The RJafroc Significance Testing Book

Dev P. Chakraborty, PhD

2023-03-11

Contents

Preface

TBA

0.1 Rationale and Organization

- Intended as an online update to my print book (Chakraborty, 2017).
- All references in this book to `RJafroc` refer to the R package with that name (case sensitive) (Chakraborty and Zhai, 2022).
- Since its publication in 2017 `RJafroc`, on which the `R` code examples in the print book depend, has evolved considerably causing many of the examples to "break" if one uses the most current version of `RJafroc`. The code will still run if one uses `RJafroc` 0.0.1 but this is inconvenient and misses out on many of the software improvements made since the print book appeared.
- This gives me the opportunity to update the print book.
- The online book has been divided into 3 books.
  - The RJafrocQuickStartBook book.
  - The RJafrocRocBook book.
  - **This book:** RJafrocSigTestBook.
  - The RJafrocFrocBook book.

0.2 TBA Acknowledgements

Dr. Xuetong Zhai

Dr. Peter Phillips

Online Latex Editor at this site

Dataset contributors

0.3 Temporary comments

This is intended to allow successful builds when a needed file is not in the build. These are indicated by, for example:

Chapter `TempComment \@ref(proper-roc-models)`

Fix these on final release.

Chapter 1

Sources of AUC variability

1.1 TBA How much finished

60%

1.2 Introduction

In previous chapters the area AUC under the ROC plot was introduced as the preferred way of summarizing performance in the ROC task, as compared to a pair of sensitivity and specificity values. It can be estimated either non-parametrically, as in `TempComment Chapter \@ref(empirical-auc)`, or parametrically, as in Chapter `TempComment \@ref(binormal-model)`, and even better ways of estimating it are described in TBA Chapter 18 and Chapter 20.

Irrespective of how it is estimated AUC is a realization of a random variable, and as such, it is subject to sampling variability. Any measurement based on a finite number of samples from a parent population is subject to sampling variability. This is because no finite sample is unique: someone else conducting a similar study would, in general, obtain a different sample. [Case-sampling variability is estimated by the binormal model in the previous chapter. It is related to the sharpness of the peak of the likelihood function, TBA §6.4.4. The sharper that the peak, the smaller the case sampling variability. This chapter focuses on general sources of variability affecting AUC, regardless of how it is estimated, and other (i.e., not binormal model based) ways of estimating it.]

Here is an outline of this chapter. The starting point is the identification of different sources of variability affecting AUC estimates. Considered next is dependence of AUC on the case-set index $\{c\}$, $c = 1, 2, ..., C$. Considered next is estimating case-sampling variability of the empirical estimate of AUC by

an analytic method. This is followed by descriptions of two resampling-based methods, namely the bootstrap and the jackknife, both of which have wide applicability (i.e., they are not restricted to ROC analysis). The methods are demonstrated using `R` code and the implementation of a calibrated simulator is shown and used to demonstrate their validity, i.e., showing that the different methods of estimating variability agree. The dependence of AUC on reader expertise and modality is considered. An important source of variability, namely the radiologist's choice of internal sensory thresholds, is described. A cautionary comment is made regarding indiscriminate usage of empirical AUC as a measure of performance.

TBA Online Appendix 7.A describes coding of the bootstrap method; Online Appendix 7.B is the corresponding implementation of the jackknife method. Online Appendix 7.C describes implementation of the calibrated simulator for single-modality single-reader ROC datasets. Online Appendix 7.D describes the code that allows comparison of the different methods of estimating case-sampling variability.

1.3 Three sources of variability

Statistics deals with variability. Understanding sources of variability affecting AUC is critical to an appreciation of ROC analysis. Three sources of variability are identified in (Swets and Pickett, 1982): case sampling, between-reader and within-reader variability.

1. Consider a single reader interpreting different case samples. Case-sampling variability arises from the finite number of cases comprising the dataset, compared to the potentially very large population of cases. [If one could sample every case there exists and have them interpreted by the same reader, there would be no case-sampling variability and the poor reader's AUC values (from repeated interpretations of the entire population) would reflect only within reader variability, see #3 below.] Each case-set $\{c\}$, consisting of K_1 non-diseased and K_2 diseased cases interpreted by the reader, yields an AUC value. The notation $\{c\}$ means different *case sets*. Thus $\{c\} = \{1\}$, $\{2\}$, etc., denote different case sets, each consisting of K_1 non-diseased and K_2 diseased cases.

There is much "data compression" in going from individual case ratings to AUC. For a single reader and given case-set $\{c\}$, the ratings can be converted to an $A_{z\{c\}}$ estimate, TBA Eqn. (6.49). The notation shows explicitly the dependence of the measure on the case-set $\{c\}$. One can conceptualize the distribution of $A_{z\{c\}}$'s over different case-sets, each of the same size $K_1 + K_2$, as a normal distribution, i.e.,

$$A_{z\{c\}} \sim N(A_{z\{\bullet\}}, \sigma^2_{\text{cs+wr}}) \tag{1.1}$$

The dot notation $\{\bullet\}$ denotes an average over all case sets. Thus, $A_{z\{\bullet\}}$ is an estimate of the case-sampling mean of A_z for a single fixed reader and $\sigma^2_{\text{cs+wr}}$ is the *case sampling plus within-reader* variance. The reason for adding the within-reader variance is explained in #3 below. The concept is that a specified reader interpreting different case-sets effectively samples different parts of the population of cases, resulting in variability in measured A_z. Sometimes easier cases are sampled, and sometimes more difficult ones. This source of variability is expected to decrease with increasing case-set size, i.e., increasing $K_1 + K_2$, which is the reason for seeking large numbers of cases in clinical trials. Case-sampling and within-reader variability also decreases as the cases become more homogenous. An example of a more homogenous case sample would be cases originating from a small geographical region with, for example, limited ethnic variability. This is the reason for seeking multi-institutional clinical trials, because they tend to sample more of the population than patients seen at a single institution.

2. Consider different readers interpreting a fixed case sample. Between-reader variability arises from the finite number of readers compared to the population of readers; the population of readers could be all board certified radiologists interpreting screening mammograms in the US. This time one envisages different readers interpreting a fixed case set $\{1\}$. The different reader's $A_{z;j}$ values (j is the reader index, $j = 1, 2, ..., J$, where J is the total number of readers in the dataset) are distributed:

$$A_{z;j} \sim N(A_{z;\bullet}, \sigma^2_{\text{br+wr}}) \tag{1.2}$$

where $A_{z;\bullet}$ is an estimate of the reader population AUC mean (the bullet symbol replacing the reader index averages over a set of readers) for the fixed case-set $\{1\}$ and $\sigma^2_{\text{br+wr}}$ is the *between-reader plus within-reader* variance. The reason for adding the within-reader variance is explained in #3 below. The concept is that different groups of J readers interpret the same case set $\{1\}$, thereby sampling different parts of the reader distribution, causing fluctuations in the measured $A_{z;j}$ of the readers. Sometimes better readers are sampled and sometimes not so good ones are sampled. This time there is no "data compression" – each reader in the sample has an associated $A_{z;j}$. However, variability of the average $A_{z;\bullet}$ over the J readers is expected to decrease with increasing J. This is the reason for seeking large reader-samples.

3. Consider a fixed reader, e.g., $j = 1$, interpreting a fixed case-sample $\{1\}$. Within-reader variability is due to variability of the ratings for the same case: the same reader interpreting the same case on different occasions

will give different ratings to it, causing fluctuations in the measured AUC. This assumes that memory effects are minimized, for example, by sufficient time between successive interpretations as otherwise, if a case is shown twice in succession, the reader would give it the same rating each time. Since this is an intrinsic source of variability (analogous to the internal noise of a voltmeter) affecting each reader's interpretations, it cannot be separated from case sampling variability, i.e., it cannot be "turned off". The last sentence needs further explanation. A measurement of case-sampling variability requires a reader, and the reader comes with an intrinsic source of variability that gets added to the case-sampling variance, so what is measured is the sum of case sampling and within-reader variances, denoted $\sigma^2_{\text{cs+wr}}$. Likewise, a measurement of between-reader variability requires a fixed case-set interpreted by different readers, each of whom comes with an intrinsic source of variability that gets added to the between-reader variance, yielding $\sigma^2_{\text{br+wr}}$. To emphasize this point, an estimate of case-sampling variability *always* includes within reader variability. Likewise, an estimate of between-reader variability *always* includes within-reader variability.

With this background, the purpose of this chapter is to delve into variability in some detail and in particular describe computational methods for estimating them. This chapter introduces the concept of resampling a dataset to estimate variability and the widely used bootstrap and jackknife methods of estimating variance are described. In a later chapter, these are extended to estimating covariance (essentially a scaled version of the correlation) between two random variables.

The starting point is the simplest scenario: a single reader interpreting a case-set.

