violations of parametric assumptions, and hence nonrobust.[10] The parametric assumptions can be tested in order to reduce the risk of applying a misleading parametric approach. Note that in order to test for parametric assumptions, a goodness-of-fit test, outlined in a later section of this chapter, can be applied. In this case, statisticians can try to verify the assumptions, while making decisions with respect to main objectives of the clinical study. This leads to very complicated topics, dealt with in multiple testing. For example, it turns out that a computation of the expected risk of making a wrong decision strongly depends on the errors that can be made by not rejecting the parametric assumptions. The complexity of this problem can increase when researchers examine various functional forms to fit the data distribution in order to apply parametric methods. A substantial body of theoretical and experimental literature has discussed the pitfalls of multiple testing, placing blame squarely on the shoulders of the many clinical investigators who examine their data before deciding how to analyze it, or neglect to report the statistical tests that may not have supported their objectives.[11] In this context, one can present different examples, both hypothetical and actual, to get to the heart of issues that especially arise in the health-related sciences. Note, also, that in many situations, due to the wide variety and complex nature of problematic real data (e.g., incomplete data subject to instrumental limitations of studies), statistical parametric assumptions are hardly satisfied, and their relevant formal tests are complicated or not readily available.[12]

Unfortunately, even clinical investigators trained in statistical methods do not always verify the corresponding parametric assumptions, nor attend to probabilistic errors of the corresponding verification, when they use well-known elementary parametric statistical methods, for example, the *t*-tests.

Thus, it is known that when the key assumptions are not met, the parametric approach may be extremely biased and inefficient when compared to its robust nonparametric counterparts. Statistical inference under the nonparametric regime offers decision-making procedures, avoiding or minimizing the use of the assumptions regarding functional forms of the data distributions.

In general, the balance between parametric and nonparametric approaches can boil down to expected efficiency versus robustness to assumptions. One very important issue is preserving the efficiency of statistical techniques through the use of robust nonparametric likelihood methods, minimizing required assumptions about data distributions.[5,13]

Remarks

A wealth of additional applied and theoretical materials related to statistical decision-making procedures may be found in a variety of scientific publications.[2,4,5,9,10,13–15]

This chapter is organized as follows: In section "R: statistical software," the statistical software R is outlined at a beginning level. The likelihood methodology is described in section "Likelihood." In section "Tests on means of continuous data," we show different tests for means of continuous data. Section "The exact likelihood ratio test for equality of two normal populations" reviews the exact likelihood ratio test for equality of two normal populations. Section "Empirical likelihood" introduces the empirical likelihood methodology. In section "Receiver operating characteristic curve analysis," we introduce common methods based on the receiver operating characteristic (ROC) curves used in biomedical and epidemiological researches to make statistical decisions.[16,17] Goodness-of-fit tests are reviewed in section "Goodness-of-fit tests." In section "Wilcoxon rank-sum tests," we review the Wilcoxon two-sample test, a nonparametric analogy to the two-sample t-test. Different tests for independence are introduced in section "Tests for independence." Section "Numerical methods for calculating critical values and powers of statistical tests" presents the method for Type I error control using Monte Carlo techniques. In section "Concluding remarks," we conclude this chapter with remarks.

R: statistical software

In section "R: statistical software," we outline the use of R, a powerful and flexible statistical software language.[18–20] Examples of statistical techniques implemented using R codes are employed in the chapter material.

R is a free case-sensitive, command line-driven software for statistical computing and graphics. Once the R program is installed via www.r-project.org and starts up, the main input window and a short introductory message (which appears a little differently on each operating system) are presented.[19] For example, Figure 31.1 shows the main input window in the operating system Windows with a few menus available at the top. Below the header a blank line is presented, with a screen prompt symbol > in the left-hand margin, showing the place where commands should be typed.

For example, we consider a simple way to input data using the `c()` function, which creates a vector, a variable with one or more values of the same type. We input a data containing 3, 5, 10, and 7, as shown below.

```
> x<-c(3,5,10,7)
```

To see the value of x, we type in the name of the vector x after the prompt symbol and press the Return key.

```
> x
```

As a result, R provides the following output:

```
[1]  3  5 10  7
```

R can perform simple statistical calculations in addition to very complex computations. Table 31.1 shows some simple commands that produce descriptive statistics of a vector x created based on a sample of measurements $X_1, \ldots, X_n$.

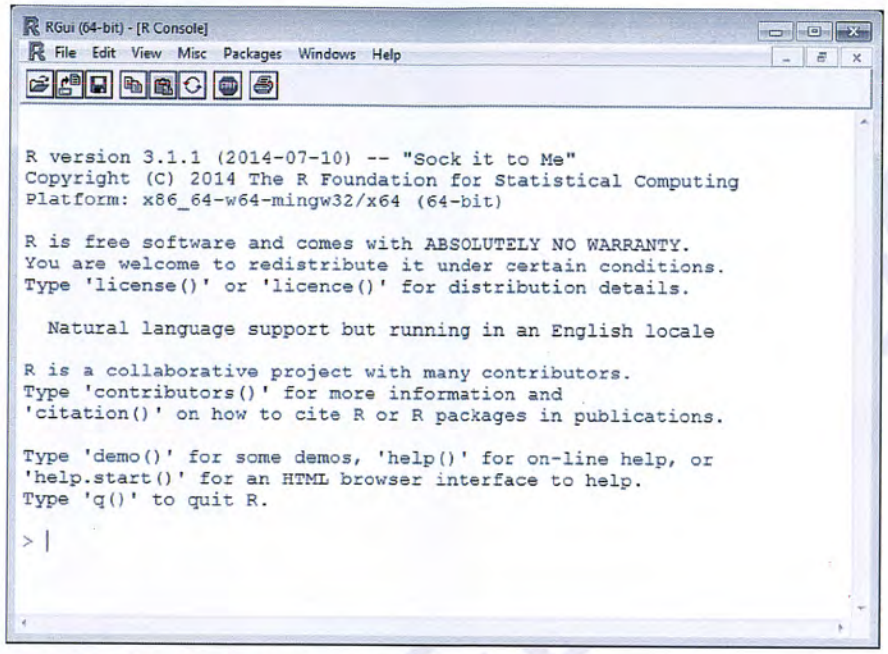

Figure 31.1 Screenshot of the R interface.

Table 31.1 The R commands that produce introductory descriptive statistics based on a numerical vector *x*.

Command	Explanations
`mean(x, na.rm = FALSE)`	Calculates the arithmetic mean of *x*.
`sd(x, na.rm = FALSE)`	Calculates the sample estimator of the standard deviation of *x*.
`var(x, na.rm = FALSE)`	Calculates the sample estimator of the variance of *x*.
`sum(x, na.rm = FALSE)`	Calculates the sum of the elements of *x*.

Instead of using the built-in functions such as those shown in Table 31.1, if a custom function need to be created to carry out some specific tasks, the `function()` command can be used. The following example shows a simple function `mymean` that determines the running mean of the first i, $i = 1, \ldots, n$ elements of a vector x, where n is the number of elements in x. Results are shown for x, specified above by applying the customized function.

```
> mymean <- function(x) {
+   tmp <- c()
+   for(i in 1:length(x)) tmp[i] <- mean(x[1:i])
+   return(tmp)
+ }
```

```
> mymean(x)
[1] 3.00 4.00 6.00 6.25
```

Note that the symbol + is shown at the left-hand side of the screen instead of > when it is working, meaning that the last command typed is incomplete. The `function()` can also be used to create complicated functions.

There are many packages in R that can be downloaded and installed from CRAN-like repositories or local files using the command `install.packages("packagename")`, where `packagename` is the name of the package to be installed and must be in quotes; single or double quotes are both fine as long as they are not mixed. Once the package is installed, it can be loaded by issuing the command `library(packagename)`, and commands available in the package can be accessed. Through an extensive help system built into R, a help entry for a specified command can be brought up via the `help(commandname)` command. As a simple example, we introduce the command `EL.means` in the `EL` library.

```
> install.packages("EL")
Installing pack-
age into 'C:/Users/xiwei/Documents/R/win-library/3.1'
(as 'lib' is unspecified)
trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.1/
EL_1.0.zip'
Content type 'application/zip' length 53774 bytes (52 Kb)
opened URL
downloaded 52 Kb

package 'EL' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
 C:\Users\xiwei\AppData\Local\Temp\Rtmp4uRCPS\downloaded
_packages
> library(EL)
> help(EL.means)
```

The use of the function `EL.means` provides possibilities to implement the empirical likelihood tests that are introduced in detail in section "Empirical likelihood."

As another concrete example, we show the `mvrnorm` command in the `MASS` library.

```
> install.packages("MASS")
Installing package into 'C:/Users/xiwei/Documents/R/win-
library/3.1'
(as 'lib' is unspecified)
trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.1/
MASS_7.3-34.zip'
Content type 'application/zip' length 1083003 bytes (1.0 Mb)
opened URL
```

```
downloaded 1.0 Mb

package 'MASS' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
 C:\Users\xiwei\AppData\Local\Temp\Rtmp4uRCPS\downloaded
_packages
> library(MASS)
> help(mvrnorm)
```

The mvrnorm command is very useful for simulating the data from a multivariate normal distribution. To illustrate, we simulate bivariate normal data with mean $(0, 0)^T$ and an identity covariance matrix with a sample size of 5.

```
> n <- 5  # define the sample size
> mu <- c(0,0)  # define the mean vector
> Sigma <- matrix(c(1,0,0,1), byrow=TRUE, ncol=2) # define
#covariance matrix
> set.seed(123)  # define the seed to fix the sample
> X <- mvrnorm (n, mu=mu, Sigma=Sigma) # generate data
> X
           [,1]        [,2]
[1,] -1.7150650 -0.56047565
[2,] -0.4609162 -0.23017749
[3,]  1.2650612  1.55870831
[4,]  0.6868529  0.07050839
[5,]  0.4456620  0.12928774
```

Likelihood

One of the traditional instruments used in medical experiments and drug development is the testing of statistical hypotheses based on the *t*-test or its different modifications. Despite the fact that these tests are straightforward with respect to their applications in clinical trials, it should be noted that there has been a huge literature on the criticism of *t*-test-type statistical tools. One major issue that has been widely recognized is the significant loss of efficiency of these procedures under different distributional assumptions. The legitimacy of *t*-test-type procedures also comes into question in the context of inflated Type I errors seen when data distributions differ from normal and the number of observations is fixed. This can pose serious problems when data based on biomarker measurements are available for statistical testing. The recent biostatistical literature has well addressed the arguments that show the values of biomarker measurements that tend to follow skewed distributions, for example, a lognormal distribution.[21] Hence, the use of *t*-test-type techniques in this setting is suboptimal and is accompanied by significant difficulties in controlling the corresponding Type I error.

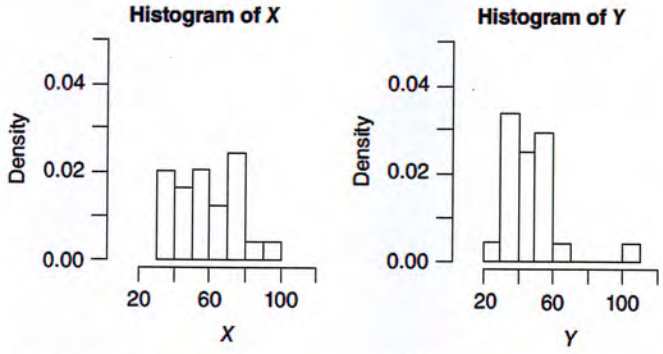

Figure 31.2 R data analysis output for measurements of HDL cholesterol levels (mg/dl) in healthy individuals.

Consider the following example based on the data from a study evaluating biomarkers related to atherosclerotic coronary heart disease:[22]

A cross-sectional population-based sample of randomly selected residents (aged 35–79) of Erie and Niagara counties of the state of New York, USA, was the focus of this experiment. The New York State Department of Motor Vehicles drivers' license rolls were employed as the sampling frame for adults between the ages of 35 and 65, whereas the elderly sample (aged 65–79) was randomly selected from the Health Care Financing Administration database. Participants provided a 12-h fasting blood specimen for biochemical analysis at baseline, and a number of characteristics were evaluated from fresh blood samples. The samples *X* and *Y* presented 50 measurements (mg/dl) of the biomarker "high-density lipoprotein (HDL) cholesterol" obtained from healthy patients. These measurements were divided into the two groups: *X* and *Y*. The following R code shows the input of the data and the construction of histograms of the data, as seen in Figure 31.2

The missed text is shown on the next page

Although one can reasonably expect the samples are from the same population, the *t*-test result shows a significant difference of their distributions, as demonstrated below via the use of the function `t.test` in R.

```
> t.test(X,Y)

  Welch Two Sample t-test

data:  X and Y
t = 2.1526, df = 47.704, p-value = 0.03644
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
  0.6657898 19.5742102
sample estimates:
mean of x mean of y
    57.20     47.08
```

The samples X and Y presented 50 measurements (mg/dl) of the biomarker "HDL (high-density lipoprotein) cholesterol" obtained from healthy patients. These measurements were divided into the two groups: X and Y. The following R code shows the input of the data and the construction of histograms of the data, as seen in Figure 28.2.

```
X<-c(37.4,70.4,52.8,46.2,74.8,96.8,41.8,55.0,83.6,63.8,63.8,52.8,46.2,37.4,5
0.6,74.8,46.2,39.6,70.4,30.8,74.8,61.6,30.8,74.8,52.8)
Y<-c(44.0,35.2,110.0,63.8,44,26.4,52.8,30.8,39.6,44,48.4,39.6,55,52.8,50.6,3
9.6,35.2,55,57.2,37.4,30.8,46.2,50.6,44,44)
> a<-min(c(X,Y))-20
> b<-max(c(X,Y))+20
> par(pty="s",mfrow=c(1,2))
> hist(X,xlim=c(a,b),ylim=c(0,0.05),freq=FALSE)
> hist(Y,xlim=c(a,b),ylim=c(0,0.05),freq=FALSE)
```

Figure 28.2 Here.

Although one can reasonably expect the samples are from the same population, the t-test result shows a significant difference of their distributions,

The Missed Text on p. 516.

Perhaps, in order to investigate reasons for this incorrect output of the t-test, the following issues may be taken into account:

The histograms displayed in Figure 31.2 indictate that the distributions of the X and Y are probably skewed. In a nonasymptotic context, when the sample sizes are relatively small, one can show that the t-test statistic is a product of likelihood ratio-type considerations, based on normally distributed observations.[14] That is, the t-test is a parametric test, and the parametric assumption seems to be violated in this example.

Thus, in many settings, it may be reasonable to propose an approach for developing statistical tests, attending to data distributions, in order to provide procedures that are as efficient as the t-test based on normally distributed observations. Toward this end, the likelihood methodology can be employed.

Likelihood ratio and its optimality

Now we turn to outlining the likelihood principle. When the forms of data distributions are assumed to be known, the likelihood principle is a central tenet for developing powerful statistical inference tools for use in clinical experiments. The *likelihood method*, or simply the *likelihood*, is arguably the most important concept for inference in parametric modeling when the underlying data are subject to different problems and limitations related to medical and epidemiological studies, for example, in the context of the analysis of survival data. Likelihood-based testing, as we know, was mainly founded and formulated in a series of fundamental papers published in the period of 1928–1938 by Jerzy Neyman and Egon Pearson.[23–26] In 1928, the authors introduced the generalized likelihood ratio test and its association with chi-squared statistics. Five years later, the Neyman–Pearson Lemma was introduced, showing the optimality of the likelihood ratio test.[24] These seminal works provided us with the familiar notions of simple and composite hypotheses and errors of the first and second kinds, thus defining formal decision-making rules for testing. Without loss of generality, the principle idea of the proof of the Neyman–Pearson Lemma can be shown by using the trivial inequality

$$(A - B)(I\{A \geq B\} - \delta) \geq 0, \tag{31.1}$$

for all A, B, where $\delta \in [0, 1]$ and $I\{\cdot\}$ denotes the indicator function. For example, suppose we would like to classify independent identically distributed (i.i.d.) biomarker measurements $\{X_i, i = 1, \ldots, n\}$ corresponding to hypotheses of the following form: H_0: X_1 is from a density function f_0, versus H_1: X_1 is from a density function f_1. In this context, to construct the likelihood ratio test statistic, we should consider the ratio between the joint density function of $\{X_1, \ldots, X_n\}$ obtained under H_1 and the joint density function of $\{X_1, \ldots, X_n\}$ obtained under H_0, and then define $\prod_{i=1}^{n} f_1(X_i)/\prod_{i=1}^{n} f_0(X_i)$ as the likelihood ratio. In this case, the likelihood ratio test is uniformly most powerful. This proposition directly follows from the expected value under H_0 of the inequality (31.1), where we define $A = \prod_{i=1}^{n} f_1(X_i)/f_0(X_i)$, B to be a test threshold (i.e., the likelihood ratio test rejects H_0 if and only if $L \geq B$), and δ is assumed to represent any decision rule based on $\{X_i, i = 1, \ldots, n\}$. The Appendix contains details of the proof. This simple proof technique was used to show optimal

aspects of different statistical decision-making policies based on the likelihood ratio concept applied in clinical experiments.[27–29]

The likelihood ratio based on the likelihood ratio test statistic is the likelihood ratio test statistic

The Neyman–Pearson test concept, fixing the probability of a Type I error, comes under some criticism by epidemiologists. One of the critical points is about the importance of paying attention to Type II errors. For example, Freiman *et al.* pointed out results of 71 clinical trials that reported no "significant" differences between the compared treatments.[2] The authors found that in the great majority of these trials, the strong effects of new treatment were reasonable. The investigators in such trials inappropriately accepted the null hypothesis as correct, which probably resulted in Type II errors. In the context of likelihood ratio-based tests, we present the following result that demonstrates an association between the probabilities of Type I and II errors.