1.4 Dependence of AUC on the case sample

Suppose a researcher conducts a ROC study with a single reader. The researcher starts by selecting a case-sample, i.e., a set of proven-truth non-diseased and diseased cases. Another researcher conducting another ROC study at the same institution selects a different case-sample, i.e., a different set of proven-truth non-diseased and diseased cases. The two case-sets contain the same numbers K_1, K_2 of non-diseased and diseased cases, respectively. Even if the same radiologist interprets the two case-sets, and the reader is perfectly reproducible, the AUC values are expected to be different. Therefore, AUC must depend on a case sample index, which is denoted $\{c\}$, where c is an integer: $c = 1, 2$, as there are two case-sets in the study as envisaged.

$$AUC \rightarrow AUC_{\{c\}} \tag{1.3}$$

Note that $\{c\}$ is not an individual *case* index, rather it is a *case-set* index, i.e., different integer values of c denote different sets, or samples, or groups, or collections of cases. [The dependence of AUC on the case sample index is not explicitly shown in the literature.]

What does the dependence of AUC on the c index mean? Different case samples differ in their *difficulty* levels. A difficult case set contains a greater fraction of difficult cases than is usual. A difficult diseased case is one where disease is difficult to detect. For example, the lesions could be partly obscured by overlapping normal structures in the patient anatomy; i.e., the lesion does not "stick out". Alternatively, variants of normal anatomy could mimic a lesion, like a blood vessel viewed end on in a chest radiograph, causing the radiologist to miss the real lesion(s) and mistake these blood vessels for lesions. An easy diseased case is one where the disease is easy to detect. For example, the lesion is projected over smooth background tissue, because of which it "sticks out", or is more conspicuous2. How does difficulty level affect non-diseased cases? A difficult non-diseased case is one where variants of normal anatomy mimic actual lesions and could cause the radiologist to falsely diagnose the patient as diseased. Conversely, an easy non-diseased case is like a textbook illustration of normal anatomy. Every structure in it is clearly visualized and accounted for by the radiologist's knowledge of the patient's non-diseased anatomy, and the radiologist is confident that any abnormal structure, *if present*, would be readily seen. The radiologist is unlikely to falsely diagnose the patient as diseased. Difficult cases tend to be rated in the middle of the rating scale, while easy ones tend to be rated at the ends of the rating scale.

1.4.1 Case sampling variability of AUC

An easy case sample will cause AUC to increase over its average value; interpreting many case-sets and averaging the AUCs determines the average value. Conversely, a difficult case sample will cause AUC to decrease. Case sampling variability causes variability in the measured AUC. How does one estimate this essential source of variability? One method, totally impractical in the clinic but easy with simulations, is to have the same radiologist interpret repeated samples of case-sets from the population of cases (i.e., patients), termed *population sampling*, or more viscerally, as the "brute force" method.

Even if one could get a radiologist to interpret different case-sets, it is even more impractical to actually acquire the different case samples of truth-proven cases. Patients do not come conveniently labeled as non-diseased or diseased. Rather, one needs to follow-up on the patients, perhaps do other imaging tests, in order to establish true disease status, or ground-truth. In screening mammography, a woman who continues to be diagnosed as non-diseased on successive yearly screening tests in the US, and has no other symptoms of breast disease, is probably disease-free. Likewise, a woman diagnosed as diseased and the diagnosis is confirmed by biopsy (i.e., the biopsy comes back showing a malignancy in the

sampled tissues) is known to be diseased. However, not all patients who are diseased are actually diagnosed as diseased: a typical false negative fraction is 20% in screening mammography3. This is where follow-up imaging can help determine true disease status at the initial screen. A false negative mistake is unlikely to be repeated at the next screen. After a year, the tumor may have grown, and is more likely to be detected. Having detected the tumor in the most recent screen, radiologists can go back and retrospectively view it in the initial screen, at which it was missed during the "live" interpretation. If one knows where to look, the cancer is easier to see. The previous screen images would be an example of a difficult diseased case. In unfortunate instances, the patient may die from the previously undetected cancer, which would establish the truth status at the initial screen, too late to do the patient any good. The process of determining actual truth is often referred to as defining the "gold standard", the *ground truth*: or simply *truthing*.

One can appreciate from this discussion that acquiring independently proven cases, particularly diseased ones, is one of the most difficult aspects of conducting an observer performance study.

There has to be a better way of estimating case-sampling variability. With a parametric model, the maximum likelihood procedure provides a means of estimating variability of each of the estimated parameters, which can be used to estimate the variability of A_z, as in Chapter `TempComment \@ref(binormal-model)`. The estimate corresponds to case-sampling variability (including an inseparable within-reader variability). If unsure about this point, the reader should run some of the examples in Chapter `TempComment \@ref(binormal-model)` with increased numbers of cases. The variability is seen to decrease.

There are other options available for estimating case-sampling variance of AUC, and this chapter is not intended to be comprehensive. Three commonly used options are described: the DeLong et al method, the bootstrap and the jackknife resampling methods.

1.5 DeLong method

If the figure-of-merit is the empirical AUC, then a procedure developed by DeLong et al4 (henceforth abbreviated to DeLong) is applicable that is based on earlier work by (Noether, 1967) and (Bamber, 1975). The author will not go into details of this procedure but limit to showing that it "works". However, before one can show that it "works", one needs to know the true value of the variance of empirical AUC. Even if data were simulated using the binormal model, one cannot use the binormal model based estimate of variance as it is an estimate, not to be confused with a true value. Estimates are realizations of random numbers and are themselves subject to variability, which decreases with increasing case-set size. Instead, a "brute-force" (i.e., simulated population sampling) approach

is adopted to determine the true value of the variance of AUC. The simulator provides a means of repeatedly generating case-sets interpreted by the same radiologist, and by sampling it enough time, e.g., $C = 10{,}000$ times, each time calculating AUC, one determines the population mean and standard deviation. The standard deviation determined this way is compared to that yielded by the DeLong method to check if the latter actually works.

```
bruteForceEstimation <-
  function(seed, mu, sigma, K1, K2) {
    # brute force method to
    # find the population
    # meanempAuc and stdDevempAuc
    empAuc <- array(dim = 10000)
    for (i in 1:length(empAuc)) {
      zk1 <- rnorm(K1)
      zk2 <- rnorm(K2, mean = mu, sd = sigma)
      empAuc[i] <- Wilcoxon(zk1, zk2)
    }
    stdDevempAuc  <-  sqrt(var(empAuc))
    meanempAuc   <-  mean(empAuc)
    return(list(
      meanempAuc = meanempAuc,
      stdDevempAuc = stdDevempAuc
    ))
  }
```

```
seed <- 1;set.seed(seed)
mu <- 1.5;sigma <- 1.3;K1 <- 50;K2 <- 52
ret <- bruteForceEstimation(seed, mu, sigma, K1, K2)
# one more trial
zk1 <- rnorm(K1)
zk2 <- rnorm(K2, mean = mu, sd = sigma)
empAuc <- Wilcoxon(zk1, zk2)
ret1  <- DeLongVar(zk1,zk2)
stdDevDeLong <- sqrt(ret1)
cat("brute force estimates:",
    "\nempAuc = ",
    ret$meanempAuc,
    "\npopulation standard deviation =",
    ret$stdDevempAuc, "\n")
#> brute force estimates:
#> empAuc =  0.819178
#> population standard deviation = 0.04176683

cat("single sample estimates = ",
    "\nempirical AUC",
```

```
    empAuc,
    "\nstandard deviation DeLong = ",
    stdDevDeLong, "\n")
#> single sample estimates =
#> empirical AUC 0.8626923
#> standard deviation DeLong =  0.03804135
```

Two functions needed for this code to work are not shown: `Wilcoxon()` calculates the Wilcoxon statistic and the `DeLongVar()` implements the DeLong variance computation method (the DeLong method also calculates co-variances, but these are not needed in the current context). Line 1 sets the `seed` of the random number generator to 1. The `seed` variable is completely analogous to the case-set index `c`. Keeping `seed` fixed realizes the same random numbers each time the program is run. Different values of `seed` result in different, i.e., statistically independent, random samples. Line 2 initialize the values (μ, σ, K_1, K_2) needed by the data simulator: the normal distributions are separated by $\mu = 1.5$, the standard deviation of the diseased distribution is $\sigma = 1.3$, and there are $K_1 = 50$ non-diseased and $K_2 = 52$ diseased cases. Line 3 calls `bruteForceEstimation`, the "brute force" method for estimating mean and standard deviation of the population distribution of AUC, returned by this function, which are the "correct" value to which the DeLong standard deviation estimate will be compared. Lines 4-9 generates a fresh ROC dataset to which the DeLong method is applied.

Two runs of this code were made, one with the smaller sample size, and the other with 10 times the sample size (the second run takes much longer). A third run was made with the larger sample size but with a different seed value. The results follow:

```
seed <- 2;set.seed(seed)
mu <- 1.5;sigma <- 1.3;K1 <- 500;K2 <- 520
ret <- bruteForceEstimation(seed, mu, sigma, K1, K2)
# one more trial
zk1 <- rnorm(K1)
zk2 <- rnorm(K2, mean = mu, sd = sigma)
empAuc <- Wilcoxon(zk1, zk2)
ret1  <- DeLongVar(zk1,zk2)
stdDevDeLong <- sqrt(ret1)
cat("brute force estimates:",
    "\nempAuc = ",
    ret$meanempAuc,
    "\npopulation standard deviation =",
    ret$stdDevempAuc, "\n")
#> brute force estimates:
#> empAuc =  0.8194988
#> population standard deviation = 0.01300203
```

```
cat("single sample estimates = ",
    "\nempirical AUC",
    empAuc,
    "\nstandard deviation DeLong = ",
    stdDevDeLong, "\n")
#> single sample estimates =
#> empirical AUC 0.8047269
#> standard deviation DeLong =  0.01356696
```

1. An important observation is that as sample-size increases, case-sampling variability decreases: 0.0417 for the smaller sample size vs. 0.01309 for the larger sample size, and the dependence is as the inverse square root of the numbers of cases, as expected from the central limit theorem.
2. With the smaller sample size (K1/K2 = 50/52; the back-slash notation, not to be confused with division, is a convenient way of summarizing the case-sample size) the estimated standard deviation (0.038) is within 10% of that estimated by population sampling (0.042). With the larger sample size, (K1/K2 = 500/520) the two are practically identical (0.01300203 vs. 0.01356696 – the latter value is for seed = 2).
3. Notice also that the one sample empirical AUC for the smaller case-size is 0.863, which is less than two standard deviations from the population mean 0.819. The "two standard deviations" comes from rounding up 1.96: as in Eqn. `TempComment \@ref(eq:binary-task-model-def-z-alpha2)`, where $z_{\alpha/2}$ was defined as the upper $1 - \alpha/2$ quantile of the unit normal distribution and $z_{0.025} = 1.96$.
4. To reiterate, with clinical data the DeLong procedure estimates case sampling plus within reader variability. With simulated data as in this example, there is no within-reader variability as the simulator yields identical values for fixed seed.

This demonstration should convince the reader that one does have recourse other than the "brute force" method, at least when the figure of merit is the empirical area under the ROC. That should come as a relief, as population sampling is impractical in the clinical context. It should also impress the reader, as the DeLong method is able to use information present in a *single dataset* to tease out its variability. [This is not magic: the MLE estimate is also able to tease out variability based on a parametric fit to a single dataset and examination of the sharpness of the peak of the log-likelihood function, Chapter `TempComment \@ref(binormal-model)`, as are the resampling methods described next.]

Next, two resampling–based methods of estimating case-sampling variance of AUC are introduced. The word "resampling" means that the dataset itself is regarded as containing information regarding its variability, which can be extracted by sampling from the original data (hence the word "resampling").

These are general and powerful techniques, applicable to any scalar statistic, not just the empirical AUC, which one might be able to use in other contexts.

1.6 Bootstrap method

The simplest resampling method, at least at the conceptual level, is the bootstrap. *The bootstrap method is based on the assumption that one can regard the observed sample as defining the population from which it was sampled.* Since by definition a population cannot be exhausted, the idea is to resample, *with replacement*, from the observed sample. Each resampling step realizes a particular bootstrap sample set denoted $\{b\}$, where $b = 1, 2, ..., B$. The curly brackets emphasize that different integer values of b denote different *sets of cases*, not individual cases. [In contrast, the notation (k) will be used to denote *removing* a specific case, k, as in the jackknife procedure to be described shortly. The index b should not be confused with the index c, the case sampling index; the latter denotes repeated sampling from the population, which is impractical in real life; the bootstrap index denotes repeated sampling from the dataset, which is quite feasible.] The procedure is repeated B times, typically B can be as small as 200, but to be safe I generally use about 1000 - 2000 bootstraps. The following example uses Table `TempComment \@ref(tab:ratings-paradigm-example-table)` from Chapter `TempComment \@ref(ratings-paradigm)`.

For convenience, let us denote cases as follows. The 30 non-diseased cases that received the 1 rating are denoted $k_{1,1}, k_{2,1}, ..., k_{30,1}$. The second index denotes the truth state of the cases. Likewise, the 19 non-diseased cases that received the 2 rating are denoted $k_{31,1}, k_{32,1}, ..., k_{49,1}$ and so on for the remaining non-diseased cases. The 5 diseased cases that received the 1 rating are denoted $k_{1,2}, k_{2,2}, ..., k_{5,2}$, the 6 diseased cases that received the 2 rating are denoted $k_{6,2}, k_{7,2}, ..., k_{11,2}$, and so on. Let us figuratively "put" all non-diseased cases (think of each case as an index card, with the case notation and rating recorded on it) into one hat (the non-diseased hat) and all the diseased cases into another hat (the diseased hat). Next, one randomly picks one case (card) from the non-diseased hat, records it's rating, and puts the case back in the hat, so that it is free to be possibly picked again. This is repeated 60 times for the non-diseased hat resulting in 60 ratings from non-diseased cases. A similar procedure is performed using the diseased hat, resulting in 50 ratings from diseased cases. The author has just described, in painful detail (one might say) the realization of the 1st bootstrap sample, denoted $\{b = 1\}$. This is used to construct the 1st bootstrap counts table, Table 1.1.

So what happened? Consider the 35 non-diseased cases with a 1 rating. If each non-diseased case rated 1 in Table `TempComment \@ref(tab:ratings-paradigm-example-table)` were picked one time, the total would have been 30, but it is 35. Therefore, some of the original non-diseased cases rated 1 must have been picked multiple times, but one must also make allowance as there is no guarantee that a specific

Table 1.1: Representative counts table.

	$r = 5$	$r = 4$	$r = 3$	$r = 2$	$r = 1$
non-diseased	0	0	9	16	35
diseased	19	8	7	9	7

case was picked at all. Still focusing on the 35 non-diseased cases with a 1 rating in the first bootstrap sample, the picked labels, reordered after the fact, with respect to the first index, might be:

$$k_{2,1}, k_{2,1}, k_{4,1}, k_{4,1}, k_{4,1}, k_{6,1}, k_{7,1}, k_{7,1}, k_{9,1}, ..., k_{28,1}, k_{28,1}, k_{30,1}, k_{30,1} \qquad (1.4)$$

In this example, case $k_{1,1}$ was not picked, case $k_{2,1}$ was picked twice, case $k_{3,1}$ was not picked, case $k_{4,1}$ was picked three times, case $k_{5,1}$ was not picked, case $k_{6,1}$ was picked once, etc. The total number of cases in Eqn. (1.4) is 35, and similarly for the other cells in Table 1.1. Next, one estimates AUC for this table. Using the Eng website referred to earlier, one gets AUC = 0.843. [It is OK to use a parametric FOM since the bootstrap is a general procedure applicable, in principle, to any FOM, not just the empirical AUC, unlike the DeLong method, which is restricted to empirical AUC.] The corresponding value for the original data, Table `TempComment \@ref(tab:ratings-paradigm-example-table)`, was AUC = 0.870. The first bootstrapped dataset yielded a smaller value than the original dataset because one happened to have picked an unusually difficult bootstrap sample.

[Notice that in the original data there were 6 + 5 = 11 diseased cases that were rated 1 and 2, but in the bootstrapped dataset there are 7 + 9 = 16 diseased cases that were rated 1 and 2; in other words, the number of incorrect decisions on diseased cases went up, which would tend to lower AUC. Counteracting this effect is the increase in number of correct decisions on diseased cases: 8 + 19 = 27 cases rated 4 and 5, as compared to 12 + 22 = 34 in the original dataset. Reinforcing the effect is that increase in the number of correct decisions on non-diseased cases, albeit minimally: 35 + 16 = 51 rated 1 and 2 vs. 30 + 19 = 49 in the original dataset, and zero counts rated 4 and 5 in the non-diseased vs. 2 + 1 = 3 in the diseased. The complexity of following this *post-facto justification* illustrates the difficulty, in fact the futility, of correctly predicting which way performance will go from comparison of the two ROC counts tables – too many numbers are changing and in the above one did not even consider the change in counts in the bin labeled 4! Hence, the need for an objective figure of merit, such as the binormal model based AUC or the empirical AUC.]

To complete the description of the bootstrap method, one repeats the procedure described in the preceding paragraphs $B = 200$ times, each time running the website calculator and the final result is B values of AUC, denoted:

$$AUC_{\{1\}}, AUC_{\{2\}}, ..., AUC_{\{B\}}$$

where $AUC_{\{1\}} = 0.843$, etc. The bootstrap estimate of the variance of AUC is defined by (Efron and Tibshirani, 1993):

$$\text{Var}\left(AUC\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(AUC_{\{b\}} - AUC_{\{\bullet\}}\right)^2 \tag{1.5}$$

The right hand side is the traditional definition of (unbiased) variance. The dot represents the average over the *replaced index*. Of course, running the website code 200 times and recording the outputs is not a productive use of time. The following code implements two methods for estimating AUC, the empirical AUC, described in `TempComment Chapter \@ref(empirical-auc)` and the binormal model estimate of AUC, described in Chapter `TempComment \@ref(binormal-model)`.