Suppose we would like to test for H_0 versus H_1, employing the likelihood ratio $L = f_{H_1}(D)/f_{H_0}(D)$ based on data D, where f_H defines a density function that corresponds to the data distribution under the hypothesis H. Say, for simplicity, we reject H_0 if $L > C$, where C is a presumed threshold. In this case, we can then show that

$$f^L_{H_1}(u) = u\, f^L_{H_0}(u), \tag{31.2}$$

where $f^L_H(u)$ is the density function of the test statistic L under the hypothesis H and $u > 0$. Details of the proof of this fact are shown in the Appendix. Thus, we can obtain the probability of a Type II error in the form of

$$\begin{aligned}\Pr\{\text{the test does not reject } H_0 | H_1 \text{ is true}\} &= \Pr\{L \le C |\, H_1 \text{ is true}\} = \int_0^C f^L_{H_1}(u)du \\ &= \int_0^C u f^L_{H_0}(u)du.\end{aligned}$$

Now, if, in order to control the Type I error, the density function $f^L_{H_0}(u)$ is assumed to be known, then the probability of the Type II error can be easily computed.

The likelihood ratio property $f^L_{H_1}(u)/f^L_{H_0}(u) = u$ can be applied to solve different issues related to performing the likelihood ratio test. For example, in terms of the bias of the test, one can request to find a value of the threshold C that maximizes

$$\Pr\{\text{the test rejects } H_0 | H_1 \text{ is true}\} - \Pr\{\text{the test rejects } H_0 | H_0 \text{ is true}\},$$

where the probability $\Pr\{\text{the test rejects } H_0 | H_1 \text{ is true}\}$ depicts the power of the test. This equation can be expressed as

$$\begin{aligned}\Pr\{L > C | H_1 \text{ is true}\} - \Pr\{L > C | H_0 \text{ is true}\} &= \left(1 - \int_0^C f^L_{H_1}(u)\,du\right) \\ &\quad - \left(1 - \int_0^C f^L_{H_0}(u)\,du\right).\end{aligned}$$

Let the derivative of this notation equal zero and solve the equation:

$$\frac{d}{dC}\left[\left(1-\int_{\infty}^{C} f_{H_1}^{L}(u)\,du\right)-\left(1-\int_{\infty}^{C} f_{H_0}^{L}(u)\,du\right)\right]=-f_{H_1}^{L}(C)+f_{H_0}^{L}(C)=0.$$

By virtue of the property (31.2), this implies $-Cf_{H_0}^{L}(C)+f_{H_0}^{L}(C)=0$ and then $C=1$, which provides the maximum discrimination between the power and the probability of a Type I error in the likelihood ratio test.

In other words, the interesting fact is that the likelihood ratio $f_{H_1}^{L}/f_{H_0}^{L}$ based on the likelihood ratio $L=f_{H_1}/f_{H_0}$ comes to be the likelihood ratio, that is, $f_{H_1}^{L}(L)/f_{H_0}^{L}(L)=L$. Interpretations of this statement in terms of information, we leave to the reader's imagination.

Exercise 31.1

Given a sample of i.i.d. measurements $X_1, \ldots, X_n$ following exponential distribution with the rate parameter λ, that is, $X_1, \ldots, X_n \sim f(x)=\lambda\exp(-\lambda x)$, derive the likelihood ratio test statistics for the simple hypothesis $H_0 : \lambda=1$ versus $H_1 : \lambda=2$.

Maximum likelihood; is it the likelihood?

Various real-world data problems require considerations of statistical hypotheses with structures that depend on unknown parameters. In this case, the maximum likelihood method proposes to approximate the most powerful likelihood ratio, employing a proportion of the maximum likelihoods, where the maximizations are over values of the unknown parameters belonging to distributions of observations under the corresponding hypotheses. We shall assume the existence of essential maximum likelihood estimators. The influential Wilks' theorem provides the basic rationale as to why the maximum likelihood ratio approach has had tremendous success in statistical applications.[30] Wilks showed that under regularity conditions, asymptotic null distributions of maximum likelihood ratio test statistics are independent of nuisance parameters. That is, a Type I error in the maximum likelihood ratio tests can be controlled asymptotically, and approximations of the corresponding p-values can be computed.

Thus, if certain key assumptions are met, one can show that parametric likelihood methods are very powerful and efficient statistical tools. We should emphasize that the role of the discovery of the likelihood ratio methodology in statistical developments can be compared to the development of the assembly line technique of mass production. The likelihood ratio principle gives clear instructions and technique manuals on how to construct efficient statistical decision rules in various complex problems related to clinical experiments. For example, Vexler *et al.* developed a likelihood ratio test for comparing populations based on incomplete longitudinal data subject to instrumental limitations.[31]

Although many statistical publications continue to contribute to the likelihood paradigm and are very important in the statistical discipline (an excellent account can be found in Lehmann and Romano[14]), several significant questions naturally arise about the maximum likelihood approach's general applicability. Conceptually,

there is an issue specific to classifying maximum likelihoods in terms of likelihoods that are given by joint density (or probability) functions based on data. Integrated likelihood functions, with respect to arguments related to data points, are equal to 1, whereas accordingly integrated maximum likelihood functions often have values that are indefinite. Thus, while likelihoods present full information regarding the data, the maximum likelihoods might lose information conditional on the observed data. Consider this simple example:

Suppose we observe X_1, which is assumed to be from a normal distribution $N(\mu, 1)$ with mean parameter μ. In this case, the likelihood has the form $(2\pi)^{-0.5}\exp(-(X_1-\mu)^2/2)$ and, correspondingly, $\int (2\pi)^{-0.5}\exp(-(X_1-\mu)^2/2)dX_1 = 1$, whereas the maximum likelihood, that is, the likelihood evaluated at the estimated μ, $\hat{\mu} = X_1$ is $(2\pi)^{-0.5}$, which clearly does not represent the data and is not a proper density. This demonstrates that since the Neyman–Pearson lemma is fundamentally founded on the use of the density-based constitutions of likelihood ratios, maximum likelihood ratios cannot be optimal in general. That is, the likelihood ratio principle is generally not robust when the hypothesis tests have corresponding nuisance parameters to consider, for example, testing a hypothesized mean given an unknown variance.

An additional inherent difficulty of the likelihood ratio test occurs when a clinical experiment is associated with an infinite-dimensional problem and the number of unknown parameters is relatively large. In this case, Wilks' theorem should be re-evaluated, and nonparametric approaches should be considered in the contexts of reasonable alternatives to the parametric likelihood methodology.[32]

The ideas of likelihood and maximum likelihood ratio testing may not be fiducial and applicable in general nonparametric function estimation/testing settings. It is also well known that when key assumptions are not met, parametric approaches may be suboptimal or biased as compared to their robust counterparts across the many features of statistical inferences. For example, in a biomedical application, Gosh proved that the maximum likelihood estimators for the Rasch model are inconsistent, as the number of nuisance parameters increases to infinity (Rasch models are often employed in clinical trials that deal with psychological measurements, e.g., abilities, attitudes, and personality traits).[33] Due to the structure of likelihood functions based on products of densities, or conditional density functions, relatively insignificant errors in classifications of data distributions can lead to vital problems related to the applications of likelihood ratio type tests.[34] Moreover, one can note that, given the wide variety and complex nature of biomedical data (e.g., incomplete data subject to instrumental limitations or complex correlation structures), parametric assumptions are rarely satisfied. The respective formal tests are complicated, or oftentimes not readily available.

Exercise 31.2

Given a sample of i.i.d. measurements $X_1, \ldots, X_n$, following an exponential distribution with the rate parameter λ, that is, $X_1, \ldots, X_n \sim f(x) = \lambda\exp(-\lambda x)$, derive the maximum likelihood ratio test statistic for the composite hypothesis $H_0 : \lambda = 1$ versus $H_1 : \lambda \neq 1$.

Tests on means of continuous data

Does a sample mean equal a pre-specified population mean, or, alternatively, do two or more samples have the same population mean? These questions can be answered by the hypothesis testing of equal means.

Likelihood ratio test for the mean of normally distributed data

Given a random sample of i.i.d. observations $X_1, \ldots, X_n$ from a normal population with the mean μ and the variance σ^2, we would like to test for the simple hypothesis

$$H_0 : \mu = \mu_0 \text{ versus } H_1 : \mu = \mu_1.$$

In this case, the likelihood function is

$$L = (\sigma\sqrt{2\pi})^{-n} \exp\left\{\sum_{i=1}^{n} (X_i - \mu)^2/(2\sigma^2)\right\}.$$

Therefore, the likelihood ratio has the form

$$\Lambda = \frac{(\sigma\sqrt{2\pi})^{-n} \exp\left\{-\sum_{i=1}^{n} (X_i - \mu_0)^2/(2\sigma^2)\right\}}{(\sigma\sqrt{2\pi})^{-n} \exp\left\{-\sum_{i=1}^{n} (X_i - \mu_1)^2/(2\sigma^2)\right\}}$$

$$= \exp\left\{\left(2(\mu_0 - \mu_1)\sum_{i=1}^{n} X_i - n(\mu_0^2 - \mu_1^2)\right)/(2\sigma^2)\right\}.$$

We reject H_0 if $\Lambda > C_\alpha$, where the constant C_α is selected for a specified value for the significance level α, the Type I error rate. To use this most powerful likelihood ratio test, we must know the values of σ^2, μ_0, and μ_1.

t-Type tests

As an example of the likelihood methodology, we present *t*-test-type decision rules, which are widely used in practice. Assuming the following constraints: (1) the ~~sample~~ data is a simple random sample from the population and each ~~sample~~ observation is independent of each other, and (2) the sample observations were drawn from a normal distribution, *t*-tests can be conducted to test means of continuous data. The one-sample *t*-test can be applied to test for the equality of the sample mean to a presumed value when the population variance is unknown or the sample size is small (no greater than 30 as a rule of thumb). To determine if two independent sets of data are significantly different from each other, the two-sample *t*-test can be applied. In the case of multivariate hypothesis testing, the multivariate *t* (Hotelling's *t*-squared) test, as a generalization of the Student's *t*-statistic, ~~should~~ can be used.

One-sample *t*-tests

In testing the null hypothesis that the population mean is equal to a specified value μ_0, ~~that is~~ i.e., $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, based on the observed data $X_1, \ldots, X_n$, one

uses the statistic

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}},$$

where n is the sample size, $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ is the sample mean, and s is the sample standard deviation. At the α significance level, the null hypothesis can be rejected if $|t| \geq t_{\alpha/2,n-1}$, where $t_{\alpha/2,n-1}$ is the $(1-\alpha/2)$th quantile of t distribution with $n-1$ degrees of freedom. Here, we define the pth quantile for a random variable as the value x, such that the probability that the random variable will be less than x is at most p and the probability that the random variable will be more than x is at least $1-p$. Note that the population does not need to be normally distributed for a large sample size (>30, as a rule of thumb). By the central limit theorem, the distribution of the population of sample means, $\overline{X}$, will be approximately normal for a sufficiently large sample size.[35]

Two-sample *t*-tests

The two-sample t-test is used to determine if two independent population means are equal, that is, $H_0 : \mu_1 = \mu_2$ versus $H_1 : \mu_1 \neq \mu_2$. Given two samples of i.i.d. observations $X_{i1}, \ldots, X_{in_i}$, $i = 1, 2$, we denote the sample size, the sample mean, and the unbiased estimator of the variance of the two samples as n_i, $\overline{X}_i = n_i^{-1}\sum_{j=1}^{n_i} X_{ij}$, and $s_i^2 = (n_i - 1)^{-1}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_i)^2$, $i = 1, 2$, respectively. The t-statistic to test whether the means are equal can be calculated as follows:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{s_d}.$$

Based on the equivalence of the population variance in two groups, the equal variances case and the unequal variances case are considered separately, and the estimate of s_d can be calculated accordingly.

Equal variances: When the two distributions are assumed to have the same variance, the estimator is $s_d = s_p\sqrt{n_1^{-1} + n_2^{-1}}$, where the pooled standard deviation $s_p = \sqrt{((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)/(n_1 + n_2 - 2)}$ is an estimator of the common standard deviation of the two samples. At the α significance level, the null hypothesis of equal means can be rejected if $|t| \geq t_{\alpha/2,n_1+n_2-2}$, where $t_{\alpha/2,n_1+n_2-2}$ is the $(1-\alpha/2)$th quantile of t distribution with $n_1 + n_2 - 2$ degrees of freedom. Note that s_p^2 is an unbiased estimator of the common variance whether the population means are the same or not.

Unequal variances (Welch's t-test): When the two population variances are not assumed to be equal, the estimator is $s_d = \sqrt{s_1^2/n_1 + s_2^2/n_2}$. Note that in this case, s_d^2 is not a pooled variance. At the α significance level, the null hypothesis of equal means can be rejected if $|t| \geq t_{\alpha/2,df}$, where $t_{\alpha/2,df}$ is the $(1-\alpha/2)$th quantile of t distribution with

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

degrees of freedom.

Paired *t*-tests

In clinical trials, the generalized treatment effect can be used to compare treatments or interventions based on the difference in mean outcomes between pre- and post-treatment measurements. In the case of one paired sample, paired *t*-tests can be conducted to test for a paired difference. Given a paired sample $X_{k1}, \ldots, X_{kn}$ of pre-treatment ($k = 1$) and post-treatment ($k = 2$) measurements, to test whether the difference μ_D in means between post- and pre-treatment measurements is μ_0, the t statistic is

$$t = \frac{\overline{X}_D - \mu_0}{s_D/\sqrt{n}},$$

where n is the number of pairs, $X_{Di} = X_{2i} - X_{1i}$, $\overline{X}_D = n^{-1}\sum_{i=1}^{n} X_{Di}$, and $s_D^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_{Di} - \overline{X}_D)^2$ is the sample mean and sample variance of differences between all pairs, respectively. At the α significance level, the null hypothesis can be rejected if $|t| \geq t_{\alpha/2,n-1}$, where $t_{\alpha/2,n-1}$ is the $(1-\alpha/2)$th quantile of t-distribution with $n-1$ degrees of freedom.

We exemplify the use of the paired *t*-test with a real-life example of the effect of asthma education on pediatric patients' acute care visits.[15]

Example 31.1

The study sample consists of 32 patients who satisfy inclusion criteria and present over a period of time. The number of acute care visits during a year is recorded. After a standardized course of asthma training, the number of acute care visits for the following year is recorded again. The change per patient, that is, the before-and-after difference in the number of visits, was 1, 1, 2, 4, 0, 5, −3, 0, 4, 2, 8, 1, 1, 0, −1, 3, 6, 3, 1, 2, 0, −1, 0,3 ,2, 1, 3, −1, −1, 1, 1, and 5. It is of interest to test if the training affects the number of visits.[15]

The following R code can be used to carry out the two-tailed test $H_0 : \mu_D = 0$ against $H_1 : \mu_D \neq 0$:

```
> # input the data: difference (before-after)
> D <-c(1,1,2,4,0,5,-3,0,4,2,8,1,1,0,-1,3,6,3,1,2,0,-1,0,3,2,
1,3,-1,-1,1,1,5)
> alpha <- 0.05 # pre-specified significance level
> # check the normality by the histogram
> hist(D,xlab="Difference",main="Histogram of before and after
difference")
>
> # calculate the test statistic
> n <- length(D)  # the sample size, i.e., the number of pairs
> t.stat <- (mean(D)-0)/(sd(D)/sqrt(n))
> t.stat
[1] 4.034031
>
> # obtain the critical value and the p-value
```

```
> crit <- qt(1-alpha/2,df=n-1)
> crit
[1] 2.039513
> pval <- 2*(1-pt(t.stat,df=n-1))  # a two-sided test
> pval
[1] 0.0003323025
```

Alternatively, one may use the built-in function `t.test` setting `paired=TRUE` and `alternative="two.sided"` in R to conduct the two-sided paired *t*test. It yields the following output:

```
> t.test(D,rep(0,n),paired=TRUE,alternative="two.sided")

 Paired t-test

data:  D and rep(0, n)
t = 4.034, df = 31, p-value = 0.0003323
alternative hypothesis: true differ-
ence in means is not equal to 0
95 percent confidence interval:
 0.818888 2.493612
sample estimates:
mean of the differences
                1.65625
```

The form of the histogram of the differences shown in Figure 31.3 suggests an approximately normal shape, satisfying the normal distribution assumption. The test

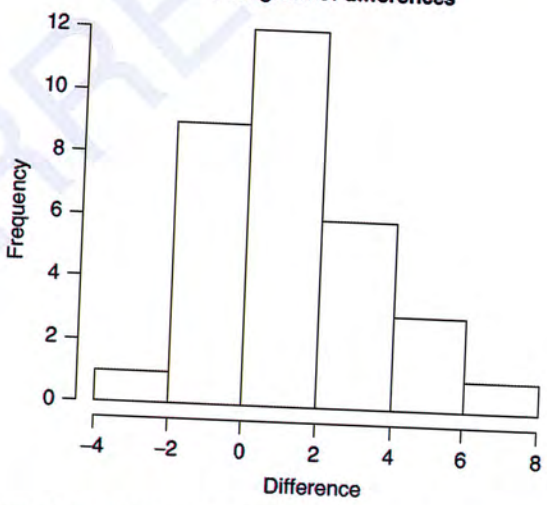

Figure 31.3 Histogram of the differences in the number of acute care visits pre- and post-asthma training.

statistic is $t = 4.034$, which is greater than the critical value $t_{\alpha/2,df=31} = 2.04$ at the $\alpha = 0.05$ significance level. We can state that we are 95% sure that the asthma training was efficacious.