1.6.1 Demonstration of the bootstrap method

To minimize clutter, several `R` functions are not shown, but they are compiled. To display them `clone` or 'fork' the book repository and look at the `Rmd` file corresponding to this output and the sourced `R` files listed below:

```
source(here("R/CH07-Variability/Transforms.R"))
source(here("R/CH07-Variability/LL.R"))
source(here("R/CH07-Variability/RocfitR.R"))
source(here("R/CH07-Variability/RocOperatingPoints.R"))
source(here("R/CH07-Variability/FixRocCountsTable.R"))
source(here("R/CH07-Variability/WilcoxonCountsTable.R"))
```

```
doBootstrap <- function(parametricFOM, B, seed, RocTable) {
  # this is the K vector
  K <- c(sum(RocTable[1,]), sum(RocTable[2,]))

  if (parametricFOM) {
    ret <- RocfitR(RocTable)
  } else {
    ret <- WilcoxonCountsTable(RocTable)
  }
  OrigAUC  <- ret$AUC

  # ready to bootstrap
  # first put the counts data into a linear array
  # convert counts table to array
```

```
  z1  <- rep(1:length(RocTable[1,]),
             RocTable[1,])
  z2  <- rep(1:length(RocTable[2,]),
             RocTable[2,])#do:
  AUC <- array(dim = B)#to save the bs AUC values
  for ( b in 1 : B){
    while (1) {
      RocTable_bs  <-
        array(dim = c(2,length(RocTable[1,])))
      # bs indices for non-diseased
      k1_b <- ceiling( runif( K[ 1 ] ) * K[ 1 ] )
      # bs indices for diseased
      k2_b <- ceiling( runif( K[ 2 ] ) * K[ 2 ] )
      bsTable <- table(z1[k1_b])
      #convert array to frequency table
      RocTable_bs[1,as.numeric(names(bsTable))] <-
        bsTable
      bsTable <- table(z2[k2_b])
      #do:
      RocTable_bs[2,as.numeric(names(bsTable))] <-
        bsTable
      #replace NAs with zeroes
      RocTable_bs[is.na(RocTable_bs )] <- 0
      if (parametricFOM) {
        temp <- RocfitR(RocTable_bs)
      } else {
        temp <- WilcoxonCountsTable(RocTable_bs)
      }
      AUC[b]  <- temp$AUC
      # a return of -1 means AUC did not converge
      if (AUC[b] != -1) break
    }
  }
  meanAUCboot <- mean(AUC)
  Var <- var(AUC)
  stdAUCboot  <- sqrt(Var)
  return(list(
    OrigAUC = OrigAUC,
    meanAUCboot = meanAUCboot,
    stdAUCboot = stdAUCboot
  ))
}
```

Since the bootstrap method is applicable to any scalar figure of merit, two options are provided in the code. If `parametricFOM` is set to `TRUE`, then the

binormal model estimate is used, and if set to `FALSE`, the empirical AUC is used. The first set of results are obtained with `parametricFOM` set to `TRUE`.

```
parametricFOM <- TRUE
B <- 200;seed <- 1;set.seed(seed)
RocTable = array(dim = c(2,5))
RocTable[1,]  <- c(30,19,8,2,1)
RocTable[2,]  <- c(5,6,5,12,22)

ret <- doBootstrap(parametricFOM, B, seed, RocTable)
OrigAUC <- ret$OrigAUC
meanAUCboot <- ret$meanAUCboot
stdAUCboot <- ret$stdAUCboot

cat("Bootstrap variance estimation:",
    "\nparametricFOM = ", parametricFOM,
    "\nseed = ", seed,
    "\nB = ", B,
    "\nOrigAUC = ", OrigAUC,
    "\nmeanAUCboot = ", meanAUCboot,
    "\nstdAUCboot = ", stdAUCboot, "\n")
#> Bootstrap variance estimation:
#> parametricFOM =  TRUE
#> seed =  1
#> B =  200
#> OrigAUC =  0.8704519
#> meanAUCboot =  0.8671713
#> stdAUCboot =  0.04380523
```

This shows that the AUC of the original data (i.e., before performing any bootstrapping) is 0.870, the mean AUC of the B = 200 bootstrapped datasets is 0.867, and the standard deviation of the 200 bootstraps is 0.0438. If one runs the website calculator referenced in the previous chapter on the dataset shown in Table `TempComment \@ref(tab:ratings-paradigm-example-table)`, one finds that the MLE of the standard deviation of the AUC of the fitted ROC curve is 0.0378. The standard deviation is itself a statistic and there is sampling variability associated with it, i.e., there exists such a beast as a standard deviation of a standard deviation; the bootstrap estimate is not too far from the MLE estimate. By setting `seed` to different values, one gets an idea of the variability of the estimate of the standard deviation of AUC. For example, with `seed` = 2, one gets:

```
#> Bootstrap variance estimation:
#> parametricFOM =  TRUE
#> seed =  2
```

```
#> B =  200
#> OrigAUC =  0.8704519
#> meanAUCboot =  0.8673155
#> stdAUCboot =  0.03815402
```

Note that both the mean of the bootstrap samples and the standard deviation have changed, but both are close to the MLE values. Examined next is the dependence of the estimates on `B`, the number of bootstraps. With `seed` = 1 and `B` = 2000 one gets:

```
#> Bootstrap variance estimation:
#> parametricFOM =  TRUE
#> seed =  1
#> B =  2000
#> OrigAUC =  0.8704519
#> meanAUCboot =  0.8674622
#> stdAUCboot =  0.03833508
```

The estimates are evidently rather insensitive to B, but the computation time was longer, ~13 seconds (running MLE 2000 times in 13 seconds is not bad!). It is always a good idea to test the stability of the results to different B and `seed` values. Unlike the DeLong et al method, which is restricted to the Wilcoxon statistic (which equals empirical AUC as per the Bamber theorem), the bootstrap is broadly applicable to other figures of merit, including non-ROC paradigm figures of merit. However, beware that it depends on the assumption that the sample itself is representative of the population. With limited numbers of cases, this could be a bad assumption. [With small numbers of cases it is relatively easy to enumerate the different outcomes of the sampling process and, more importantly, their respective probabilities, leading to what is termed the *exact bootstrap*. It is "exact" in the sense that there is no seed variable or number of bootstrap dependence.]

Finally, here is the output when using non-parametric AUC, with `seed` = 1.

```
#> Bootstrap variance estimation:
#> parametricFOM =  FALSE
#> seed =  1
#> B =  200
#> OrigAUC =  0.8606667
#> meanAUCboot =  0.8604575
#> stdAUCboot =  0.04125475
```

1.7 Jackknife method

The second resampling method, termed the *jackknife*, is computationally less demanding, but as was seen with the bootstrap, with modern personal computers computational limitations are no longer that important, at least for the types of analyses that this book is concerned with.

In this method, the first case is removed, or jackknifed, from the set of cases and the MLE (or empirical estimation) is conducted on the resulting dataset, which has one less case. Let us denote by $AUC_{(1)}$ the resulting value of AUC. The parentheses around the subscript 1 are meant to emphasize that the AUC value corresponds to that with the first case *removed* from the original dataset. Next, the first case is replaced, and now the second case is removed, the new dataset is analyzed yielding $AUC_{(2)}$, and so on, yielding K (K is the total number of cases; $K = K_1 + K_2$) *jackknife AUC values*:

$$AUC_{(k)} \qquad k = 1, 2, ..., K \tag{1.6}$$

The corresponding jackknife pseudovalues Y_k are defined by:

$$Y_k = K \times AUC - (K-1) \times AUC_{(k)} \tag{1.7}$$

Here AUC denotes the estimate using the entire dataset, i.e., not removing any cases. The jackknife pseudovalues will turn out to be of central importance in TBA Chapter 09. The *jackknife AUC values*, defined by Eqn. (1.6), should not be confused with jackknife derived psuedovalues, defined by Eqn. (1.7).

The jackknife estimate of the variance is defined by (Efron and Tibshirani, 1993):

$$\mathrm{Var}_{\mathrm{jack}} = \frac{(K-1)^2}{K} \frac{1}{K-1} \sum_{k=1}^{K} \left(AUC_{(k)} - AUC_{(\bullet)} \right)^2 \tag{1.8}$$

Since variance of K scalars is defined by:

$$\mathrm{Var}\,(x) = \frac{1}{K-1} \sum_{k=1}^{K} (x_k - x_\bullet)^2 \tag{1.9}$$

It follows that:

$$\mathrm{Var}_{\mathrm{jack}}\,(\mathrm{AUC}) = \frac{(K-1)^2}{K} \mathrm{Var}\,(\mathrm{AUC}) \tag{1.10}$$

In Eqn. (1.8) I have deliberately not simplified the right hand side by canceling out $K-1$. The purpose is to show, Eqn. (1.10), that the usual expression for

the variance (of the jackknife FOM values) needs to be multiplied by a **variance inflation factor** $\frac{(K-1)^2}{K}$, which is approximately equal to K, in order to obtain the correct jackknife estimate of variance of AUC. This factor was not necessary when one used the bootstrap method. That is because the bootstrap samples are more representative of the actual spread in the data. The jackknife samples are more restricted than the bootstrap samples, so the spread of the data is smaller; hence the need for the variance inflation factor (Efron and Tibshirani, 1993).

```
doJackknife <- function(parametricFOM, RocTable) {
  # this is the K vector
  K <- c(sum(RocTable[1,]), sum(RocTable[2,]))

  if (parametricFOM) {
    ret <- RocfitR(RocTable)
  } else {
    ret <- WilcoxonCountsTable(RocTable)
  }
  OrigAUC  <- ret$AUC

  # first put the counts data into a linear array
  z1  <- rep(1:length(RocTable[1,]),
             RocTable[1,])
  z2  <- rep(1:length(RocTable[1,]),
             RocTable[2,])

  AUC_jack  <- array(dim = sum(K))
  Y_k <- array(dim = sum(K))
  z_jk  <- array(dim = sum(K))
  # ready to jackknife
  for ( k in 1 : sum(K)){
    RocTable_jk  <- array(dim = c(2,length(RocTable[1,])))
    if ( k <= K[ 1 ]){
      z1_jk <- z1[ -k ]
      z2_jk <- z2
    }else{
      z1_jk <- z1
      z2_jk <- z2[ -(k - K[ 1 ]) ]
    }
    #convert array to frequency table
    RocTable_jk[1,1:length(table(z1_jk))]  <-
      table(z1_jk)
    RocTable_jk[2,1:length(table(z2_jk))]  <-
      table(z2_jk)
    #replace NAs with zeroes
```

```
    RocTable_jk[is.na(RocTable_jk)] <- 0
    # AUC_jack for observed data
    if (parametricFOM) {
      temp <- RocfitR(RocTable_jk)
    } else {
      temp <- WilcoxonCountsTable(RocTable_jk)
    }
    AUC_jack[k]  <- temp$AUC
    Y_k[k] <- sum(K)*OrigAUC - (sum(K)-1)*AUC_jack[k]
    if (AUC_jack[k] == -1)
      stop("RocfitR did not converge in jackknife loop")
  }
  meanAUCjack <- mean(AUC_jack)
  #Efron and Stein's paper, include jackknife inflation factor
  Var_jack <- var(AUC_jack) * ( sum(K) - 1)^2 / sum(K)
  stdAUCjack  <- sqrt(Var_jack)
  return(list(
    OrigAUC = OrigAUC,
    meanAUCjack = meanAUCjack,
    stdAUCjack = stdAUCjack
  ))
}
```

Since the jackknife method is applicable to any scalar figure of merit, two options are provided in the code. If `parametricFOM` is set to `TRUE`, then the binormal model estimate is used, and if set to `FALSE`, the empirical AUC is used. The first set of results are obtained with `parametricFOM` set to `TRUE`. Notice that the code does not use a `set.seed()` statement, as no random number generator is needed in the jackknife method. Systematically removing and replacing each case in sequence, one at a time, is not random sampling, which should further explain the need for the variance inflation factor in Eqn. (1.10).

```
parametricFOM <- TRUE
RocTable = array(dim = c(2,5))
RocTable[1,]  <- c(30,19,8,2,1)
RocTable[2,]  <- c(5,6,5,12,22)

ret <- doJackknife(parametricFOM, RocTable)
OrigAUC <- ret$OrigAUC
meanAUCjack <- ret$meanAUCjack
stdAUCjack <- ret$stdAUCjack

cat("Jackknife variance estimation:",
    "\nparametricFOM = ", parametricFOM,
    "\nOrigAUC = ", OrigAUC,
```

```
    "\nmeanAUCjack = ", meanAUCjack,
    "\nstdAUCjack = ", stdAUCjack, "\n")
#> Jackknife variance estimation:
#> parametricFOM =  TRUE
#> OrigAUC =  0.8704519
#> meanAUCjack =  0.8704304
#> stdAUCjack =  0.03861591
```

The next output is with the non-parametric figure of merit:

```
#> Jackknife variance estimation:
#> parametricFOM =  FALSE
#> OrigAUC =  0.8606667
#> meanAUCjack =  0.8606667
#> stdAUCjack =  0.03689264
```

It may be noticed that the mean of the jackknife figure of merit values, i.e., 0.8606667, exactly equals the original figure of merit 0.8606667 (i.e., that calculated including all cases). This can be shown analytically to be true so long as the figure of merit is the empirical AUC. A similar relation is not true for the bootstrap.

1.8 Calibrated simulator

1.8.1 The need for a calibrated simulator

The population sampling method used previously, 1.5, to compare the DeLong method to a known standard used arbitrarily set simulator values, i.e., $\mu = 1.5$ and $\sigma = 1.3$. One does not know if these values actually represent real clinical data. In this section a simple method of implementing population sampling using a *calibrated simulator* is described. A calibrated simulator is one whose parameters are chosen to match those of an actual clinical dataset. This way one has some assurance that the simulator is realistic and therefore its verdict on a proposed method or analysis (in our case method of estimating AUC variability) is likely to be correct.

1.8.2 Implementation of a simple calibrated simulator

The simple simulator described here is limited to a single reader single modality dataset. A more complex simulator describing multiple readers in multiple modalities is described in a later chapter (TBA). Consider a clinical dataset, such as in Table `TempComment \@ref(tab:ratings-paradigm-example-table)`.

Analyzed by MLE, this yields binormal model parameters, `a`, `b` and the thresholds $\zeta_1, \zeta_2, \zeta_3, \zeta_4$. After conversion to $\mu = a/b$ and $\sigma = 1/b$ and new zetas $\zeta = \zeta/b$, the values are (in the same order): 2.173597, 1.646099, 0.01263423, 1.475351, 2.494901, 3.945221 (see code output below):

```
# mu_sigma is the mu-sigma notation
mu_sigma <- c(2.173597, 1.646099, 0.01263423, 1.475351, 2.494901, 3.945221)
# ab is the a-b notation
ab <- c(1.320453, 0.607497, 0.007675259, 0.8962713, 1.515645, 2.39671)
ab[1]/ab[2] # this is mu
#> [1] 2.173596
1/ab[2]     # this is sigma
#> [1] 1.646099
ab[3:6]/ab[2] # this is zeta in mu-sigma notation
#> [1] 0.01263423 1.47535099 2.49490121 3.94522113
```

[The reason for dividing ζ by b is that when re-scaling the decision variable axis by b one must also re-scale the cutoffs.] The values μ, σ, ζ define the calibrated simulator, in the sense that the parameter values are calibrated to match the dataset in Table `TempComment \@ref(tab:ratings-paradigm-example-table)`.

Here is the function `doCalSimulator()` that will be used to perform the initial calibration followed by population sampling from the calibrated simulator:

```
1  doCalSimulator <- function(P, parametricFOM, RocCountsTable) {
2    K <- c(sum(RocCountsTable[1,]),
3           sum(RocCountsTable[2,]))
4    # perform the initial calibration
5    ret <- RocfitR(RocCountsTable) # AUC for observed data
6    a <- ret$a
7    b  <- ret$b
8    zetas <- ret$zeta
9    mu  <- a/b
10   sigma <- 1/b
11   zetas <- zetas/b # need to also scale zetas
12    # AUC for observed data
13   if (parametricFOM) {
14     OrigAUC <- RocfitR(RocCountsTable)$AUC
15   } else {
16     OrigAUC <- WilcoxonCountsTable(RocCountsTable)$AUC
17   }
18   # perform the population sampling
19   AUC <- array(dim = P)
20   for ( p in 1 : P){
21     while (1) {
```

```
22
23        RocCountsTableSimPop <-
24          SimulateRocCountsTable(K, mu, sigma, zetas)
25        if (parametricFOM) {
26
27          # AUC for fitted curve
28          temp <- RocfitR(RocCountsTableSimPop)
29          # a return of -1 means RocFitR did not converge
30          if (temp[1] != -1) {
31            AUC[p]  <- temp$AUC
32            break
33          }
34        } else {
35          AUC[p] <- (WilcoxonCountsTable(RocCountsTableSimPop))$AUC
36          break
37        }
38      }
39    }
40    AUC <- AUC[!is.na(AUC)]
41    meanAUC  <- mean(AUC)
42    stdAUC  <- sqrt(var(AUC))
43    return(list(
44      mu = mu, # these define the calibration simulator
45      sigma = sigma, #do:
46      zetas = zetas, #do:
47      OrigAUC = OrigAUC,
48      meanAUC = meanAUC,
49      stdAUC = stdAUC
50    ))
51  }
```