Multivariate *t*-tests

Hypothesis testing of the equality of means can be constructed in the multivariate case, say, p-variate. Assuming that the samples are independently drawn from two independent multivariate normal distributions with the same covariance, that is, for the ith sample, $\mathbf{X}_{ij} \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i = 1, 2$, $j = 1, \ldots, n_i$, the hypothesis is $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$. Define $\overline{\mathbf{X}}_i = n_i^{-1} \sum_{j=1}^{n_i} \mathbf{X}_{ij}$, $i = 1, 2$, as the sample means and $\mathbf{W} = (n_1 + n_2 - 2)^{-1} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)^T$ as the unbiased pooled covariance matrix estimate, then the Hotelling's two-sample T^2 statistic is

$$t^2 = \frac{n_1 n_2}{n_1 + n_2} (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)^T \mathbf{W}^{-1} (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2).$$

Note that the Hotelling's two-sample T^2 statistic follows the Hotelling's T^2 distribution with parameters p and $n_1 + n_2 - 2$; that is, $t^2 \sim T^2(p, n_1 + n_2 - 2)$. The null hypothesis is rejected if $t^2 > T^2(\alpha; p, n_1 + n_2 - 2)$ at the α significance level.

In the following example, we will show how to simulate data in R using the function `mvrnorm` introduced in Section "R: statistical software" and conduct the Hotelling's two-sample T^2 test.

Example 31.2

We consider an example in which age and measurements of weight (in kg) are recorded for each patient in two independent groups. Assume that for the first group, $n_1 = 50$ patients are randomly sampled from the population $N_2\left(\begin{pmatrix}25\\65\end{pmatrix}, \begin{pmatrix}5 & 1\\1 & 9\end{pmatrix}\right)$, while for the second group, $n_2 = 70$ patients are randomly sampled from the population $N_2\left(\begin{pmatrix}25\\70\end{pmatrix}, \begin{pmatrix}5 & 1\\1 & 9\end{pmatrix}\right)$. We simulate the data and test if the means are equal.

For each group, we have bivariate normal data with a common population covariance matrix. The following R code can simulate bivariate normal data and conduct the Hotelling's two-sample T^2 test for the equality of means:

```
> # Check if packages are aleady installed.
> check.pkg <- c("ICSNP", "MASS") %in% rownames(installed
.packages())
> if(any(!check.pkg)) install.packages(c("ICSNP", "MASS")
[!check.pkg])
> # load packages
> library(ICSNP)
> library(MASS)
> # simulate data
> set.seed(123)
> n1 <- 50
```

```
> n2 <- 70
> Sigma <- matrix(c(5,1, 1,9), byrow=TRUE, ncol=2) # common
covariance matrix
> X1    <- mvrnorm (n1, mu=c(25, 65), Sigma=Sigma)
> X2  <- mvrnorm (n2, mu=c(25, 70), Sigma=Sigma)
> X <- rbind(X1, X2)
> Group <- factor(rep(1:2, c(n1,n2)))
> HotellingsT2(X ~ Group)
```

It yields the following output:

```
Hotelling's two sample T2-test

data:  X by Group
T.2 = 35.3461, df1 = 2, df2 = 117, p-value = 9.823e-13
alternative hypothesis: true location difference is not equal
to c(0,0)
```

With a p-value far less than 0.001, we reject the null hypothesis at the 0.05 significance level. We can conclude that we are 95% sure there is significant difference in the means of two bivariate data.

Exercise 31.3

Generate data from bivariate normal distributions based on $n_1 = 130$, $n_2 = 100$, $\boldsymbol{\mu}_1 = (50\ 89)^T$, $\boldsymbol{\mu}_2 = (45\ 92)^T$ and the common covariance matrix $\boldsymbol{\Sigma} = \begin{pmatrix} 15 & -2 \\ -2 & 9 \end{pmatrix}$. Conduct the Hotelling's two-sample T^2 test at the $\alpha = 0.05$ significance level.

The exact likelihood ratio test for equality of two normal populations

Testing the equality of two independent normal populations is of practical importance. Let $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_n\}$ be two independent random samples from normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. It is of interest to test

$$H_0 : \mu_1 = \mu_2 \text{ and } \sigma_1^2 = \sigma_2^2 \text{ versus } H_1 : \mu_1 \neq \mu_2 \text{ or } \sigma_1^2 \neq \sigma_2^2.$$

Pearson and Neyman[36] considered the likelihood ratio test

$$\lambda_{n,m} = \frac{\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2/n\right]^{n/2}\left[\sum_{j=1}^{m}\left(Y_j - \overline{Y}\right)^2/m\right]^{m/2}}{\left\{\left[\sum_{i=1}^{n}\left(X_i - u\right)^2 + \sum_{j=1}^{m}\left(Y_j - u\right)^2\right]/(n+m)\right\}^{(n+m)/2}},$$

where $\overline{X}$, $\overline{Y}$, and u are the sample means of the X sample, the Y sample, and the combined sample, respectively. For $\lambda \in (0, 1)$, Zhang *et al.*[37] derived the exact distribution of $\lambda_{n,m}$ as

$$\Pr(\lambda_{n,m} \le \lambda) = 1 - C \int\int_D w_1^{(n-1)/2-1} w_2^{(m-1)/2-1} \Big/ \sqrt{1 - w_1 - w_2}\, dw_1 dw_2$$

$$= 1 - C \int_{r_1}^{r_2} w_1^{(n-3)/2} \int_{z/w_1^{n/m}}^{1-w_1} w_2^{(m-3)/2} \Big/ \sqrt{1 - w_1 - w_2}\, dw_2 dw_1,$$

where $z = \dfrac{\lambda^{2/m} n^{n/m} m}{(n+m)^{(n+m)/m}}$, $C = \dfrac{\Gamma((n+m-1)/2)}{\Gamma((n-1)/2)\Gamma((m-1)/2)\Gamma(1/2)}$,

$D = \left\{ (w_1, w_2) : w_1 > 0, w_2 > 0, w_1 + w_2 < 1, w_1^{n/2} w_2^{m/2} (n+m)^{(n+m)/2} / (n^{n/2} m^{m/2}) > \lambda \right\}$,

and $r_1 < r_2$ are the two roots (for the variable w_1) of

$$1 - w_1 - z/w_1^{n/m} = 0.$$

Note that the double integral can be computed using Gaussian quadrature, implemented with the R function `plrt`.

Empirical likelihood

One very important issue is to preserve efficiency of the statistical inference through the use of robust likelihood-type methods, while concurrently minimizing assumptions about the underlying distribution. Toward this end, the recent biostatistical literature has shifted focus toward robust and efficient nonparametric and semiparametric developments of various "artificial" or "approximate" likelihood techniques. These methods have a wide variety of applications related to clinical experiments. Many nonparametric and semiparametric approximations to powerful parametric likelihood procedures have been used routinely in both statistical theory and practice. Well-known examples include the quasi-likelihood method, approximations of parametric likelihoods via orthogonal functions, techniques based on quadratic artificial likelihood functions, and the local maximum likelihood methodology.[38–41] Various studies have shown that artificial or approximate likelihood-based techniques efficiently incorporate information expressed through the data, and have many of the same asymptotic properties as those derived from the corresponding parametric likelihoods. The empirical likelihood (EL) method is one of a growing array of artificial or approximate likelihood-based methods currently in use in statistical practice.[42] Interest in and the resulting impact of EL methods continue to grow rapidly. Perhaps more importantly, EL methods now have various vital applications in a large and expanding number of areas of clinical studies.

A question of major interest to this section turns on the performance of EL constructs relative to ordinary parametric likelihood ratio-based procedures in the context of clinical experiments. Our desire to incorporate several recent developments and applications in these areas in an easy-to-use manner provides one of the main impetuses for this section. The EL method for testing has been dealt with extensively in the literature within a variety of settings.[42–47]

Classical empirical likelihood

As background for the development of EL-type techniques, we first outline the classical EL approach. The simple classical EL takes the form $\prod_{i=1}^{n}(F(X_i) - F(X_i-))$, which is a functional of the cumulative distribution function F and i.i.d. observations X_i, $i = 1, \ldots, n$. This EL technique is *"distribution function-based."* [42] In the distribution-free setting, an empirical estimator of this likelihood may take the form of $L_p = \prod_{i=1}^{n} p_i$, where the components p_i, $i = 1, \ldots, n$, the estimators of the probability weights, should maximize the likelihood L_p, provided that $\sum_{i=1}^{n} p_i = 1$ and empirical constraints based on $X_1, \ldots, X_n$ hold. For example, suppose we would like to test the hypothesis

$$H_0 : E(g(X_1, \theta)) = 0 \text{ versus } H_1 : E(g(X_1, \theta)) \neq 0,$$

where $g(.,.)$ is a given function and θ is a parameter. Then, in a nonparametric manner, we define the EL function of the form $\mathrm{EL}(\theta) = L(X_1, \ldots, X_n|\theta) = \prod_{i=1}^{n} p_i$, where $\sum_{i=1}^{n} p_i = 1$. Under the null hypothesis, the maximum likelihood approach requires one to find the values of the p_i that maximize the EL given the empirical constraints $\sum_{i=1}^{n} p_i = 1$ and $\sum_{i=1}^{n} p_i g(X_i, \theta) = 0$ that present an empirical version of the condition under H_0 that $E(g(X_1, \theta)) = 0$ (the null hypothesis is assumed to be rejected when there are no $0 < p_1, \ldots, p_n < 1$ to satisfy the empirical constraints). In this case, using Lagrange multipliers, one can show that

$$\mathrm{EL}(\theta) = \sup_{0<p_1,p_2,\ldots,p_n<1, \sum p_i=1, \sum p_i g(X_i,\theta)=0} \prod_{i=1}^{n} p_i = \prod_{i=1}^{n} (n + \lambda g(X_i, \theta))^{-1}, \tag{31.3}$$

where λ is a root of $\sum g(X_i, \theta)(n + \lambda g(X_i, \theta))^{-1} = 0$. Since under H_1, the only constraint under consideration is $\sum p_i = 1$, we have

$$\mathrm{EL} = \sup_{0<p_1,p_2,\ldots,p_n<1, \sum p_i=1} \prod_{i=1}^{n} p_i = \prod_{i=1}^{n} n^{-1} = (n)^{-n}. \tag{31.4}$$

Combining equations (31.3) and (31.4), we obtain the EL ratio (ELR) test statistic $\mathrm{ELR}(\theta) = \mathrm{EL}/\mathrm{EL}(\theta)$ for the hypothesis test of H_0 versus H_1. For example, when the function $g(u, \theta) = u - \theta$, the null hypothesis corresponds to the expectation.

Owen showed that the nonparametric test statistic $2 \log \mathrm{ELR}(\theta)$ has an asymptotic chi-square distribution under the null hypothesis.[42] This result illustrates that Wilks' theorem-type results continue to hold in the context of this infinite-dimensional problem. Consequently, there are techniques for correcting forms of ELRs to improve the convergence rate of the null distributions of ELR test statistics to chi-square distributions. These techniques are similar to those applied in the field of parametric maximum likelihood ratio procedures.[45] The statement of the hypothesis testing above can easily be inverted with respect to providing nonparametric confidence interval estimators.

In terms of the accessibility of this method, it should be noted that the number of EL software packages continues to expand, particularly the R software packages. For example, the links to `library(emplik)` and `library(EL)` R packages that include the R function `el.test()` and `EL.test()`. These simple R functions can be very useful for the EL analysis of data from clinical studies.

For illustrative example, we revisit the HDL cholesterol data shown in Figure 31.2. Now, we use the empirical likelihood ratio test for means. The following R output

shows the result of the empirical likelihood comparison between the means of the groups: X and Y.

```
> library(EL)
> EL.means(X,Y)

 Empirical likelihood mean difference test

data:  X  and  Y
-2 * LogLikelihood = 3.547, p-value = 0.05965
95 percent confidence interval:
 -0.4900842 19.0138090
sample estimates:
Mean difference
       10.17393
```

Perhaps, in this example, the ELR test outperforms the t-test that claims to reject the hypothesis $E(X) = E(Y)$, when X and Y are the measurements related to the same group of patients.

The classical EL methodology has been shown to have properties that make it attractive for testing hypotheses regarding parameters (e.g., moments) of distributions.[34,43,48] However, practicing statisticians working on clinical experiments, for example, case–control studies, commonly face a variety of distribution-free comparisons and/or evaluations over all distribution functions of complete and incomplete data subject to different types of measurement errors. In this framework, the *density-based* empirical likelihood methodology figures prominently.

Exercise 31.4

Generate a sample of i.i.d. measurements $X_1, \ldots, X_{25}$ from the following distributions: (1) normal distribution $N(0,1)$, and (2) $F(x) = (1 - \lambda \exp(-\lambda(x+1)))I\{x+1 > 0\}$, where λ is the rate parameter. Test $H_0 : E(X) = 0$ versus $H_1 : E(X) \neq 0$ at the $\alpha = 0.05$ significance level. In the latter case, is the result of the t-test valid? Explain the reason.

Density-based empirical likelihood

According to the Neyman–Pearson lemma, density-based likelihood ratios can provide uniformly most powerful tests. Using this as a starting point, Vexler *et al.* proposed an alternative to the "*distribution function-based*" EL methodology.[49] The authors employed the approximate density-based likelihood, which has the following form:

$$L_f = \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} f_i, \quad f_i = f(X_{(i)}),$$

where $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are the order statistics based on $X_1, \ldots, X_n$, and $f_1, \ldots, f_n$ take on the values that maximize L_f given the empirical constraint corresponding to

Table 31.2 Comparison of the classical EL and density-based EL approaches.

Characteristics	Owen[48]; Yu *et al.*[54]	The density-based EL method
Construction of the likelihood function	Distribution based	Density based
Usage of Lagrange multipliers method	Yes	Yes
Usage of constraints for maximization	Yes	Yes
Common focus of the test	Parameters (e.g., moments)	Overall distributions
Critical value	Asymptotic	Exact
The form of the test statistic	Numeric approach is required to calculate values of Lagrange multipliers	No numeric approach

$\int f(u)du = 1$. This density-based EL approach was used successfully in order to construct efficient entropy-based goodness-of-fit test procedures.[29,50] The density-based EL methodology has been satisfactorily applied to develop a test for symmetry based on paired data. This test significantly outperforms classical procedures.[49] Gurevich and Vexler extended the density-based EL approach to a two-sample nonparametric likelihood ratio test.[51] Vexler and Yu used the density-based EL concept to present two group comparison principles based on bivariate data with a missing pattern as a consequence of data collection procedures.[52] Furthermore, the density-based EL methods were used to efficiently address nonparametric problems of complex composite hypothesis testing in children, in social/behavioral studies based on randomized prospective experiments.[53] In many practical settings, the density-based ELRs can provide simple and exact tests. Some distinctive characteristics of the density-based EL method test statistic as compared to the typical EL approach are summarized in Table 31.2.[48,54]

We note that Table 31.2 cannot correspond to all relevant EL constructions. For example, Hall and Owen developed large-sample methods for constructing *"distribution function-based"* EL confidence bands in problems of nonparametric density estimation.[55] Einmahl and McKeague proposed to localize the *"distribution function-based"* EL approach using one or more "time" variables implicit in the given null hypothesis.[56] Integrating the log-likelihood ratio over those variables, the authors constructed exact-test procedures for detecting a change in distribution, testing for symmetry about zero, testing for exponentiality, and testing for independence.

It is a common practice to conduct medical trials in order to compare a new therapy with a standard of care based on paired data consisting of pre- and post-treatment measurements. In such cases, there is often great interest in identifying treatment effects within each therapy group, as well as detecting a between-group difference. Nonparametric comparisons between distributions of new therapy and control groups, as well as detecting treatment effects within each group, may be based on multiple-hypothesis tests. To this end, one can create relevant tests combining, for example, the Kolmogorov–Smirnov test and the Wilcoxon signed-rank test. The

use of the classical procedures commonly requires complex considerations about combining the known nonparametric tests and preserving the Type I error control and reasonable power of the resulting test. Alternatively, the density-based ELR technique provides a direct distribution-free approach for efficiently analyzing a variety of tasks occurring in clinical trials. The density-based EL method can easily be applied to test nonparametrically for different composite hypotheses. In this case, the density-based EL approach implies a standard scheme to develop highly efficient procedures, approximating nonparametrically the most powerful Neyman–Pearson test rules, given the aims of clinical studies. For example, Vexler *et al.* developed a density-based ELR methodology that was efficiently used to compare two therapy strategies for treating children's attention-deficit/hyperactivity disorder and severe mood dysregulation.[53] It was demonstrated that various composite hypotheses in a paired data setting (e.g., before vs. after treatment) can be tested with the density-based ELR tests, which give more emphasis to the overall distributional difference rather than to certain location parameter differences.