In the function `doCalSimulator(P, parametricFOM, RocCountsTable)`, `P` is the desired number of population samples, `parametricFOM` is a logical, if set to `TRUE` the binormal model is used to calculate *fitted* AUC and otherwise the Wilcoxon statistic is used to calculate *empirical* AUC, and `RocCountsTable` contains the ROC data, such as Table `TempComment \@ref(tab:ratings-paradigm-example-table)`, to which the simulator is to be calibrated to. Lines 2-3 construct the K-vector, containing K_1, K_2. Line 5 performs the maximum likelihood fit, using function `RocfitR(RocCountsTable)`. The returned variable contains a, b, ζ as a `list`, which are extracted at lines 6-8. Lines 9-11 converts these to the mu-sigma notation. In essence, lines 5 - 11 calibrates the simulator and the calibrated values of the simulator are contained in μ, σ, ζ. Lines 13-17 calculates `OrigAUC`, the AUC of the original data, using parametric `RocfitR` or the Wilcoxon statistic, as appropriate, depending on the value of `parametricFOM`. After

defining a length `P` array, at line 19, to hold the sampled `AUC` values, lines 20-39 begins and ends a `for` loop to conduct the `P` population samples. Each pass through the `for` loop yields K_1 samples from the non-diseased distribution and K_2 samples from the diseased distribution, returned in the variable `RocCountsTableSimPop`, which is similar in structure to a counts table like Table `TempComment \@ref(tab:ratings-paradigm-example-table)`. Within the `for` loop there is an endless `while` loop, needed because `RocfitR` can sometimes fail to converge, signaled by the first member of the returned `list` being minus 1, in which case another iteration of the `while` loop is performed (see line 30) and otherwise the `break` statement (line 32) causes program execution to proceed to the next iteration of the `for` loop. After entering the `while` loop, lines 22-23, a new ROC counts table is generated. The returned `list` is saved to `temp` at line 28, and if `temp[1] != -1` (i.e., `RocfitR` did converge) the AUC value is saved to `AUC[p]`, line 31. Upon exiting the code one has `P` values of AUC in the array `AUC`.

1.8.2.1 Parametric AUC results

The following code uses the function just described and prints out the results.

```
parametricFOM <- TRUE
seed <- 1
set.seed(seed)
P <- 2000
RocCountsTable = array(dim = c(2,5))
RocCountsTable[1,]  <- c(30,19,8,2,1)
RocCountsTable[2,]  <- c(5,6,5,12,22)
ret <- doCalSimulator(P, parametricFOM, RocCountsTable)
mu <- ret$mu
sigma <- ret$sigma
zetas <- ret$zetas
meanAUC_1_2000 <- ret$meanAUC
stdAUC_1_2000 <- ret$stdAUC
```

After setting `parametricFOM` to `TRUE` (for a parametric fit), `seed` to 1 and `P` to 2000, the ROC counts table is defined and the function `doCalSimulator()` is called. The returned `list` contains the parameter values for the calibrated simulator: $\mu = 2.1735969$, $\sigma = 1.6460988$ and $\zeta = 0.0126342, 1.4753512, 2.4949012, 3.9452209$. It also contains `OrigAUC`, the AUC of the original data, calculated by `RocfitR()`, in this case `OrigAUC` $= 0.8704519$, and the mean and standard deviation of the 2000 AUC values, equal to 0.8676727 and 0.0403331, respectively.

The simulations were repeated with `seed` $= 2$. This time the mean and standard deviation of the 2000 AUC values, are equal to 0.8681855 and 0.0405516,

respectively. The respective values corresponding to the two `seed` values are quite close to each other (to within a percent).

More variability is observed, as expected, when the above two simulations are repeated with `P` = 200:

For `seed` = 1 and `P` = 200 the mean and standard deviation of the 200 AUC values, are 0.8727151 and 0.0355281, respectively.

For `seed` = 2 and `P` = 200 the mean and standard deviation of the 200 AUC values, are 0.8649385 and 0.0450947, respectively. Note the greater variability induced by the change in `seed`, as compared to `P` = 2000.

1.8.2.2 Non-parametric AUC results

The next simulation is with `seed` = 1 and `P` = 2000, but this time `parametricFOM` is set to `FALSE`. The calibration proceeds as before, using `RocfitR` to determine the parameters of the simulation model, calibrating the simulator requires a parametric fit, but this empirical AUC is used to obtained the 2000 AUC samples. The mean and standard deviation of the AUC values, are 0.8497634 and 0.0367476, respectively. Note that these are smaller than the corresponding parametric estimates. The empirical AUC is expected to be smaller than the corresponding parametric AUC as joining adjacent points with straight lines will underestimate the area under the smooth ROC curve. Repeating with `seed` = 2, the mean and standard deviation of the AUC values, are 0.8503732 and 0.0369091, respectively, which are close to the `seed` = 1 values.

1.9 Discussion

This chapter focused on the factors affecting variability of AUC, namely case-sampling and between-reader variability, each of which contain an inseparable within-reader contribution. The only way to get an estimate of within-reader variability is to have the same reader re-interpret the same case-set on multiple occasions, after a sufficient time delay to minimize memory effects. This is rarely done and is unnecessary, in the ROC context, to sound experimental design and analysis. Some early publications have suggested that such re-interpretations are needed, but modern methods, described in the next part of the book, does not require re-interpretations. Indeed, it is a waste of precious reader-time resources. Rather than have the same readers re-interpret the same case-set on multiple occasions, it makes much more sense to recruit more readers and/or collect more cases, guided by a systematic sample size estimation method. Another reason I am not in favor of re-interpretations is that the within-reader variance is usually smaller than case-sampling and between-reader variances. Re-interpretations would minimize a quantity that is already small, which is not good practice.

The bootstrap and jackknife methods described in this chapter have wide applicability. Later they will be extended to estimating the covariance (essentially a scaled correlation) between two random variables. Also described was the DeLong method, applicable to the empirical AUC. Using a real dataset and simulators, all methods were shown to agree with each other, especially when the numbers of cases is large, Table 7.3 (row-D).

The concept of a calibrated simulator was introduced as a way of "anchoring" a simulator to a real dataset. While relatively easy for a single dataset, the concept has yet to be extended to where it would be useful, namely designing a simulator calibrated to a dataset consisting of interpretations by multiple readers in multiple modalities of a common dataset. Just as a calibrated simulator allowed comparison of the different variance estimation methods to a known standard, obtained by population sampling, a more general calibrated simulator would allow better testing the validity of the analysis described in the next few chapters.

This concludes Part A of this book. The next chapter begins Part B, namely the statistical analysis of multiple-reader multiple-case (MRMC) ROC datasets.

TBA: what to do with removed sections?

1.10 Chapter References

Significance Testing

Chapter 2

Hypothesis Testing

2.1 TBA How much finished

60%

2.2 Introduction

The problem addressed here is how to decide whether an estimate of AUC is consistent with a pre-specified value. One example of this is when a single-reader rates a set of cases in a single-modality, from which one estimates AUC, and the question is whether the estimate is statistically consistent with a pre-specified value. From a clinical point of view, this is generally not a useful exercise, but its simplicity is conducive to illustrating the broader concepts involved in this and later chapters. The clinically more useful analysis is when multiple readers interpret the same cases in two or more modalities. [With two modalities, for example, one obtains an estimate AUC for each reader in each modality, averages the AUC values over all readers within each modality, and computes the inter-modality difference in reader-averaged AUC values. The question forming the main subject of this book is whether the observed difference is consistent with zero.]

Each situation outlined above admits a binary (yes/no) answer, which is different from the estimation problem that was dealt with in connection with the maximum likelihood method in (book) Chapter 06, where one computed numerical estimates (and confidence intervals) of the parameters of the fitting model.

Hypothesis testing is the process of dichotomizing the possible outcomes of a statistical study and then using probabilistic arguments to choose one option over the other.

The two options are termed the *null hypothesis* (NH) and the *alternative hypothesis* (AH). The hypothesis testing procedure is analogous to the jury trial system in the US, with 20 instead of 12 jurors, with the NH being the presumption of innocence and the AH being the defendant is guilty. The decision rule is to assume the defendant is innocent unless all 20 jurors agree the defendant is guilty. If even one juror disagrees, the defendant is deemed innocent (equivalent to choosing an α – defined below - of 0.05, or 1/20).