The R software can be employed in order to implement a computer program that realizes a density-based EL strategy. For example, programs of this type are presented in the *Statistics in Medicine* journal's Web domain http://onlinelibrary.wiley.com/doi/10.1002/sim.4467/suppinfo.Miecznikowski, Vexler, and Shepherd developed the R package "dbEmpLikeGOF" for nonparametric density-based likelihood ratio tests for goodness of fit and two-sample comparisons.[58] See also http://cran.r-project.org/web/packages/dbEmpLikeNorm/ for the R package "dbEmpLikeNorm: Test for joint assessment of normality," developed by Drs. Shepherd, Tsai, Vexler, and Miecznikowski. The group of coauthors Tanajian, Vexler, and Hutson presented a package entitled "Novel and efficient density-based empirical likelihood procedures for symmetry and K-sample comparisons" in STATA, a general-purpose statistical software language.[59] It is available over the web at http://sphhp.buffalo.edu/biostatistics/research-and-facilities/software/stata.html.

In order to exemplify the density-based empirical likelihood method, we employ data from the clinical study that is mentioned in section "Likelihood." This study was designed as a case–control study of biomarkers for coronary heart disease. In accordance with the biomedical literature, the HDL biomarker has been suggested as having strong discriminatory ability for myocardial infarction (MI). To define cases, we consider the sample Y that consists of 25 measurements of the HDL biomarker on individuals who recently survived an MI. In order to represent controls, 25 HDL biomarker measurements on healthy subjects are denoted as $X_1, \ldots, X_{25}$. The following R code inputs the data and constructs the histograms of the data, as shown in Figure 31.4:

```
> X<-c(96.8,57.2,37.4,44.0,55.0,41.8,46.2,41.8,41.8,59.4,44.0,
52.8,33.0,52.8,41.8,44.0,52.8,59.4,37.4,77.0,39.6,57.2,57.2,
41.8,39.6)
> Y<-c(26.4,33.0,30.8,35.2,44.0,48.4,61.6,41.8,26.4,28.6,55.0,
61.6,63.8,24.2,37.4,48.4,52.8,46.2,57.2,68.2,46.2,37.4,46.2,
52.8,35.2)
> a<-min(c(X,Y))-20
> b<-max(c(X,Y))+20
> par(pty="s",mfrow=c(1,2),oma=c(0,0,0,0),mar=c(0,4,0,0))
```

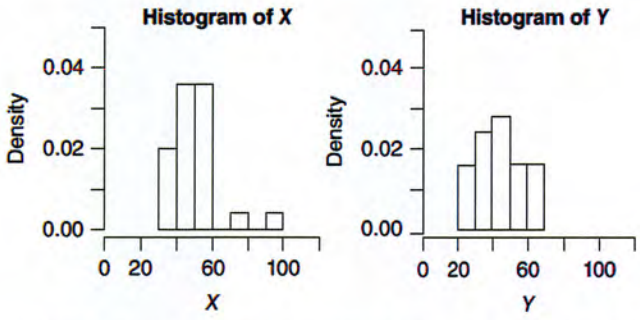

Figure 31.4 R data analysis output for measurements of HDL cholesterol levels (mg/dl) X and Y in ~~individuals with~~ the disease and healthy individuals, respectively.

```
> hist(X,xlim=c(a,b),ylim=c(0,0.05),freq=FALSE)
> hist(Y,xlim=c(a,b),ylim=c(0,0.05),freq=FALSE)
```

The classical empirical likelihood ratio test can be conducted via the R function `EL.means`. With a p-value of 0.101 as shown below, we fail to reject that $E(X) = E(Y)$ at the 0.05 significance level.

```
> EL.means(X,Y)

 Empirical likelihood mean difference test

data:  X  and  Y
-2 * LogLikelihood = 2.6898, p-value = 0.101
95 percent confidence interval:
 -1.066513 13.850197
sample estimates:
Mean difference
       5.720065
```

Thus, in this example, the ELR test cannot be used to demonstrate the discriminatory ability of the HDL biomarker with respect to the MI disease. In this case, the two-sample density-based empirical likelihood ratio test[51] shows the p-value <0.043, supporting rejection of the hypothesis regarding equivalency of distributions of X and Y. For the sake of completeness, the Appendix presents an example of R procedures for executing the two-sample density-based ELR test. In addition, Vexler *et al.* proposed a simple, but very efficient, density-based empirical likelihood ratio test for independence and provided the R code to run the procedure.[60]

Combinations of likelihoods to assemble composite tests and archive full information regarding data

Strictly speaking, "distribution-function/density-based" EL techniques and parametric likelihood methods are closely related concepts. This provides the impetus for an

impressive expansion in the number of EL developments, based on combinations of likelihoods of different types.[61]

Consider a simple example, where we assume to observe independent couples given as (X, Y). In this case, the likelihood function can be denoted as $L(X, Y)$. Suppose the observations for the X's are observed completely, whereas a proportion of the observed data for the Y's is incomplete. Assume a model of Y given X, that is, $Y|X$, is well defined, for example, $Y_i = \beta X_i + \varepsilon_i$, where β denotes the model parameter and ε_i is a normally distributed error term, for $i = 1, \ldots, n$. Then, we refer to Bayes' theorem to represent $L(X, Y) = L(Y|X)L(X)$, where $L(X)$ can be substituted by the EL to avoid parametric assumptions regarding distributions of X's.

In this context, Qin shows an inference on incomplete bivariate data using a method that combines the parametric model and ELs.[62] This method also incorporates auxiliary information from variables in the form of constraints, which can be obtained from reliable resources such as census reports. This approach makes it possible to use all available bivariate data, whether completely or incompletely observed. In the context of a group comparison, constraints can be formed based on null and alternative hypotheses, and these constraints are incorporated into the EL. This result was extended and applied to the following practical issues:

Malaria remains a major epidemiological problem in many developing countries. In endemic areas, an individual may have symptoms attributable either to malaria or to other causes. From a clinical viewpoint, it is important to attend to the next tasks: (i) to correctly diagnose an individual who has developed symptoms, so that the appropriate treatments can be given; (ii) to determine the proportion of malaria-affected cases in individuals who have symptoms, so that policies on intervention program can be developed. Once symptoms have developed in an individual, the diagnosis of malaria can be based on the analysis of the parasite levels in blood samples. However, even a blood test is not conclusive, as in endemic areas many healthy individuals can have parasites in their blood slides. Therefore, data from this type of study can be viewed as coming from a mixture distribution, with the components corresponding to malaria and nonmalaria cases. Qin and Leung constructed new EL procedures to estimate the proportion of clinical malaria using parasite-level data from a group of individuals with symptoms attributable to malaria.[63] Yu *et al.* and Vexler *et al.* proposed two-sample EL techniques based on incomplete data to analyze a Pneumonia Risk Study in an ICU Setting.[46,47] In the context of this study, the initial detection of ventilator-associated pneumonia (VAP) for inpatients at an intensive care unit requires composite symptom evaluation, using clinical criteria such as the clinical pulmonary infection score (CPIS). When CPIS is above a threshold value, bronchoalveolar lavage (BAL) is performed to confirm the diagnosis by counting actual bacterial pathogens. Thus, CPIS and BAL results are closely related, and both are important indicators of pneumonia, whereas BAL data are incomplete. Yu *et al.* and Vexler *et al.* derived EL methods to compare the pneumonia risks among treatment groups for such incomplete data.[46,47] In semi- and nonparametric contexts, including EL settings, Qin and Zhang showed that the full likelihood can be decomposed into the product of a conditional likelihood and a marginal likelihood, in a similar manner to the parametric likelihood considerations.[61] These techniques augment the study's power by enabling researchers to use any observed data and relevant information.

Receiver operating characteristic curve analysis

The ROC curves are useful visualization tools for illustrating the discriminant ability of biomarkers to distinguish between two populations: diseased and nondiseased. The ROC curve methodology was originally developed for radar signal detection theory and was extensively employed in psychological and, most importantly, medical research and epidemiology.

Assume, without loss of generality, that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are measurements from the diseased and nondiseased populations, respectively. The observations $X_1, \ldots, X_n$ are i.i.d. and independent of i.i.d. measurements $Y_1, \ldots, Y_m$. Let F and G denote the CDFs of X and Y, respectively. The ROC curve $R(t)$ can be defined as $R(t) = 1 - F(G^{-1}(1-t))$, where $t \in [0, 1]$.[65] It plots sensitivity (true positive rate, $1 - F(t)$) against 1 minus specificity (true negative rate, $1 - G(t)$) for various values of the threshold t. As an example, we consider three biomarkers with their corresponding ROC curves presented in Figure 31.5, whose underlying distributions are $F_1 \sim N(0,1)$, $G_1 \sim N(0,1)$ for biomarker A (the diagonal line), $F_2 \sim N(0,1)$, $G_2 \sim N(1,1)$ for biomarker B (in a dashed line), and $F_3 \sim N(0,1)$, $G_3 \sim N(10,1)$ for biomarker C (in a dotted line), respectively.

The following R code plots the ROC curve, as shown Figure 31.5.

```
> t<-seq(0,1,0.001)
> R1<-1-pnorm(qnorm(1-t,0,1),0,1)   # biomarker 1
> R2<-1-pnorm(qnorm(1-t,1,1),0,1)   # biomarker 2
> R3<-1-pnorm(qnorm(1-t,10,1),0,1)  # biomarker 3
```

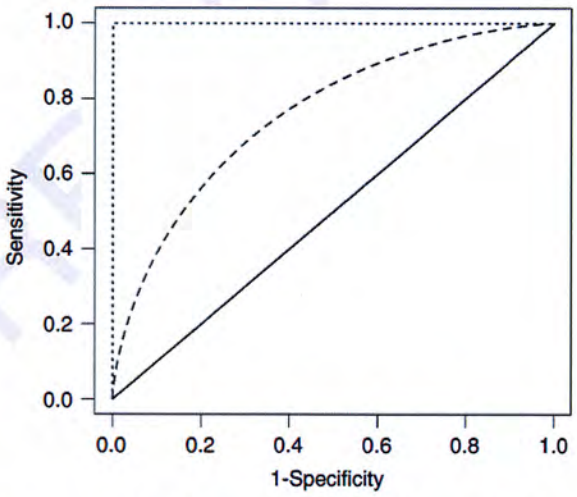

Figure 31.5 ROC curves related to the biomarkers. The solid diagonal line corresponds to the ROC curve of biomarker *A*, where $F_1 \sim N(0,1)$ and $G_1 \sim N(0,1)$. The dashed line displays the ROC curve of biomarker *B*, where $F_2 \sim N(0,1)$ and $G_2 \sim N(1,1)$. The dotted line close to the upper left corner plots the ROC curve for biomarker *C*, where $F_3 \sim N(0,1)$ and $G_3 \sim N(10,1)$.

```
> plot(R1,t,type="l",lwd=1.5,lty=1,cex.lab=1.1,ylab=
"Sensitivity",xlab="1-Specificity")
> lines(R2,t,lwd=1.5,lty=2)
> lines(R3,t,lwd=1.5,lty=3)
```

It can be seen that the farther apart the two distributions F and G fall, the more the ROC curve curves up to the top left corner. A perfect biomarker would have the ROC curve come close to the top left corner, and a biomarker without discriminability would result in a diagonal line in the ROC curve. We also observe that there exists a trade-off between specificity and sensitivity.

There exists extensive research on estimating the ROC curves from the parametric and nonparametric perspectives.[65–67] Assuming both the diseased and nondiseased populations are normally distributed, that is, $F \sim N(\mu_1, \sigma_1^2)$ and $G \sim N(\mu_2, \sigma_2^2)$, the corresponding ROC curve can be expressed as

$$\mathrm{ROC}(t) = \Phi[a + b\Phi^{-1}(t)],$$

where $a = (\mu_1 - \mu_2)/\sigma_1$, $b = \sigma_2/\sigma_1$, and Φ is the standard normal CDF. The estimated ROC curve is obtained by substituting the maximum likelihood estimators (MLEs) of the normal parameters μ_1, μ_2, σ_1, and σ_2 into the formula. The nonparametric estimate of the ROC curve used the empirical distribution functions.[66,67] Define the empirical distribution function of F as

$$\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le t\},$$

where $I\{\cdot\}$ denotes the indicator function and the empirical distribution function $\hat{G}_m$ of G can be defined similarly. Substituting F and G by corresponding empirical estimates $\hat{F}_n$ and $\hat{G}_m$, respectively, the empirical estimator of the ROC curve is given by

$$\hat{R}(t) = 1 - \hat{F}_n(\hat{G}_m^{-1}(1-t)),$$

which converges to $R(t)$ for a large sample.[66]

Figure 31.6 presents the nonparametric estimators of the ROC curves with a sample size 1000 for three biomarkers described in Figure 31.5, that is, $F_1 \sim N(0,1)$, $G_1 \sim N(0,1)$ for biomarker A (the diagonal line), $F_2 \sim N(0,1)$, $G_2 \sim N(1,1)$ for biomarker B (in a dashed line), and $F_3 \sim N(0,1)$, $G_3 \sim N(10,1)$ for biomarker C (in a dotted line), respectively, using the following R code:

```
> if(!("pROC" %in% rownames(installed.packages()))) install
.packages("pROC")
> library(pROC)
> n<-1000
> set.seed(123)  # set the seed
> # Simulate data from the normal distribution
> X1<-rnorm(n,0,1)
> Y1<-rnorm(n,1,1)
> group<-cbind(rep(1,n),rep(0,n))
> measures<-c(X1,Y1)
```

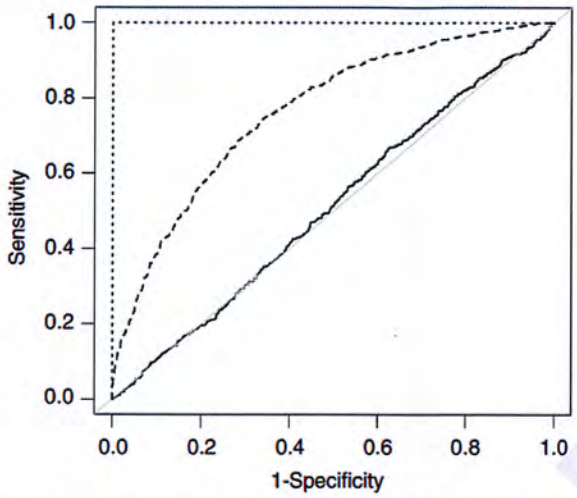

Figure 31.6 The nonparametric estimators of ROC curves of three different biomarkers based on samples of sizes 1000. The solid diagonal line corresponds to the nonparametric estimator of the ROC curve of biomarker *A*, where $F_1 \sim N(0, 1)$ and $G_1 \sim N(0, 1)$. The dashed line displays the nonparametric estimator of the ROC curve of biomarker *B*, where $F_2 \sim N(0, 1)$ and $G_2 \sim N(1, 1)$. The dotted line close to the upper left corner plots the nonparametric estimator of the ROC curve for biomarker *C*, where $F_3 \sim N(0, 1)$ and $G_3 \sim N(10, 1)$.

```
> roc1<-roc(group, measures)
> plot(1-roc1$specificities,roc1$sensitivities,type="l",
ylab="Sensitivity",xlab="1-Specificity")
> abline(a=0,b=1,col="grey")  # add the diagonal line for
reference
```

It should be noted that for large sample sizes, the ROC curves are well approximated by the nonparametric estimators.

In health-related studies, the ROC curve methodology is commonly related to case–control studies. As a type of observational study, case–control studies differentiate and compare two existing groups differing in outcome on the basis of some supposed causal attribute. For example, based on factors that may contribute to a medical condition, subjects can be grouped as the cases, for patients with condition/disease, and the controls, for patients without the condition/disease. For independent populations, for example, cases and controls, various parametric and nonparametric approaches have been proposed to evaluate the performance of biomarkers.[65–70]

Area under the ROC curve

A rough idea of the performance of the biomarkers can be obtained with the ROC curve. However, judgments solely based on the ROC curves are far from enough to precisely describe the diagnostic accuracy of biomarkers. The area under the ROC curve (AUC) is a common index of the diagnostic performance of a continuous biomarker

biomarker. It measures the ability to discriminate between the control and the disease groups.[68,69] Bamber noted that the area under this curve is equal to $Pr(X \geq Y)$.[70] We prove this result in the Appendix. Values of AUCs can range from 0.5, in the case of no difference between distributions, to 1, where the two distributions are perfectly discriminated. For more details, see Kotz *et al.* for wide discussions regarding evaluations of the AUC-type objectives.[71]

Parametric approach for AUC testing

Under the normal assumptions, a closed form of the AUC is presented as

$$A = \Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right),$$

where for the diseased population $X \sim N(\mu_1, \sigma_1^2)$ and for the nondiseased population $Y \sim N(\mu_2, \sigma_2^2)$, $\mu_1 \geq \mu_2$.[67,71] We provide the proof in the Appendix. By substituting maximum likelihood estimators for μ_i and σ_i^2, $i = 1, 2$ into the above formula, the maximum likelihood estimator of the AUC can be obtained correspondingly. Given the estimator of the AUC under the normal distribution assumption, one can easily construct confidence interval-based tests for the AUC using the delta method; see Kotz *et al.* for details.[71] In several cases, to achieve a fit between data distributions and the normal assumptions, a transformation of observations, for example, the Box–Cox transformation, can be recommended, before the above parametric approach is applied.[73] In general, when data distributions are different from the normal distribution function, the AUC evaluated that the given forms of F and G can be expressed as $\Pr(X > Y) = \int G(x)dF(x)$, in a similar manner to the technique above.[72]

Exercise 31.5

Biomarker levels were measured from diseased and healthy populations, providing i.i.d. observations $X_1 = 0.39$, $X_2 = 1.97$, $X_3 = 1.03$, $X_4 = 0.16$, which are assumed to be from a normal distribution, as well as i.i.d. observations $Y_1 = 0.42$, $Y_2 = 0.29$, $Y_3 = 0.56$, $Y_4 = -0.68$, $Y_5 = -0.54$, which are also assumed to be from a normal distribution, respectively. Please define the ROC curve. Obtain a formal notation of the AUC and estimate the AUC. What can be concluded about the discriminating ability of the biomarker with respect to the disease?