2.3 Single-modality single-reader ROC study

The binormal model described in Chapter 06 can be used to generate sets of ratings to illustrate the methods being described in this chapter. To recapitulate, the model is described by:

$$Z_{k_1 1} \sim N(0,1)$$
$$Z_{k_2 2} \sim N(\mu, \sigma^2)$$

The following code chunk encodes the `Wilcoxon` function:

```
Wilcoxon <- function (zk1, zk2)
{
  K1 = length(zk1)
  K2 = length(zk2)
  W <- 0
  for (k1 in 1:K1) {
    W <- W + sum(zk1[k1] < zk2)
    W <- W + 0.5 * sum(zk1[k1] == zk2)
  }
  W <- W/K1/K2
  return (W)
}
```

In the next code chunk we set $\mu = 1.5$ and $\sigma = 1.3$ and simulate $K_1 = 50$ non-diseased cases and $K_2 = 52$ diseased cases. The `for`-loop draws 50 samples from the $N(0,1)$ distribution and 52 samples from the $N(\mu, \sigma^2)$ distribution, calculates the empirical AUC using the Wilcoxon, and the process is repeated 10,000 times, the AUC values are saved to a huge array `AUC_c` (the c-subscript is for case sample, where each case sample represents 102 cases). After exit from the `for`-loop we calculate the mean and standard deviation of the `AUC` values.

```
seed <- 1;set.seed(seed)
mu <- 1.5;sigma <- 1.3;K1 <- 50;K2 <- 52

# cheat to find the population mean and std. dev.
AUC_c <- array(dim = 10000)
for (c in 1:length(AUC_c)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu, sd = sigma)
  AUC_c[c] <- Wilcoxon(zk1, zk2)
}
meanAUC   <-  mean(AUC_c);sigmaAUC  <-  sd(AUC_c)
cat("pop mean AUC_c = ", meanAUC,
    ", pop sigma AUC_c = ", sigmaAUC, "\n")
#> pop mean AUC_c =  0.819178 , pop sigma AUC_c =  0.04176683
```

By the simple (if unimaginative) approach of sampling 10,000 times, one has estimates of the *population* mean and standard deviation of empirical AUC, denoted below by AUC_{pop} and σ_{AUC}, respectively.

The next code-chunk simulates one more independent ROC study with the same numbers of cases, and the resulting area under the empirical curve is denoted AUC in the code.

```
# one more trial, this is the one we want
# to compare to meanAUC
zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu, sd = sigma)
AUC <- Wilcoxon(zk1, zk2)
cat("New AUC = ", AUC, "\n")
#> New AUC =  0.8626923

z <- (AUC - meanAUC)/sigmaAUC
cat("z-statistic = ", z, "\n")
#> z-statistic =  1.04184
```

Is the new value, 0.8626923, sufficiently different from the population mean, 0.819178, to reject the null hypothesis NH : $\text{AUC} = \text{AUC}_{pop}$? Note that the answer to this question can be either yes or no: equivocation is not allowed!

The new value is "somewhat close" to the population mean, but how does one decide if "somewhat close" is close enough? Needed is the statistical distribution of the random variable AUC under the hypothesis that the true mean is AUC_{pop}. In the limit of a large number of cases, the pdf of AUC under the null hypothesis is a normal distribution $N\left(\text{AUC}_{pop}, \sigma^2_{\text{AUC}}\right)$:

$$\text{pdf}_{\text{AUC}}\left(\text{AUC} \mid \text{AUC}_{pop}, \sigma_{\text{AUC}}\right) = \frac{1}{\sigma_{\text{AUC}}\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{\text{AUC} - \text{AUC}_{pop}}{\sigma_{\text{AUC}}}\right)^2\right)$$

The translated and scaled value is distributed as a unit normal distribution, i.e.,

$$Z \equiv \frac{\text{AUC} - \text{AUC}_{pop}}{\sigma_{\text{AUC}}} \sim N(0,1)$$

[The Z notation here should not be confused with z-sample, decision variable or rating of a case in an ROC study; the latter, when sampled over a set of non-diseased and diseased cases, yield a realization of AUC. The author trusts the distinction will be clear from the context.] The observed magnitude of z is 1.0418397. [Upper-case for random variable, lower-case for realized or observed value.]

The ubiquitous p-value is the probability that the observed magnitude of z, or larger, occurs under the null hypothesis (NH) that the true mean of Z is zero. Stated somewhat differently, but equivalently, it is the probability that a random sample from $N(0,1)$ exceeds z.

The p-value corresponding to an observed z of 1.0418397 is given by:

$$\begin{aligned}\Pr(|Z| \geq |z| \mid Z \sim N(0,1)) &= \Pr(|Z| \geq 1.042 \mid Z \sim N(0,1)) \\ &= 2\Phi(-1.042) \\ &= 0.2975\end{aligned}$$

To recapitulate statistical notation, $\Pr(|Z| \geq |z| \mid Z \sim N(0,1))$ is parsed as $\Pr(A \mid B)$, that is, the probability $|Z| \geq |z|$ given that $Z \sim N(0,1)$. The second line in the preceding equation follows from the symmetry of the unit normal distribution, i.e., the area above 1.042 must equal the area below -1.042.

Since z is a continuous variable, there is zero probability that a sampled value will exactly equal the observed value. Therefore, one must pose the statement as above, namely the probability that Z is at least as extreme as the observed value (by "extreme" I mean further from zero, in either positive or negative directions). If the observed was $z = 2.5$ then the corresponding p-value would be $2\Phi(-2.5)$=0.01242, which is smaller than 0.2975. Under the zero-mean null hypothesis, the larger the magnitude of the observed value z, the smaller the p-value, and the more unlikely that the data supports the NH. **The p-value can be interpreted as the degree of unlikelihood that the data is consistent with the NH.**

By convention one adopts a fixed value of the probability, denoted and usually $\alpha = 0.05$, which is termed *the significance level* of the test, and the decision rule is to reject the null hypothesis if the observed p-value $< \alpha$. α is also referred to as the *size* of the test.

$$p < \alpha \Rightarrow \text{Reject NH}$$

If the p-value is exactly 0.05 (unlikely with ROC analysis, but one needs to account for it), then one does not reject the NH. In the 20-juror analogy, of one juror insists the defendant is not guilty, the observed p-value is 0.05, and one does not reject the NH that the defendant is innocent (the double negatives, very common in statistics, can be confusing; in plain English, the defendant goes home).

According to the previous discussion, the critical magnitude of z that determines whether to reject the null hypothesis is given by:

$$z_{\alpha/2} = -\Phi^{-1}(\alpha/2)$$

For $\alpha = 0.05$ this evaluates to 1.95996 (which is sometimes rounded up to two, good enough for "government work" as the saying goes) and the decision rule is to reject the null hypothesis only if the observed magnitude of z is larger than $z_{\alpha/2}$.

The decision rule based on comparing the observed z to a critical value is equivalent to a decision rule based on comparing the observed p-value to α. It is also equivalent, as will be shown later, to a decision rule based on a $(1-\alpha)$ confidence interval for the observed statistic. One rejects the NH if the closed confidence interval does not include zero.

2.4 Type-I errors

Just because one rejects the null hypothesis does not mean that the null hypothesis is false. Following the decision rule puts an upper limit on, or "caps", the probability of incorrectly rejecting the null hypothesis at α. In other words, by agreeing to reject the NH only if $p \leq \alpha$, one has set an upper limit, namely α, on errors of this type, termed *Type-I* errors. These could be termed false positives in the hypothesis testing sense, not to be confused with false positive occurring on individual case-level decisions. According to the definition of α:

$$\Pr(\text{Type I error} \mid \text{NH}) = \alpha$$

To demonstrate the ideas one needs to have a very cooperative reader interpreting new sets of independent cases not just one more time, but 2000 more times (the reason for the 2000 trials will be explained below). The simulation code follows:

```
seed <- 1;set.seed(seed)
mu <- 1.5;sigma <- 1.3;K1 <- 50;K2 <- 52
```

```
nTrials <- 2000
alpha <- 0.05 # size of test
reject = array(0, dim = nTrials)
for (trial in 1:length(reject)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu, sd = sigma)
  AUC <- Wilcoxon(zk1, zk2)
  z <- (AUC - meanAUC)/sigmaAUC
  p <- 2*pnorm(-abs(z)) # p value for individual trial
  if (p < alpha) reject[trial] = 1
}

CI <- c(0,0); width <- -qnorm(alpha/2)
ObsvdTypeIErrRate <- sum(reject)/length(reject)
CI[1] <- ObsvdTypeIErrRate -
  width*sqrt(ObsvdTypeIErrRate*(1-ObsvdTypeIErrRate)/nTrials)
CI[2] <- ObsvdTypeIErrRate +
  width*sqrt(ObsvdTypeIErrRate*(1-ObsvdTypeIErrRate)/nTrials)
cat("alpha = ", alpha, "\n")
#> alpha =  0.05
cat("ObsvdTypeIErrRate = ", ObsvdTypeIErrRate, "\n")
#> ObsvdTypeIErrRate =  0.0535
cat("95% confidence interval = ", CI, "\n")
#> 95% confidence interval =  0.04363788 0.06336212
exact <- binom.test(sum(reject),n = 2000,p = alpha)
cat("exact 95% CI = ", as.numeric(exact$conf.int), "\n")
#> exact 95% CI =  0.04404871 0.06428544
```

The population means were calculated in an earlier code chunk. One initializes `NTrials` to 2000 and α to 0.05. The `for-loop` describes our captive reader interpreting independent sets of cases 2000 times. *Each completed interpretation of 102 cases is termed a trial.* For each trial one calculates the observed value of `AUC`, the observed `z` statistic and the the observed p-value. The observed p-value is compared against the fixed value α and one sets the corresponding `reject[trial]` flag to unity if $p < \alpha$. In other words, if the trial-specific p-value is less than α one counts an instance of rejection of the null hypothesis. The process is repeated 2000 times.