Hint: Values that may help you to approximate the estimated AUC: $\Pr(\xi < x) \approx 0.56$, when $x = 1$, $\xi = N(0.7, 4)$; $\Pr(\xi < x) \approx 0.18$, when $x = 1$, $\xi = N(0.9, 1)$; $Pr(\xi < x) = 0.21$, when $x = 0$, $\xi = N(0.8, 1)$; $\Pr(\xi < x) \approx 0.18$, when $x = 0$, $\xi = N(0.9, 1)$.

Nonparametric approach for AUC testing

Conversely, a nonparametric estimator for the AUC can be obtained as

$$\hat{A} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} I(X_i > Y_j),$$

where X_i, $i = 1, \ldots, n$ and Y_j, $j = 1, \ldots, m$ are the observations for diseased and nondiseased populations, respectively.[74] It is equivalent to the well-known Mann–Whitney statistic, and the variance of this empirical estimator can be obtained using U-statistic theory.[75] The empirical likelihood method to construct the confidence interval estimation for the AUC was introduced by Qin and Zhou.[76] Replacing the indicator function by a kernel function, one can obtain a smoothed ROC curve.[77]

Exercise 31.6

Biomarker levels were measured from diseased and healthy populations, providing i.i.d. observations $X_1 = 0.39$, $X_2 = 1.97$, $X_3 = 1.03$, $X_4 = 0.16$, which are assumed to be from a continuous distribution, as well as i.i.d. observations $Y_1 = 0.42$, $Y_2 = 0.29$, $Y_3 = 0.56$, $Y_4 = -0.68$, $Y_5 = -0.54$, which are also assumed to be from a continuous distribution, respectively. Please define the ROC curve. Estimate the AUC nonparametrically. What can be concluded regarding the discriminating ability of the biomarker with respect to the disease?

Nonparametric comparison between two ROC curves

It is of great importance for the researchers to compare two biomarkers. If we use both diagnostic markers on the same m controls and n cases, we can represent the bivariate outcomes as (X_{1j}, X_{2j}) $(j = 1, \ldots, m)$ and (Y_{1k}, Y_{2k}) $(k = 1, \ldots, n)$, respectively. We denote the respective bivariate distributions by $F(x_1, x_2)$ and $G(y_1, y_2)$, and the marginals by $F_i(x_i)$ and $G_i(y_i)$, $i = 1, 2$, and we assume that the $m + n$ bivariate vectors are mutually independent. Denote the sensitivity at specificity p by $S_i(p)$, $i = 1, 2$, and define

$$\Delta = \int (S_1 p) - S_2(p)) dW(p),$$

where W is a probability measure on the open unit interval. The parameter Δ allows one to compare sensitivities on a predefined range of specificities of clinical interest by adjusting the weight function W accordingly. When $W(p)$ is the uniform distribution on $(0, 1)$, the parameter Δ equals the difference of AUCs between two biomarkers.

Wieand *et al.* considered a nonparametric estimate of Δ in the form of

$$\hat{\Delta} = \int (\hat{S}_1 p) - \hat{S}_2(p)) dW(p),$$

where $\hat{S}_i(p) = 1 - \hat{G}_i(\hat{\xi}_{ip})$, $\hat{G}_i$ is the empirical distribution of G_i and the sample quantile $\hat{\xi}_{ip}$ is the $[mp]$th order statistic among the m values of X_i, where $[mp]$ is the smallest integer that equals or exceeds mp.[67] Assume that W is a probability measure in $(0, 1)$ and that there exists $\varepsilon > 0$ such that W has a bounded derivative in $(0, \varepsilon)$ and $(1 - \varepsilon, 1)$. Suppose further that $G_i(\xi_{ip})$, for $i = 1, 2$, has continuous derivatives in $(0, 1)$, which are monotone in $(0, \varepsilon)$ and $(1 - \varepsilon, 1)$. Define $s_i(p) = S_i'(p) = -G_i'(\xi_{ip})/F_i'(\xi_{ip})$. Then, as $N = n + m$ tends to ∞ with $m/N \to \lambda$, for $0 < \lambda < 1$, $N^{1/2}(\hat{\Delta} - \Delta)$ tends to a normal distribution with variance $\sigma^2 = \sigma^{11} - 2\sigma^{12} + \sigma^{22}$, where

$$\sigma_{ii} = \int_0^1 \int_0^1 \{(1 - \lambda)^{-1} S_i(\max(p, q))(1 - S_i(\min(p, q))) + \lambda^{-1} s_i(p) s_i(q)(\min(p, q) - pq)\} dW(p) dW(q),$$

and

$$\sigma_{12} = (1-\lambda)^{-1}\int\int (G(\xi_{1p},\xi_{2q}) - (1-S_1(p))(1-S_2(q))dW(p)dW(q)$$
$$+ \lambda^{-1}\int\int (G(\xi_{1p},\xi_{2q}) - pq)s_1(p)s_2(q)dW(p)dW(q).$$

Based on this asymptotic distribution of $\hat{\Delta}$, one can conduct a nonparametric procedure for testing $H_0 : \Delta = 0$ versus $H_1 : \Delta > 0$.[67] Note that the test proposed by Wieand *et al.* requires the estimation of densities, and selection of a satisfactory smoothing parameter may be problematic.

Best linear combinations based on values of multiple biomarkers

In practice, different markers are usually related to the disease, showing treatment effects in various magnitudes and different directions. For example, low levels of high-density lipoprotein (HDL)-cholesterol and high levels of thiobarbuturic acid reacting substances (TBARS), biomarkers of oxidative stress and antioxidant status, are indicators of coronary heart disease.[22] When multiple biomarkers are available, it is of great interest to seek a simple best linear combination (BLC) of biomarkers, such that the combined score achieves the maximum AUC or the maximum treatment effect over all possible linear combinations. Consider a study with d continuous-scale biomarkers yielding measurements $\mathbf{X}_i = (X_{1i}, \ldots, X_{di})^T$, $i = 1, \ldots, n$, on n diseased patients, and measurements $\mathbf{Y}_j = (Y_{1j}, \ldots, Y_{dj})^T$, $j = 1, \ldots, m$, on m nondiseased patients, respectively. It is of interest to construct effective one-dimensional combined scores of biomarkers' measurements, that is, $X(\mathbf{a}) = \mathbf{a}^T\mathbf{X}$ and $Y(\mathbf{a}) = \mathbf{a}^T\mathbf{Y}$, such that the AUC based on these scores is maximized over all possible linear combinations of biomarkers. Define $A(\mathbf{a}) = \Pr(X(\mathbf{a}) > Y(\mathbf{a}))$; the statistical problem is to estimate the maximum AUC defined as $A = A(\mathbf{a}_0)$, where $\mathbf{a}_0$ are the BLC coefficients satisfying $\mathbf{a}_0 = \arg\max_{\mathbf{a}} A(\mathbf{a})$. For simplicity, we assume that the first component of the vector $\mathbf{a}$ equals 1.[78] For example, in the case of two biomarkers, that is, $d = 2$, the AUC can be defined as $A(a) = \Pr(X_1 + aX_2 > Y_1 + aY_2)$.

Parametric method

Assuming $\mathbf{X}_i \sim N(\boldsymbol{\mu}_\mathbf{X}, \boldsymbol{\Sigma}_\mathbf{X})$, $i = 1, \ldots, n$ and $\mathbf{Y}_j \sim N(\boldsymbol{\mu}_\mathbf{Y}, \boldsymbol{\Sigma}_\mathbf{Y})$, $j = 1, \ldots, m$, Su and Liu derived the BLC coefficients $\mathbf{a}_0 \propto \boldsymbol{\Sigma}_C^{-1}\boldsymbol{\mu}$ and the corresponding optimal AUC as $\Phi(\omega^{1/2})$, where $\boldsymbol{\mu} = \boldsymbol{\mu}_\mathbf{X} - \boldsymbol{\mu}_\mathbf{Y}$, $\boldsymbol{\Sigma}_C = \boldsymbol{\Sigma}_\mathbf{X} + \boldsymbol{\Sigma}_\mathbf{Y}$, $\omega = \boldsymbol{\mu}^T\boldsymbol{\Sigma}_C^{-1}\boldsymbol{\mu}$, and Φ is the standard normal CDF.[79]

Based on Su and Liu's point estimator, we can derive the confidence interval estimation for the BLC-based AUC under multivariate normality assumptions.[80]

Exercise 31.7

Consider the simple bivariate normal case where $\mathbf{X}_i \sim N\left(\begin{pmatrix}\mu_{X_1}\\ \mu_{X_2}\end{pmatrix}, \begin{pmatrix}1 & \rho_X\\ \rho_X & 1\end{pmatrix}\right)$, $i = 1, \ldots, n$ and $\mathbf{Y}_j \sim N\left(\begin{pmatrix}\mu_{Y_1}\\ \mu_{Y_2}\end{pmatrix}, \begin{pmatrix}1 & \rho_Y\\ \rho_Y & 1\end{pmatrix}\right)$, $j = 1, \ldots, m$. Derive the best linear combination and the corresponding maximum AUC.

Nonparametric method

Chen *et al.* proposed to use kernels to construct the EL-based confidence interval estimation for the BLC-based maximum AUC via construction of the empirical likelihood ratio (ELR) test statistic for testing the hypothesis $H_0 : A = A_0$ versus $H_1 : A \neq A_0$.[81]

Let k be a symmetric kernel function and define $K_h(x) = \int_{-\infty}^{x/h} k(u)du$, $v_i(\mathbf{a}) = m^{-1}\sum_{j=1}^{m} K_h(\mathbf{a}^T\mathbf{X}_i - \mathbf{a}^T\mathbf{Y}_j)$, $i = 1, \ldots, n$, where h is the bandwidth parameter. Regarding to the kernel estimation, we refer the reader to the textbook of Silverman.[82] Let $\mathbf{p} = (p_1, p_2, \ldots, p_n)^T$ be a probability weight vector, $\sum_{i=1}^{n} p_i = 1$ and $p_i \geq 0$ for all $i = 1, \ldots, n$. The EL for the BLC-based AUC evaluated at the true value A_0 of AUC can be defined as

$$L(A_0) = \sup\left\{\prod_{i=1}^{n} p_i : \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i v_i\left(\hat{\mathbf{a}}_0\right) = A_0\right\},$$

where $\hat{\mathbf{a}}_0$ satisfies $\sum_{i=1}^{n} p_i \partial v_i(\mathbf{a})/\partial \mathbf{a}|_{\mathbf{a}=\hat{\mathbf{a}}_0} = \mathbf{0}$. One can show (using Lagrange multipliers) that

$$p_i = \frac{1}{n}\frac{1}{1 + \lambda(v_i(\hat{\mathbf{a}}_0) - A_0)}, \quad i = 1, \ldots, n,$$

where the Lagrange multiplier λ is the root of

$$\frac{1}{n}\sum_{i=1}^{n}\frac{v_i(\hat{\mathbf{a}}_0)}{1 + \lambda(v_i(\hat{\mathbf{a}}_0) - A_0)} = A_0.$$

Under the alternative hypothesis, we have just the constraint $\sum_{i=1}^{n} p_i = 1$, and hence $L(A_0) = (1/n)^n$ at $p_i = 1/n$. Therefore, the empirical log-likelihood ratio test statistic is

$$l(A_0) = -2\log \mathrm{ELR}(A_0) = 2\sum_{i=1}^{n}\log(1 + \lambda(v_i(\hat{\mathbf{a}}_0) - A_0)).$$

We define $\hat{\mathbf{a}}_K = \arg\max_{\mathbf{a}} A^K_{m,n}(\mathbf{a})$, where $A^K_{m,n}(\mathbf{a}) = \sum_{i=1}^{n} v_i(\mathbf{a})/n$. Under some general conditions (see Chen *et al.*[81] for details), the asymptotic distribution of $l(A_0)$ under $H_0 : A = A_0$ is a scaled chi-square distribution with one degree of freedom, that is,

$$\gamma(A_0)l(A_0) \xrightarrow{d} \chi^2_1, \quad \text{as } n, m \to \infty,$$

where

$$\gamma(A_0) = \frac{m\hat{\sigma}^2}{(m+n)s^2}, \quad \hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\left(v_i\left(\hat{\mathbf{a}}_K\right) - n^{-1}\sum_{j=1}^{n} v_j(\hat{\mathbf{a}}_K)\right)^2, \quad s^2 = \frac{m\hat{\sigma}^2_{10} + n\hat{\sigma}^2_{01}}{m+n},$$

$$\sigma^2_{10} = \mathrm{Cov}(K_h(\mathbf{a_0}^T\mathbf{X}_1 - \mathbf{a_0}^T\mathbf{Y}_1),\ K_h(\mathbf{a_0}^T\mathbf{X}_1 - \mathbf{a_0}^T\mathbf{Y}_2)),$$

$$\sigma^2_{01} = \mathrm{Cov}(K_h(\mathbf{a_0}^T\mathbf{X}_1 - \mathbf{a_0}^T\mathbf{Y}_1),\ K_h(\mathbf{a_0}^T\mathbf{X}_2 - \mathbf{a_0}^T\mathbf{Y}_1)),$$

and $\hat{\sigma}^2_{10}$ and $\hat{\sigma}^2_{01}$ are the corresponding estimates.

Based on the asymptotic distribution of the statistic $l(A_0)$, the $100(1-\alpha)\%$ empirical likelihood-based confidence interval for the maximum AUC can be constructed as

$$R_\alpha = \left\{A_0 : \gamma\left(A_0\right) l(A_0) \leq \chi^2_1(1-\alpha)\right\},$$

where $\chi_1^2(1-\alpha)$ is the $(1-\alpha)$th quantile of the chi-square distribution with one degree of freedom. It gives a confidence interval with asymptotically correct coverage probability $1-\alpha$, that is, $\Pr(A_0 \in R_a) = 1-\alpha+o(1)$. R code related to the problem described above can be found at following the Web domain: http://www.sciencedirect.com/science/article/pii/S0167947314002710.

Goodness-of-fit tests

Many statistical procedures are, strictly speaking, only appropriate when a parametric assumption about data distribution is made. If the distribution under the null hypothesis is completely known, then testing for goodness of fit is equivalent to testing for uniformity. In general, testing distribution assumptions for normality and uniformity is suggested and has been one of the major areas of continuing statistical research both theoretically and practically. Normal distributions are commonly assumed in applications of statistical procedures. For instance, in most situations, parametric linear regression analysis can be done where the errors are assumed to be normally distributed. Thus, tests for goodness of fit, especially tests for normality, have a very important role in clinical experiments. Testing composite hypotheses of normality (or other specified), that is, H_0: the population is normally distributed versus H_1: the population is not normally distributed, is well addressed in statistical literature.[29,80,81] Included in the coverage are the Shapiro–Wilk test, the Kolmogorov–Smirnov test, and the Anderson–Darling test, among which the Shapiro–Wilk test is highly efficient and has the best power for a given significance.[82,83,84]

The Shapiro–Wilk test employs the null hypothesis principle to check whether a sample of i.i.d. observations $X_1, \ldots, X_n$ came from a normally distributed population. The test statistic is

$$W = \left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2 \Big/ \sum_{i=1}^{n} (X_i - \overline{X})^2,$$

where $X_{(i)}$ is the ith order statistic, that is, the ith smallest number in the sample, and $\overline{X}$ is the sample mean; the constants a_i are given by

$$(a_1, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}},$$

where $m = (m_1, \ldots, m_n)^T$ and $m_1, \ldots, m_n$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution and V is the covariance matrix of those order statistics. For example, when the sample size is 10, it can be obtained that $a_1 = 0.5739$, $a_2 = 0.3291$, $a_3 = 0.2141$, $a_4 = 0.1224$, and $a_5 = 0.0399$. The null hypothesis is rejected if W is below a predetermined threshold. Note that the Shapiro–Wilk test does not depend on the distribution of data under the null hypothesis and the corresponding critical value can be calculated numerically using Monte Carlo techniques, see section "Numerical methods for calculating critical values and powers of statistical tests" for details.

It should be noted that these tests for normality can be subject to low power, especially when the sample size is small. Thus, a Q–Q plot is recommended for verification in addition to the test.[85]

The function `shapiro.test` in R makes it possible to implement the Shapiro–Wilk test for normality.

Example 31.3

We generate data from $N(0, 1)$, Uniform$(0, 1)$, and exp(1) with the sample size $n = 100$ using R and check the data for normality.

The following program shows how to simulate the normal distribution, the uniform distribution, and the exponential distribution in R, respectively. The function `shapiro.test` in R conducts the Shapiro–Wilk test for normality, yielding the test statistic and the corresponding p-value, as seen below:

```
> n<-100
>
> # N(0,1)
> x<-rnorm(n,mean=0,sd=1)
> shapiro.test(x)

  Shapiro-Wilk normality test

data:  x
W = 0.9876, p-value = 0.4825


> # Uniform(0,1)
> y<-runif(n,min=0,max=1)
> shapiro.test(y)

  Shapiro-Wilk normality test

data:  y
W = 0.9545, p-value = 0.001674

> # exp(1)
> z<-rexp(n,rate=1)
> shapiro.test(z)

  Shapiro-Wilk normality test

data:  z
W = 0.8113, p-value = 5.574e-10
```

For the data simulated from $N(0, 1)$, we fail to reject the null hypothesis at the 0.05 significance level with a p-value of 0.4825, stating that there is no sufficient evidence

p-value = 0.4825

to conclude the data are not from the normal distribution. For the data simulated from either Uniform(0, 1) or exp(1), we reject the null hypothesis at the 0.05 significance level with a p-value of 0.0017 and <0.0001, respectively, stating that there is sufficient evidence to conclude the data are not from the normal distribution. For more details regarding goodness-of-fit tests, we refer the reader to Vexler *et al.* and Claeskens *et al.*[29,39,50]

Exercise 31.8

Simulate data from $N(1,2)$, Gamma$(1/2,1)$, $t_{\mathrm{df}=5}$ with the sample size $n = 15$, 25, and 50, respectively, and check for normality.

Wilcoxon rank-sum tests

As a nonparametric analog to the two-sample t-test, Wilcoxon rank-sum test (also called the Mann–Whitney U test or the Mann–Whitney–Wilcoxon test) can be used primarily when investigators do not want to, or cannot, assume that data distributions are normal.

Suppose that we have two samples of observations, containing i.i.d. measurements $X_1, \ldots, X_m$ and i.i.d. measurements $Y_1, \ldots, Y_n$, respectively. In practical applications, we often want to test the hypothesis that two populations are the same in the context of no location shift. To formulate, assuming that $X_1, \ldots, X_m \sim F(x-u)$, $Y_1, \ldots, Y_n \sim F(y-v)$, that all $m+n$ observations are independent, and that $F(\cdot)$ is symmetric about zero, it is of interest to test $H_0 : \Delta = v - u = 0$ against $H_1 : \Delta = v - u > 0$. Let R_i be the rank of Y_i among all $m+n$ observations, where the rank refers to the ordinal number of the corresponding observation among a pre-ordered data set in ascending order. The Mann–Whitney statistic $W = \sum_{i=1}^{n} R_i - n(n+1)/2 = \sum_{i=1}^{n}\sum_{i=1}^{n} I\{X_i < Y_j\}$ rejects H_0 for large values of W. It follows from one-sample U-statistics theory that, under the null hypothesis, W is asymptotically normal,[86] that is,

$$\frac{W - mn/2}{\sqrt{mn(m+n+1)/12}} \xrightarrow{d} N(0,1).$$

Therefore, a cutoff value for the α-level test can be found as $K_\alpha = mn/2 + 1/2 + z_\alpha\sqrt{mn(m+n+1)/12}$, where the additional $1/2$ is added for a continuity correction.

The Wilcoxon rank-sum test has greater efficiency than the t-test on non-normal distributions, and it is nearly as efficient as the t-test on normal distributions.

Example 31.4

We consider the HDL data described in Figure 31.2 and conduct a Wilcoxon rank-sum test for equality of means in two groups.

The function `wilcox.test` in R conducts the two-sample Wilcoxon test for equality on means. Note that the alternative can be revised to "less" or "greater" in terms of a one-sided test.

```
> X<-c(37.4,70.4,52.8,46.2,74.8,96.8,41.8,55.0,83.6,63.8,63.8,
52.8,46.2,37.4,50.6,74.8,46.2,39.6,70.4,30.8,74.8,61.6,30.8,
74.8,52.8)
> Y<-c(44.0,35.2,110.0,63.8,44,26.4,52.8,30.8,39.6,44,48.4,
39.6,55,52.8,50.6,39.6,35.2,55,57.2,37.4,30.8,46.2,50.6,44,44)
> wilcox.test(X,Y)

  Wilcoxon rank sum test with continuity correction

data:  X and Y
W = 432, p-value = 0.02065
alternative hypothesis: true location shift is not equal to 0
```

For the HDL data, the p-value of the Wilcoxon rank-sum test is 0.02065. Thus, we reject the null hypothesis at the 0.05 significance level and conclude a significant difference in the mean between two groups.

Tests for independence

Evaluations of relationships between pairs of variables, including testing for independence, are increasingly important. In this section, we introduce nonparametric tests for independence, including the Pearson correlation coefficient ρ, the Spearman's rank correlation coefficient ρ_S, the Kendall's rank correlation coefficient, data-driven rank techniques, the empirical likelihood-based method, and the density-based empirical likelihood ratio test.[56,60,87] Note that the Pearson correlation coefficient, the Spearman's rank correlation coefficient, and the Kendall's rank correlation coefficient focus on specific interdependence, for example, linear and/or monotone. In reality, the dependence structure may be more complex. For example, in the models $Y = 1/X$, $Y = \varepsilon/X$, or $Y = \varepsilon/X^2$, where ε is a random variable, X and Y are dependent in an inverse manner, and in the second and third cases, $E(XY)$ can be nonexistent. In this lies the difficulty of interpreting the correlation as a measure of dependence in general.

Assume we obtain a random sample of n pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from a continuous bivariate population. Let $F_X(x)$ and $F_Y(y)$ denote the marginal distribution functions of X and Y, respectively, and let $F_{XY}(x,y) = \Pr(X \le x, Y \le y)$ be the joint distribution function of the (X, Y) pairs. For future use, let R_i and S_i denote the rank of X_i and Y_i, $i = 1, \ldots, n$, respectively, and F_{Xn}, F_{Yn}, and F_n be the empirical distribution functions of F_X, F_Y, and F_{XY}, respectively. The null hypothesis of bivariate independence between X and Y can be formally stated as $H_0 : F(x,y) = F(x)F(y)$ for all $(x,y) \in R^2$ versus $H_1 : F(x,y) \neq F(x)F(y)$ for some $(x,y) \in R^2$.

Pearson correlation coefficient

Pearson's correlation coefficient ρ is the most familiar dependence concept that measures the linear dependence between a pair of variables (X, Y). It is defined in terms of moments as

$$\rho = \rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E((X - \mu_X)(Y - \mu_Y))}{\sqrt{E((X - \mu_X)^2)E((Y - \mu_Y)^2)}},$$

where $\text{cov}(X, Y)$ represents the covariance of X and Y and σ_X, $\sigma_Y > 0$ denotes the standard deviations of X and Y, respectively. The Pearson correlation is defined only if both of the standard deviations are finite and nonzero. Applying the Cauchy–Schwarz inequality to the definition of covariance, it can be easily shown that $-1 \le \rho \le 1$, where $\rho = 1$ indicates a perfect increasing linear relationship and $\rho = -1$ shows a perfect decreasing linear relationship. By substituting corresponding moment estimators, the sample Pearson correlation coefficient can be obtained as

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}, \quad -1 \le r \le 1.$$

If two variables are from bivariate normal distribution or the sample size is not very small, $t = \sqrt{(n-2)/(1-r^2)}r$, which has an asymptotic t-distribution with $n - 2$ degrees of freedom under H_0.[86] Accordingly, we reject H_0 if $|t| \ge t_{\alpha/2,n-2}$, where $t_{\alpha/2,n-2}$ is the $(1 - \alpha/2)$th quantile of t-distribution with $n - 2$ degrees of freedom.

Spearman's rank correlation coefficient

Spearman's rank correlation coefficient ρ_S is the Pearson correlation between ranks of X and Y, that is, $\rho_S = \rho_S(X, Y) = \rho(F_X(X), F_Y(Y))$, $-1 \le \rho_S \le 1$. It accesses a monotonic relationship between two variables. Testing for independence is equivalent to the test $H_0 : \rho_S = 0$ versus $H_1 : \rho_S \ne 0$. The sample Spearman's rank correlation coefficient is

$$r_s = 1 - 6\sum_{i=1}^{n}(R_i - S_i)^2/(n(n^2 - 1)).$$

Note that if there are tied X values and/or tied Y values, each observation in the tied group is assigned with the average of the ranks associated with the tied group.

At the significance level α, we reject H_0 if $|r_s| \ge r_{s,\alpha/2}$, where $r_{s,\alpha/2}$ can be found by *qSpearman* in R.[89] For large sample sizes, we can also conduct the test based on the asymptotic t-distribution of the Pearson correlation coefficient between the ranked variables.

Kendall's rank correlation coefficient

The Kendall's rank correlation coefficient is a distribution-free measure of independence based on signs of products of differences, where

$$\tau = \Pr((X_1 - X_2)(Y_1 - Y_2) > 0) - \Pr((X_1 - X_2)(Y_1 - Y_2) < 0), \quad -1 \le \rho_S \le 1.$$

It is a measure of the relative difference between Pr{concordance} and Pr{discordance}. Testing for independence is equivalent to the test $H_0 : \tau = 0$ versus $H_1 : \tau \neq 0$. The Kendall statistic can be defined as

$$K = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{sgn}\{(Y_j - Y_i)(X_j - X_i)\},$$

where $\text{sgn}\{x\} = 1$, if $x > 0$; 0, if $x = 0$; and -1, if $x < 0$.

Accordingly, at the significance level α, an exact test can be conducted, and we reject H_0 if $\overline{K} \geq k_{\alpha/2}$, where $\overline{K} = K/(n(n-1)/2)$ and $k_{\alpha/2}$ can be found by *qKendall* in R.[89] Alternatively, the test can be conducted based on the asymptotic standard normal distribution of the standardized $K^* = (n(n-1)(2n+5)/18)^{-1/2}K$ under H_0. The null hypothesis is rejected if $|K^*| \geq z_{\alpha/2}$, where $z_{\alpha/2}$ is the $100(1-\alpha/2)$th quantile of the standard normal distribution.

Example 31.5

We consider the HDL data described in Figure 31.2 and test for independence between X and Y.

The function `cor.test` conducts the test for independence between two samples. Note that the alternative can be revised to "less" or "greater" in terms of a one-sided test. The method option can be "Pearson," "Kendall," or "Spearman," based on the method chosen. The following shows the Pearson correlation test for the HDL data:

```
> cor.test(X,Y,alternative = "two.sided",method="pearson")

  Pearson's product-moment correlation

data:  X and Y
t = -1.5704, df = 23, p-value = 0.13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.62897919  0.09571223
sample estimates:
      cor
-0.3111874
```

The p-value of the Pearson correlation test for the HDL data is 0.13. Thus, we fail to reject the null hypothesis at the 0.05 significance level and state that there is not sufficient evidence to conclude X and Y are independent.

Data-driven rank tests

Kallenberg *et al.* expressed the dependence between X and Y via Fourier coefficients of the grade representation against a very wide class of alternatives.[87] To this end, the distribution of $(F_X(X), F_Y(Y))$ is considered as exponential families with respect

to the Lebesgue measure $[0,1]\times[0,1]$. It is assumed that the observed samples are distributed according to the joint density function given as

$$h(F_X(x), F_Y(y)) = c(\theta)exp\left\{\sum_{j=1}^{k}\theta_j b_j(x^*)\, b_j(y^*)\right\},$$

where b_j denotes the jth orthonormal Legendre polynomial, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$, and $c(\theta)$ is a normalizing constant.

Within exponential families, the null hypothesis corresponds to $\boldsymbol{\theta} = 0$, and the score test for testing $\boldsymbol{\theta} = 0$ against $\boldsymbol{\theta} \neq 0$ is given by rejecting for large values of $\left\{n^{-1/2}\sum_{i=1}^{n} b_r\left(F_X\left(X_i\right)\right) b_s(F_Y(Y_i))\right\}^2$.

A smooth test statistic

$$T_k = \sum_{j=1}^{k} V(j,j),$$

where

$$V(r,s) = \left\{n^{-\frac{1}{2}}\sum_{i=1}^{n} b_r\left(\frac{R_i - 1/2}{n}\right) b_s\left(\frac{S_i - 1/2}{n}\right)\right\}^2,$$

can be obtained by replacing unknown distribution functions F_X and F_Y by corresponding empirical distribution functions and applying a correction for continuity. Accordingly, two different test statistics $TS2$ and V were proposed based on different selection of the order k.

A "diagonal" test statistic $TS2 = T_{S2}$ is useful in the case of the "diagonal" model, which contains only products of Legendre polynomials with the same order in both variables. Let $d(n)$ be a sequence of numbers tending to infinity as $n \to \infty$. In a similar manner to the modified Schwarz's rule,[90] the order is chosen as $S2 = \text{argmin}_k\{T_k - k\log(n) \geq T_j - j\log(n), 1 \leq j,k \leq d(n)\}$. The score test for testing $H_0 : \boldsymbol{\theta} = 0$ against $H_a : \boldsymbol{\theta} \neq 0$ in the exponential family is given by rejecting for large values of TS2.

Otherwise, the "mixed" products are involved and a "mixed" statistic can be used. Let $|\Lambda|$ denote the cardinality of Λ and $T_\Lambda = \sum_{(r,s)\in\Lambda} V(r,s)$, and search for a model $\Lambda^* = \arg\max_\Lambda\{T_\Lambda - |\Lambda|\log(n)\}$. If Λ^* is not unique, the first among those Λ^*'s that have smallest cardinality is chosen. Then, the "mixed" statistic of H_0 is $V = T_{\Lambda^*}$.

Empirical likelihood-based method

Einmahl *et al.* constructed a test statistic by localizing the empirical likelihood.[56] Let $L(\widetilde{F}_{XY}) = \prod_{i=1}^{n}\widetilde{P}(\{(X_i, Y_i)\})$, where $\widetilde{P}$ is the probability measure corresponding to F_{XY}. For $(x,y) \in R^2$, the local likelihood ratio test statistic is

$$R(x,y) = \frac{\sup\{L(\widetilde{F}_{XY}) : \widetilde{F}_{XY}(x,y) = \widetilde{F}_X(x)\widetilde{F}_Y(y)\}}{\sup\{L(\widetilde{F}_{XY})\}}.$$

Then,

$$\log R(x,y) = nP_n(A_{11})\log\frac{F_{Xn}(x)F_{Yn}(y)}{P_n(A_{11})} + nP_n(A_{12})\log\frac{F_{Xn}(x)(1-F_{Yn}(y))}{P_n(A_{12})}$$
$$+ nP_n(A_{21})\log\frac{(1-F_{Xn}(x))F_{Yn}(y)}{P_n(A_{21})} + nP_n(A_{22})\log\frac{(1-F_{Xn}(x))(1-F_{Yn}(y))}{P_n(A_{22})},$$

where P_n is the empirical probability measure, F_{Xn} and F_{Yn} are the corresponding marginal distribution functions, and $A_{11} = (-\infty, x] \times (-\infty, y]$, $A_{12} = (-\infty, x] \times (y, \infty)$, $A_{21} = (x, \infty) \times (-\infty, y]$, $A_{22} = (x, \infty) \times (y, \infty)$, and $0 \log(\cdot/0) = 0$. Then, the distribution-free test statistic T_n of testing for independence, that is,

$$T_n = -2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \log R(x, y) dF_{Xn}(x) dF_{Yn}(y),$$

can be obtained by forming integrals of the log-likelihood ratio statistic.

Density-based empirical likelihood ratio test

Vexler *et al.* considered a density-based empirical likelihood approach for creating nonparametric test statistics, which approximate parametric Neyman–Pearson statistics, to test the null hypothesis of bivariate independence against a wide class of alternatives.[60] The test statistic is defined as

$$VT_n = \prod_{i=1}^{n} n^{1-\beta_2} \widetilde{\Delta}_i([0.5n^{\beta_2}, 0.5n^{\beta_2}]),$$

where the function $[x]$ denotes the nearest integer to x, $0.75 < \beta_2 < 0.9$, and

$$\widetilde{\Delta}_i(m, r) = (F_{Xn}(X_{(R_i+r)}) - F_{Xn}(X_{(R_i-r)}))^{-1}(F_n(X_{(R_i+r)}, Y_{(i+m)}) - F_n(X_{(R_i-r)}, Y_{(i+m)}) - F_n(X_{(R_i+r)}, Y_{(i-m)}) + F_n(X_{(R_i-r)}, Y_{(i-m)}) + n^{-\beta_1}).$$

Note that $0 < \beta_1 < 0.5$ ensures the consistency of the proposed test. The proposed test is exact. The null hypothesis is rejected if $\log(VT_n) > C_\alpha$, where C_α is an α-level test threshold. It follows that $\Pr_{H_0}(\log(VT_n) > C_\alpha) = \Pr_{\{X_i\}_{i=1}^n, \{Y_i\}_{i=1}^n \sim \text{Uniform}[0,1]}(\log(VT_n) > C_\alpha | H_0)$. The critical values for the proposed test can be accurately approximated using Monte Carlo techniques, see [59, 91] and the next section for details.

Numerical methods for calculating critical values and powers of statistical tests

Many statistical tests, including Shapiro–Wilk tests, data-driven rank techniques, the empirical likelihood-based method, the density-based empirical likelihood ratio test, t-test-type tests, and likelihood ratio tests are exact. Exact tests are well known to be simple, efficient, and reliable and to have finite sample Type I error control. Under the null hypothesis, distributions of test statistics for exact tests are independent of the underlying data distributions. For example, we consider the two-sample t-test with equal variances introduced in section "Multivariate t-tests." When the assumptions of normality and homogeneity of variance are satisfied, the test is exact, that is, the sampling distribution of $t = (\overline{X}_1 - \overline{X}_2)/(s_p \sqrt{n_1^{-1} + n_2^{-1}})$ under a true null hypothesis would be given exactly by the t-distribution with degrees of freedom $n_1 + n_2 - 2$ (we refer the notations to Section "Multivariate t-tests."). Another concrete example of exact tests is the Wilcoxon rank-sum test, in which the test statistics are based on indicator functions $I\{\cdot\}$. Noticing the fact that $I\{Z_1 < Z_2\} = I\{F_Z(Z_1) < F_Z(Z_2)\}$ and

$$\tilde{\Delta}_i(m,r) \equiv \left(F_{\ddot{X}n}(X_{(\mathrm{M}_i+r)}) - F_{\ddot{X}n}(X_{(\mathrm{M}_i-r)})\right)^{-1} \left(F_n(X_{(\mathrm{M}_i+r)}, Y_{(i+m)})\right.$$
$$\left. - F_n(X_{(\mathrm{M}_i-r)}, Y_{(i+m)}) - F_n(X_{(\mathrm{M}_i+r)}, Y_{(i-m)}) + F_n(X_{(\mathrm{M}_i-r)}, Y_{(i-m)}) + n^{-\beta_1}\right),$$

With that

M_i is an integer number such that $X_{(\mathrm{M}_i)} = X_{t(i)}$ ($X_{t(i)}$ is the concomitant of the i-th order statistic $Y_{(i)}$, see [60] for details); $X_{(\mathrm{M}_i+r)} = X_{(n)}$, if $M_i + r > n$; $X_{(\mathrm{M}_i-r)} = X_{(1)}$, if $M_i - r < 1$

1. David, H. A., and Nagaraja, H. N. (2003), *Order Statistics*. New York: Wiley

$I\{Z_1 < -Z_2\} = I\{F_Z(Z_1) < F_Z(-Z_2)\} = I\{F_Z(Z_1) < 1 - F_Z(Z_2)\}$, where the random variables $F_Z(Z_1)$ and $F_Z(Z_2)$ have a uniform distribution under H_0, the distributions of the test statistics for the Wilcoxon rank-sum test are independent of the distributions of observations. Due to the independence of the null distribution of test statistics on the data distribution, the critical values of exact tests can be computed exactly, without using asymptotic approximations.

Methods that can be used to calculate the critical values of exact tests include a classical technique based on Monte Carlo (MC) evaluations, an interpolation technique based on tabulated critical values, and a hybrid of the MC and the interpolation methods. The classical Monte Carlo strategy is a well-known approach for obtaining accurate approximations to the critical values of exact tests. The critical values can be calculated by simulating data for a relatively large number of MC repetitions; say, for example, from a standard normal distribution for one-sample tests and a Uniform(0,1) distribution for two-sample and three-sample tests. The generated values of the test statistic L of the exact test of interest can be used to determine the critical value C_α at the desired significance level α. Assuming that the decision rule is to reject H_0 for large values of L, that is, when $L > C_\alpha$, then the critical value C_α can be obtained via calculating the $1 - \alpha$ quantile of the MC null distribution of L. However, the use of the MC technique can be computationally intensive in some testing situations. For example, a relatively large number of MC repetitions, which we define as M, are needed to evaluate critical values that correspond to the 1% significance level, since in this case the common 95% confidence interval of such evaluation can be calculated as $[0.01 \pm 1.96\sqrt{0.01(1-0.01)/M}]$. Another standard method applied in various statistical software routines is the interpolation technique based on tabulated critical values. Interpolation differs from MC method in that tables of critical values are calculated beforehand for an exact test of interest and for various sample sizes and significance levels. Therefore, the execution speed of the testing algorithm improves when the tables are provided for use within the testing algorithm. However, the interpolation method becomes less reliable when real data characteristics (e.g., sample sizes) differ from those used to tabulate the critical values. As an outgrowth of the methods described above, a hybrid method combines both interpolation and MC by means of the nonparametric Bayes concept. The hybrid method can be applied in a broad setting and is shown to be very efficient in the context of exact-test computations of critical values and powers.[57,91]

Concluding remarks

The necessity and danger of testing the statistical hypothesis

The ubiquitous use of statistical decision-making procedures' findings in the current medical literature displays the vital role that statistical hypothesis testing plays in clinical trials in different branches of biomedical sciences. The benefits and fruits of statistical tests based on mathematical probabilistic techniques in epidemiology or other health-related disciplines strongly depend on successful formal presentations of statements of problems and a description of their nature. Oftentimes, certain assumptions about the observations used for the tests provide the probability statements that are

required for the statistical tests. These assumptions do not come for free, and ignoring their appropriateness can cause serious bias or inconsistency of statistical inferences, even when the test procedures themselves are carried out without mistakes. The sensitivity of the probabilistic properties of a test to the assumptions is referred to as the lack of robustness of the test.[92]

Various statistical techniques require parametric assumptions to define forms of data distributions to be known up to parameters' values. For example, in the *t*-test, the assumptions are that the observations of different individuals are realizations of independent, normally distributed random variables, with the same expected value and variance for all individuals within the investigated group. Such assumptions are not automatically satisfied, and for some assumptions, it may be doubted whether they are ever satisfied exactly. The null hypothesis H_0 and alternative hypothesis H_1 are statements that, strictly speaking, imply these assumptions, and which therefore are not each other's complement. There is a possibility that the assumptions are invalid, and neither H_0 nor H_1 is true. Thus, we can reject a statement related to clinical trials' interests just because the assumptions are not met. This issue is an impetus to depart from parametric families of data distributions and employ nonparametric test-strategies. Wilk and Gnanadesikan described and discussed graphical techniques based on the primitive empirical CDF and on quantile (*Q*–*Q*) plots, percent (*P*–*P*) plots, and hybrids of these, which are useful in assessing one-dimensional samples.[85] Statistical techniques such as the likelihood ratio test, the maximum likelihood ratio test, the ROC curve methodology, *t*-tests, and so on, can really assist in solving challenging epidemiology and biomedical problems. Several components regarding tests based on incomplete data or data subject to instrument limitation can be found in Vexler *et al.*[12] Reiser developed a corrected confidence interval for the AUC, adjusted for measurement errors.[93] In this chapter, we considered retrospective statistical problems, that is, issues based on already collected data. There are statistical mechanisms based on sequentially observed measurement.[94]

Appendix

The most powerful test: As discussed in section "Likelihood," the most powerful statistical decision rule is to reject H_0 if and only if $\prod_{i=1}^{n} f_1(X_i)/f_0(X_i) \geq B$. The term "most powerful" induces us to formally define how to compare statistical tests. Without loss of generality, since the ability to control the Type I error (TIE) of statistical tests has an essential role in statistical decision-making, we compare tests with equivalent probabilities of the TIE, $\Pr_{H_0}\{\text{test rejects } H_0\} = \alpha$, where the subscript H_0 indicates that we consider the probability given that the hull hypothesis is correct. The level of significance α is the probability of making a TIE. In practice, the researcher should choose a value of α, for example, $\alpha = 0.05$, before performing the test. Thus, we should compare the likelihood ratio test with δ, any decision rule based on $\{X_i, i = 1, \ldots, n\}$, setting up $\Pr_{H_0}\{\delta \text{ rejects } H_0\} = \alpha$ and $\Pr_{H_0}\left\{\prod_{i=1}^{n} f_1\left(X_i\right)/f_0(X_i) \geq B\right\} = \alpha$. This comparison is with respect to the power $\Pr_{H_1}\{\text{test rejects } H_0\}$. Notice that to derive the mathematical expectation, in the context of a problem related to testing statistical hypotheses, one must define whether the expectation should be conducted under

H_0- or H_1-regime. For example,

$$E_{H_1}\varphi(X_1, X_2, \ldots X_n) = \int \varphi(x_1, x_2, \ldots x_n) f_1(x_1, x_2, \ldots x_n) dx_1 dx_2 \ldots dx_n$$
$$= \int \varphi(x_1, x_2, \ldots x_n) \prod_{i=1}^{n} f_1(x_i) \prod_{i=1}^{n} dx_i,$$

where the expectation is considered under the alternative hypothesis. The indicator $I\{C\}$ of the event C can be considered as a random variable with values 0 and 1. By virtue of the definition, the expected value of $I\{C\}$ is $\mathrm{EI}\{C\} = 0 \times \Pr\{I\{C\} = 0\} + 1 \times \Pr\{I\{C\} = 1\} = \Pr\{I\{C\} = 1\} = \Pr\{C\}$.

Taking into account the comments mentioned above, we derive the expectation under H_0 of the inequality (31.1), where $A = \prod_{i=1}^{n} f_1(X_i)/f_0(X_i)$, B is a test threshold, and δ represents any decision rule based on $\{X_i, i = 1, \ldots, n\}$. One can assume that $\delta = 0, 1$, and when $\delta = 1$ we reject H_0. Thus, we obtain

$$E_{H_0}\left(\left(\prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)} - B\right) I\left\{\frac{f_1(X_i)}{f_0(X_i)} \geq B\right\}\right) \geq E_{H_0}\left(\left(\prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)} - B\right)\delta\right).$$

And hence,

$$E_{H_0}\left(\prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)} I\left\{\frac{f_1(X_i)}{f_0(X_i)} \geq B\right\}\right) - BE_{H_0}\left(I\left\{\frac{f_1(X_i)}{f_0(X_i)} \geq B\right\}\right)$$
$$\geq E_{H_0}\left(\prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)}\delta\right) - BE_{H_0}(\delta),$$

where $E_{H_0}(\delta) = E_{H_0}(I\{\delta = 1\}) = \Pr_{H_0}\{\delta = 1\} = \Pr_{H_0}\{\delta \text{ rejects } H_0\}$. Since we compare the tests with the fixed level of significance

$$E_{H_0}\left(I\left\{\frac{f_1(X_i)}{f_0(X_i)} \geq B\right\}\right) = \Pr_{H_0}\left\{\frac{f_1(X_i)}{f_0(X_i)} \geq B\right\} = \Pr_{H_0}\{\delta \text{ rejects } H_0\} = \alpha,$$

we have

$$E_{H_0}\left(\prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)} I\left\{\frac{f_1(X_i)}{f_0(X_i)} \geq B\right\}\right) \geq E_{H_0}\left(\prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)}\delta\right). \tag{A.1}$$

Consider

$$E_{H_0}\left(\prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)}\delta\right) = E_{H_0}\left(\prod_{i=1}^{n} \frac{f_1(X_i)}{f_0(X_i)}\delta(X_1, \ldots, X_n)\right)$$
$$= \int \prod_{i=1}^{n} \frac{f_1(x_i)}{f_0(x_i)}\delta(x_1, \ldots, x_n) f_0(x_1, \ldots, x_n) dx_1 \ldots dx_n$$
$$= \int \frac{\prod_{i=1}^{n} f_1(x_i)}{\prod_{i=1}^{n} f_0(x_i)}\delta(x_1, \ldots, x_n) \prod_{i=1}^{n} f_0(x_i) dx_1 \ldots dx_n = \int \delta(x_1, \ldots, x_n) \prod_{i=1}^{n} f_1(x_i) dx_1 \ldots dx_n$$
$$= E_{H_1}\delta = \Pr_{H_1}\{\delta \text{ rejects } H_0\}. \tag{A.2}$$

Since δ represents any decision rule based on $\{X_i, i = 1, \dots, n\}$, including the likelihood ratio based test, equation (A.2) implies

$$E_{H_0}\left(\prod_{i=1}^{n}\frac{f_1(X_i)}{f_0(X_i)}I\left\{\frac{f_1(X_i)}{f_0(X_i)} \geq B\right\}\right) = \Pr_{H_1}\left\{\prod_{i=1}^{n}\frac{f_1(X_i)}{f_0(X_i)}I\left\{\frac{f_1(X_i)}{f_0(X_i)} \geq B\right\}\right\}.$$

Applying this equation and (A.2) to (A.1), we complete to prove that the likelihood ratio test is a most powerful statistical decision rule.

The likelihood ratio property $f^L_{H_1}(u) = f^L_{H_0}(u)u$: in order to obtain this property, we consider

$$\Pr_{H_1}\{u - s \leq L \leq u\} = E_{H_1} I\{u - s \leq L \leq u\} = \int I\{u - s \leq L \leq u\} f_{H_1}$$
$$= \int I\{u - s \leq L \leq u\}\frac{f_{H_1}}{f_{H_0}} f_{H_0} = \int I\{u - s \leq L \leq u\} L f_{H_0}.$$

This implies the inequalities

$$\Pr_{H_1}\{u - s \leq L \leq u\} \leq \int I\{u - s \leq L \leq u\} u f_{H_0} = u\Pr_{H_0}\{u - s \leq L \leq u\}$$

and

$$\Pr_{H_1}\{u - s \leq L \leq u\} \geq \int I\{u - s \leq L \leq u\}(u - s) f_{H_0} = (u - s)\Pr_{H_0}\{u - s \leq L \leq u\}.$$

Dividing these inequalities by s and employing $s \to 0$, we get $f^L_{H_1}(u) = f^L_{H_0}(u)u$, where $f^L_{H_0}(u)$ and $f^L_{H_1}(u)$ are the density functions of the statistic $L = f_{H_1}/f_{H_0}$ under H_0 and H_1, respectively.

The general form of the AUC: By the definition of the AUC (the area under the ROC curve) and the fact that F and G are CDFs of X and Y, respectively, the AUC can be expressed as

$$\int_0^1 \mathrm{ROC}(t)dt = \int_0^1 (1 - FG^{-1}(1 - t))dt = \int_{-\infty}^{\infty}(1 - F(w))dG(w)$$
$$= 1 - \int_{-\infty}^{\infty} F(w)dG(w) = 1 - \Pr(X \leq Y) = \Pr(X > Y).$$

The form of the AUC under the normal data distribution assumption: Assume $X \sim N(\mu_1, \sigma_1^2)$ and, for the nondiseased population, $Y \sim N(\mu_2, \sigma_2^2)$. Note that X and Y are independent. Consequently, we can obtain that

$$A = \Pr(X > Y) = \Pr(X - Y > 0) = 1 - \Pr\left\{\frac{(X - Y) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}} \leq -\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right\}$$
$$= 1 - \Phi\left(-\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right) = \Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right).$$

R code: the two-sample (X and Y) density-based ELR test (for the details, see Ref. 51):

```
############sample data with the sample sizes
n1=n2=25############
n1=25
n2=25
x<-sample(control,n1)
y<-sample(case,n2)
delta<-0.1
z<-c(x,y)
sx<-sort(x)
sy<-sort(y)
sz<-sort(z)
#####################################################
#######obtaining the ELR based on the sample X###
#####################################################
m<-c(round(n1^(delta+0.5)):min(c(round((n1)^(1-delta)),
round(n1/2))))  ###generate a vector      of "m"
a<-replicate(n1,m)           ###store repeated values of the
vector "m"
rm<-as.vector(t(a))          ###transpose the previous
length(m)*n1 matrix and make it to be a vector
#rm<-rep(m, each = n1)       ###repeat the vector of "m"
#n1 times
L<-c(1:n1)- rm               ###order from (1-m) to (n1-m)
LL<-replace(L, L <= 0, 1 )   ###replace values that are
#<=0 with 1 when (1-m) <=0
U<-c(1:n1)+ rm               ###order from (1+m) to (n1+m)
UU<-replace(U, U > n1, n1)   ###replace values that are n1
#with n1 when (n1+m)>n1
xL<-sx[LL]   ###obtain x(i-m)
xU<-sx[UU]   ###obtain x(i+m)
F<-ecdf(z)(xU)-ecdf(z)(xL)   ### the empirical distribution
#function
F[F==0]<-1/(n1+n2
I<-2*rm/(n1*F)      ### a (n1*length(m)) vector of (2*m)/
#(n1*empirical distribution function)
ux<-array(I, c(n1,length(m)))    ### make the previous vector
#as a n1*length(m) matrix
tstat1<-log(min(apply(ux,2,prod)))              ###get the
#part of the test statistic based on the sample X
#####################################################
#######obtaining the ELR based on the sample Y#######
#####################################################
m<-c(round(n2^(delta+0.5)):min(c(round((n2)^(1-delta)),
round(n2/2))))   ###generate a vector of "m"
```

```
a<-replicate(n2,m)            ###store repeated values of the
#vector "m"
rm<-as.vector(t(a))          ###transpose the previous
#length(m)*n2 matrix and make it to be a vector
#rm<-rep(m, each = n2)          ###repeat the vector of "m"
#n2 times
L<-c(1:n2)-rm                 ###order from (1-m) to (n2-m)
LL<-replace(L, L <= 0, 1 )     ###replace values that are
#<=0 with 1 when (1-m) <=0
U<-c(1:n2)+ rm                 ###order from (1+m) to (n2+m)
UU<-replace(U, U > n2, n2)     ###replace values that are
#>n2 with n2 when (n2+m)>n2
yL<-sy[LL]  ###obtain y(i-m)
yU<-sy[UU]  ###obtain y(i+m)
F<-ecdf(z)(yU)-ecdf(z)(yL) ###the empirical distribution
#function
F[F==0]<-1/(n1+n2)
I<-2*rm/(n2*F)     ### the (n2*length(m)) vector of (2*m)/
#(n2*empirical distribution fuction)
uy<-array(I, c(n2,length(m)))              ### make the
#previous vector as a n2*length(m) matrix
tstat2<-log(min(apply(uy,2, prod)))        ###get ELR_Y
finalts<-tstat1+tstat2                   ### the final test
statistic log(V)
```

Multiple choice questions

1 Investigators are interested in evaluating a treatment A. They target to show that A can significantly improve health conditions of patients. How can the investigators formulate the null statistical hypothesis?

- **a.** The corresponding distribution function based on measurements from a group of patients without treatment A is different from that based on measurements from a group of patients with treatment A
- **b.** The corresponding distribution function based on measurements from a group of patients without treatment A equals to that based on measurements from a group of patients with treatment A
- **c.** The corresponding distribution function based on measurements from a group of patients without treatment A is smaller than that based on measurements from a group of patients with treatment A

2 Assume we observe i.i.d. measurements from a known distribution that depends on a simple parameter θ. We would like to test that $\theta = 0$ versus $\theta \neq 0$. What kind of testing strategies could you propose:

- **a.** Likelihood ratio
- **b.** *t*-Test
- **c.** Maximum likelihood ratio.
- **d.** Empirical likelihood ratio
- **e.** Density-based empirical likelihood ratio

3 Assume we observe independent measurements from normal distributions with the variances $\sigma_X^2 = \sigma_Y^2 = 1$. The diseased population is presented by a sample $X_1, \ldots, X_n$, the nondiseased population is presented by a sample $Y_1, \ldots, Y_m$. To evaluate the area under the ROC curve, it is better to use the following formal notation:

Q6

a. $\Phi\left(\frac{\overline{X} - \overline{Y}}{\sqrt{2}}\right)$

b. $\frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} I\{X_i > Y_j\}$

c. $\frac{(\overline{X} - \overline{Y})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{j=1}^{m}(Y_j - \overline{Y})^2}$

References

Q7

1 Rothman, K.J., Greenland, S. & Lash, T.L. (2008) *Modern Epidemiology*. Lippincott Williams & Wilkins.

2 Freiman, J.A., Chalmers, T.C., Smith, H. Jr., *et al.* (1978) The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. *The New England Journal of Medicine*, 299(13), 690–694.

3 Goodman, S.N. (1999) Toward evidence-based medical statistics. 1: The P value fallacy. *Annals of Internal Medicine*, 130(12), 995–1004.

Q8

4 Berger, J.O. & Mohan, D. (1987) Testing precise hypotheses. *Statistical Science*, 2(3), 317–335.

5 Gibbons, J.D. & Chakraborti, S. (2011) *Nonparametric Statistical Inference*. Springer.

6 Sackrowitz, H. & Samuel-Cahn, E. (1999) P values as random variables – expected P values. *The American Statistician*, 53(4), 326–331.

7 Berger, J.O. (1985) *Statistical Decision Theory and Bayesian Analysis*. Springer.

8 Vexler, A., Tao, G. & Hutson, A.D. (2014) Posterior expectation based on empirical likelihoods. *Biometrika*, 101(3), 711–718.

9 Lindsey, J. (1996) *Parametric Statistical Inference*. Oxford Science Publications.

10 Freedman, D. (2009) *Statistical Models: Theory and Practice*. Cambridge University Press.

11 Austin, P.C., Mamdani, M.M., Juurlink, D.N. *et al.* (2006) Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *Journal of Clinical Epidemiology*, 59(9), 964–969.

12 Vexler, A., Tao, G. & Chen, X. (2014) A toolkit for clinical statisticians to fix problems based on biomarker measurements subject to instrumental limitations: From repeated measurement techniques to a hybrid pooled-unpooled design. In: *Advanced Protocols in Oxidative Stress III*. Humana Press.

13 Wilcox, R.R. (2012) *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press.

14 Lehmann, E.L. & Romano, J.P. (2006) *Testing Statistical Hypotheses*. Springer.

15 Riffenburgh, R.H. (2012) *Statistics in Medicine*, 3rd edn. Academic Press.

16 Pepe, M.S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

17 Vexler, A., Schisterman, E.F. & Liu, A. (2008) Estimation of ~~roc~~ ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures. *Statistics in Medicine*, 27(2), 280–296.

18 R Development Core Team. *R: A Language and Environment for Statistical Computing*. 2013. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. ISBN 3-900051-07-0.

19 Gardener, M. (2012) *Beginning R: The Statistical Programming Language*. John Wiley & Sons.

capital ROC

20 Crawley, M.J. (2012) *The R Book*. John Wiley & Sons.

21 Limpert, E., Stahel, W.A. *et al.* (2001) Log-normal distributions across the sciences: Keys and clues on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability-normal or log-normal: That is the question. *BioScience*, 51(5), 341–352.

22 Schisterman, E.F., Faraggi, D., Browne, R. *et al.* (2001) Tbars and cardiovascular disease in a population-based sample. *Journal of Cardiovascular Risk*, 8(4), 219–225.

23 Neyman, J. & Pearson, E.S. (1928) On the use and interpretation of certain test criteria for purposes of statistical inference: Part II. *Biometrika*, 20, 263–294.

24 Neyman, J., Pearson, E. S. (1933) The testing of statistical hypotheses in relation to probabilities a priori. In *Mathematical Proceedings of the Cambridge Philosophical Society*, 29(4), 492–510. Cambridge University Press.

25 Neyman, J. & Pearson, E.S. (1936) *Contributions to the Theory of Testing Statistical Hypotheses I*. University Press, pp. 1–37.

26 Neyman, J. & Pearson, E.S. (1938) *Contributions to the Theory of Testing Statistical Hypotheses II*. University Press.

27 Vexler, A., Wu, C. & Yu, K.F. (2010) Optimal hypothesis testing: from semi to fully bayes factors. *Metrika*, 71(2), 125–138.

28 Vexler, A. & Wu, C. (2009) An optimal retrospective change point detection policy. *Scandinavian Journal of Statistics*, 36(3), 542–558.

29 Vexler, A. & Gurevich, G. (2010) Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy. *Computational Statistics & Data Analysis*, 54(2), 531–545.

30 Wilks, S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1), 60–62.

31 Vexler, A., Yu, J. & Hutson, A.D. (2011) Likelihood testing populations modeled by autoregressive process subject to the limit of detection in applications to longitudinal biomedical data. *Journal of Applied Statistics*, 38(7), 1333–1346.

32 Fan, J., Zhang, C. & Zhang, J. (2001) Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics*, 153–193.

33 Ghosh, M. (1995) Inconsistent maximum likelihood estimators for the rasch model. *Statistics & Probability Letters*, 23(2), 165–170.

34 Gurevich, G. & Vexler, A. (2010) Retrospective change point detection: from parametric to distribution free policies. *Communications in Statistics-Simulation and Computation*, 39(5), 899–920.

35 Box, G.E.P., Hunter, J.S. & Hunter, W.G. (1978) *Statistics for Experimenters*. Wiley, pp. 144.

36 Pearson, E.S. & Neyman, J. (1930) *On the Problem of Two Samples*. Imprimerie de l'university.

37 Zhang, L., Xu, X. & Chen, G. (2012) The exact likelihood ratio test for equality of two normal populations. *The American Statistician*, 66(3), 180–184.

38 Wedderburn, R.W.M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, 61(3), 439–447.

39 Claeskens, G. & Hjort, N.L. (2004) Goodness of fit via non-parametric likelihood ratios. *Scandinavian Journal of Statistics*, 31(4), 487–513.

40 Wang, J. (2006) Quadratic artificial likelihood functions using estimating functions. *Scandinavian Journal of Statistics*, 33(2), 379–390.

41 Fan, J., Farmen, M. & Gijbels, I. (1998) Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(3), 591–608.

42 Owen, A. (1990) Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1), 90–120.

43 Qin, J. & Lawless, J. (1994) Empirical likelihood and general estimating equations. *The Annals of Statistics*, 300–325.

Q9

44 Lazar, N. & Mykland, P.A. (1998) An evaluation of the power and conditionality properties of empirical likelihood. *Biometrika*, 85(3), 523–534.
45 Vexler, A., Liu, S., Kang, L. *et al.* (2009) Modifications of the empirical likelihood interval estimation with improved coverage probabilities. *Communications in Statistics-Simulation and Computation*, 38(10), 2171–2183.
46 Yu, J., Vexler, A. & Tian, L. (2010) Analyzing incomplete data subject to a threshold using empirical likelihood methods: an application to a pneumonia risk study in an ICU setting. *Biometrics*, 66(1), 123–130.
47 Vexler, A., Yu, J., Tian, L. *et al.* (2010) Two-sample nonparametric likelihood inference based on incomplete data with an application to a pneumonia study. *Biometrical Journal*, 52(3), 348–361.
48 Owen, A.B. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2), 237–249.
49 Vexler, A., Gurevich, G. & Hutson, A.D. (2013) An exact density- based empirical likelihood ratio test for paired data. *Journal of Statistical Planning and Inference*, 143(2), 334–345.
50 Vexler, A., Shan, G., Kim, S. *et al.* (2011) An empirical likelihood ratio based goodness-of-fit test for inverse Gaussian distributions. *Journal of Statistical Planning and Inference*, 141(6), 2128–2140.
51 Gurevich, G. & Vexler, A. (2011) A two-sample empirical likelihood ratio test based on samples entropy. *Statistics and Computing*, 21(4), 657–670.
52 Vexler, A. & Yu, J. (2011) Two-sample density-based empirical likelihood tests for incomplete data in application to a pneumonia study. *Biometrical Journal*, 53(4), 628–651.
53 Vexler, A., Tsai, W.M., Gurevich, G. *et al.* (2012) Two-sample density-based empirical likelihood ratio tests based on paired data, with application to a treatment study of attention-deficit/hyperactivity disorder and severe mood dysregulation. *Statistics in Medicine*, 31(17), 1821–1837.
54 Yu, J., Vexler, A., Kim, S.E. *et al.* (2011) Two-sample empirical likelihood ratio tests for medians in application to biomarker evaluations. *Canadian Journal of Statistics*, 39(4), 671–689.
55 Hall, P. & Owen, A.B. (1993) Empirical likelihood confidence bands in density estimation. *Journal of Computational and Graphical Statistics*, 2(3), 273–289.
56 Einmahl, J.H.J. & McKeague, I.W. (2003) Empirical likelihood based hypothesis testing. *Bernoulli*, 9(2), 267–290.
57 Vexler, A., Tsai, W.M. & Malinovsky, Y. (2012) Estimation and testing based on data subject to measurement errors: from parametric to non-parametric likelihood methods. *Statistics in Medicine*, 31(22), 2498–2512.
58 Miecznikowski, J.C., Vexler, A. & Shepherd, L. (2013) dbemplikegof: An R package for nonparametric likelihood ratio tests for goodness-of-fit and two sample comparisons based on sample entropy. *Journal of Statistical Software*, 54(3), 1–19.
59 Vexler, A., Tanajian, H. & Hutson, A.D. (2014) Density-based empirical likelihood procedures for testing symmetry of data distributions and K-sample comparisons. *The Stata Journal*, 14(2), 304–328.
60 Vexler, A., Tsai, W.M. & Hutson, A.D. (2014) A simple density-based empirical likelihood ratio test for independence. *The American Statistician*, 68(3), 158–169.
61 Qin, J. & Zhang, B. (2005) Marginal likelihood, conditional likelihood and empirical likelihood: connections and applications. *Biometrika*, 92(2), 251–270.
62 Qin, J. (2000) Miscellanea. Combining parametric and empirical likelihoods. *Biometrika*, 87(2), 484–490.
63 Qin, J. & Leung, D.H.Y. (2005) A semiparametric two-component compound mixture model and its application to estimating malaria attributable fractions. *Biometrics*, 61(2), 456–464.
64 Green, D.M. & Swets, J.A. (1966) *Signal Detection Theory and Psychophysics*. Vol. 1. Wiley, New York.

Q10

65 Pepe, M.S. (1997) A regression modelling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika*, 84(3), 595–608.
66 Hsieh, F., Turnbull, B.W. *et al.* (1996) Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, 24(1), 25–40.
67 Wieand, S., Gail, M.H., James, B.R. *et al.* (1989) A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika*, 76(3), 585–592.
68 Pepe, M.S. & Thompson, M.L. (2000) Combining diagnostic test results to increase accuracy. *Biostatistics*, 1(2), 123–140.
69 McIntosh, M.W. & Pepe, M.S. (2002) Combining several screening tests: optimality of the risk score. *Biometrics*, 58(3), 657–664.
70 Bamber, D. (1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4), 387–415.
71 Kotz, S., Lumelskii, Y. & Pensky, M. (2003) The stress-strength model and its generalizations. In: *Theory and Applications*. World Scientific, Singapore.
72 Metz, C.E., Herman, B.A. & Shen, J.H. (1998) Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuousl-distributed data. *Statistics in Medicine*, 17(9), 1033–1053.
73 Box, G.E.P. & Cox, D.R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society, Series B (Methodological)*, 26(2), 211–252.
74 Zhou, X.H., Obuchowski, N.A. & McClish, D.K. (2011) *Statistical Methods in Diagnostic Medicine*. Vol. 712. John Wiley & Sons.
75 Sering, R.J. (2009) *Approximation Theorems of Mathematical Statistics*. Vol. 162. John Wiley & Sons.
76 Qin, G. & Zhou, X.H. (2006) Empirical likelihood inference for the area under the ROC curve. *Biometrics*, 62(2), 613–622.
77 Zou, K.H., Hall, W.J. & Shapiro, D.E. (1997) Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, 16(19), 2143–2156.
78 Pepe, M.S., Cai, T. & Longton, G. (2006) Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, 62(1), 221–229.
79 Su, J.Q. & Liu, J.S. (1993) Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association*, 88(424), 1350–1355.
80 Reiser, B. & Faraggi, D. (1997) Confidence intervals for the generalized ROC criterion. *Biometrics*, 53, 644–652.
81 Chen, X., Vexler, A. & Markatou, M. (2015) Empirical likelihood ratio confidence interval estimation of best linear combinations of biomarkers. *Computational Statistics & Data Analysis*, 82, 186–198.
82 Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Vol. 26. CRC Press.
83 Shapiro, S.S. & Wilk, M.B. (1965) An analysis of variance test for normality (complete samples). *Biometrika*, 52, 591–611.
84 Razali, N.M. & Wah, Y.B. (2011) Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1), 21–33.
85 Wilk, M.B. & Gnanadesikan, R. (1968) Probability plotting methods for the analysis of data. *Biometrika*, 55(1), 1–17.
86 Hettmansperger, T.P. & McKean, J.W. (1978) Statistical inference based on ranks. *Psychometrika*, 43(1), 69–79.
87 Kallenberg, W.C.M. & Ledwina, T. (1999) Data-driven rank tests for independence. *Journal of the American Statistical Association*, 94(445), 285–301.
88 Cox, D.R. & Hinkley, D.V. (1979) *Theoretical Statistics*. CRC Press.
89 Wheeler, B. (2013) *SuppDists: Supplementary distributions. R Package Version* 1.1-9.1. http://CRAN.R-project.org/package=SuppDists

90 Inglot, T., Kallenberg, W.C.M. & Ledwina, T. (1997) Data driven smooth tests for composite hypotheses. *The Annals of Statistics*, 25(3), 1222–1250.
91 Vexler, A., Kim, Y.M., Yu, J. *et al.* (2014) Computing critical values of exact tests by incorporating Monte Carlo simulations combined with statistical tables. *Scandinavian Journal of Statistics*, 41(4), 1013–1030.
92 Wilcox, R.R. (1998) The goals and strategies of robust methods. *British Journal of Mathematical and Statistical Psychology*, 51(1), 1–39.
93 Reiser, B. (2000) Measuring the effectiveness of diagnostic markers in the presence of measurement error through the use of ROC curves. *Statistic in Medicine*, 19, 2115–2129.
94 Jennison, C. & Turnbull, B.W. (2010) *Group Sequential Methods with Applications to Clinical Trials*. CRC Press.

Webliographies

http://onlinelibrary.wiley.com/doi/10.1002/sim.4467/suppinfo – R programs of realization of the two-sample density-based empirical likelihood ratio tests based on paired data.
http://cran.r-project.org/web/packages/dbEmpLikeNorm/ – The R package "dbEmpLikeNorm": Test for joint assessment of normality.
http://cran.r-project.org/web/packages/dbEmpLikeGOF/index.html – The R package "dbEmpLikeGOF" for nonparametric density-based likelihood ratio tests for goodness of fit and two-sample comparisons.
http://sphhp.buffalo.edu/biostatistics/research-and-facilities/software/stata.html – The STATA package entitled "Novel and efficient density-based empirical likelihood procedures for symmetry and k-sample comparisons."
http://www.sciencedirect.com/science/article/pii/S0167947314002710 – The R function for obtaining the empirical likelihood ratio confidence interval estimation of best linear combinations of biomarkers.