Upon exit from the for-loop, one calculates the observed Type-I error rate, denoted `ObsvdTypeIErrRate` by summing the reject array and dividing by 2000. One calculates a 95% confidence interval for `ObsvdTypeIErrRate` based on the binomial distribution, as in (book) Chapter 03.

The observed Type-I error rate is a realization of a random variable, as is the estimated 95% confidence interval. The fact that the confidence interval includes $\alpha = 0.05$ is no coincidence - it shows that the hypothesis testing procedure is working as expected. To distinguish between the selected α (a fixed value) and

that observed in a simulation study (a realization of a random variable), the term *empirical* α is sometimes used to denote the observed rejection rate.

It is a mistake to state that one wishes to minimize the Type-I error probability. The minimum value of α (a probability) is zero. Run the software with this value of α: one finds that the NH is never rejected. The downside of minimizing the expected Type-I error rate is that the NH will never be rejected, even when the NH is patently false. The aim of a valid method of analyzing the data is not minimizing the Type-I error rate, rather, the observed Type-I error rate should equal the specified value of α (0.05 in our example), allowance being made for the inherent variability in it's estimate. This is the reason 2000 trials were chosen for testing the validity of the NH testing procedure. With this choice, the 95% confidence interval, assuming that observed value is close to 0.05, is roughly ± 0.01 as explained next.

Following analogous reasoning to (book) Chapter 03, Eqn. (3.10.10), and defining f as the observed rejection fraction over T trials, and as usual, F is a random variable and f a realized value,

$$\sigma_f = \sqrt{f(1-f)/T} F \sim N\left(f, \sigma_f^2\right)$$

An approximate $(1-\alpha)100$ percent CI for f is:

$$CI_f = \left[f - z_{\alpha/2}\sigma_f, f + z_{\alpha/2}\sigma_f\right]$$

If f is close to 0.05, then for 2000 trials, the 95% CI for f is $f \pm 0.01$, i.e., `qnorm(alpha/2) * sqrt(.05*(.95)/2000)` $= 0.009551683 \sim 0.01$.

The only way to reduce the width of the CI, and thereby run a more stringent test of the validity of the analysis, is to increase the number of trials T. Since the width of the CI depends on the inverse square root of the number of trials, one soon reaches a point of diminishing returns. Usually $T = 2000$ trials are enough for most statisticians and me, but studies using more simulations have been published.

2.5 One vs. two sided tests

The test described above is termed 2-tailed. Here, briefly, is the distinction between 2-tailed vs. 1-tailed p-values:

```
alpha  <- 0.05
# Example 1
# p value for two-sided AH
p2tailed <- pnorm(-abs(z)) + (1-pnorm(abs(z)))
```

```
cat("pvalue 2-tailed, AH: z ne 0 = ", p2tailed, "\n")
#> pvalue 2-tailed, AH: z ne 0 =  0.2943993

# Example 2
# p value for one-sided AH gt 0
p1tailedGT <- 1-pnorm(z)
cat("pvalue 1-tailed, AH: z gt 0 = ", p1tailedGT, "\n")
#> pvalue 1-tailed, AH: z gt 0 =  0.8528004

# Example 2
# p value for one-sided AH lt 0
p1tailedLT <- pnorm(z)
cat("pvalue 1-tailed, AH: z lt 0 = ", p1tailedGT, "\n")
#> pvalue 1-tailed, AH: z lt 0 =  0.8528004

df <- data.frame(p2tailed = p2tailed,
                 p1tailedGT = p1tailedGT,
                 p1tailedGT = p1tailedGT)
print(df)
#>    p2tailed p1tailedGT p1tailedGT.1
#> 1 0.2943993  0.8528004    0.8528004
```

The only difference between these tests is in how the alternative hypotheses is stated.

- For a two-tailed test the alternative hypothesis is AUC $\neq$ AUC$_{pop}$. Large deviations, in either direction, cause rejection of the NH.
- For the first one-tailed test the alternative hypothesis is AUC $>$ AUC$_{pop}$. Large positive observed values of z result in rejection of the NH. Large negative values do not.
- For the second one-tailed test the alternative hypothesis is AUC $<$ AUC$_{pop}$. Large negative observed values of z result in rejection of the NH. Large positive values do not.
- The last two statements are illustrated below with the following code-fragments:

```
# p1tailedGT
1-pnorm(1) # do not reject
#> [1] 0.1586553
1-pnorm(2) # reject
#> [1] 0.02275013
1-pnorm(-2) # do not reject
#> [1] 0.9772499
```

```
# p1tailedGT
pnorm(-1) # do not reject
#> [1] 0.1586553
pnorm(-2) # reject
#> [1] 0.02275013
pnorm(2) # do not reject
#> [1] 0.9772499
```

Note that the p-value of the 1-tailed tests are half that of the 2-tailed test. Further discussion of the difference between 2-tailed and 1-tailed tests, and when the latter might be appropriate, is given below.

If the null hypothesis is rejected anytime the magnitude of the observed value of z exceeded the critical value $-\Phi^{-1}(\alpha/2)$. This is a statement of the alternative hypothesis (AH) AUC $\neq$ AUC$_{pop}$, in other words too high or too low values of z *both* result in rejection of the null hypothesis. This is referred to as a two-sided AH and the resulting p-value is termed a *two-sided* p-value. This is the most common one used in the literature.

Suppose the additional trial performed by the radiologist was performed after an intervention following which the radiologist's performance is expected to increase. To make matters clearer, assume the interpretations in the 10,000 trials used to estimate AUC$_{pop}$ were performed with the radiologist wearing an old pair of eye-glasses, possibly out of proper strength, and the additional trial is performed after the radiologist gets a new set of prescription eye-glasses. Because the radiologist's eyesight has improved, the expectation is that performance should increase. In this situation, it is appropriate to use the one-sided alternative hypothesis AUC $>$ AUC$_{pop}$. Now, large positive values of z result in rejection of the null hypothesis, but large negative values do not. The critical value of z is defined by $z_\alpha = \Phi(1-\alpha)$, which for $\alpha = 0.05$ is 1.645 (i.e., `qnorm(1-alpha) = 1.644854`). Compare 1.64 to the critical value $-\Phi^{-1}(\alpha/2)$ = 1.96 for a two-sided test. If the change is in the expected direction, it is more likely that one will reject the NH with a one-sided than with a two-sided test. The p-value for a one-sided test is given by:

$$\Pr(Z \geq 1.042 \mid \text{NH}) = \Phi(-1.042) = 0.1487$$

Notice that this is half the corresponding two-sided test p-value; this is because one is only interested in the area under the unit normal that is above the observed value of z. If the intent is to obtain a significant finding, it is tempting to use one-sided tests. The down side of a one-sided test is that even with a large excursion of the observed z in the other direction one cannot reject the null hypothesis. So if the new eye-glasses are so bad as to render the radiologist practically blind (think of a botched cataract surgery) the observed z would be large and negative, but one cannot reject the null hypothesis AUC = AUC$_{pop}$.

The one-sided test could be run the other way, with the alternative hypothesis being stated as AUC $<$ AUC$_{pop}$. Now, large negative excursions of the observed value of AUC cause rejection of the null hypothesis, but large positive excursions do not. The critical value is defined by $z_\alpha = \Phi^{-1}(\alpha)$, which for $\alpha = 0.05$ is -1.645. The p-value is given by (note the reversed sign compared to the previous one-sided test:

$$\Pr(Z \leq 1.042 \mid \text{NH}) = \Phi(1.042) = 1 - 0.1487 = 0.8513$$

This is the complement of the value for a one-sided test with the alternative hypothesis going the other way: obviously the probability that Z is smaller than the observed value (1.042) plus the probability that Z is larger than the same value must equal one.

2.6 Statistical power

So far, focus has been on the null hypothesis. The Type-I error probability was introduced, defined as the probability of incorrectly rejecting the null hypothesis, the control, or "cap" on which is α, usually set to 0.05. What if the null hypothesis is actually false and the study fails to reject it? This is termed a Type-II error, the control on which is denoted β, the probability of a Type-II error. **The complement of β is called statistical power.**

The following table summarizes the two types of errors and the two correct decisions that can occur in hypothesis testing. In the context of hypothesis testing, a Type-II error could be termed a false negative, not to be confused with false negatives occurring on individual case-level decisions.

Truth	Fail to reject NH	Reject NH
NH is True	1 - α	α (FPF)
NH is False	β (FNF)	Power = 1 - β

This resembles the 2 x 2 table encountered in (book) Chapter 02, which led to the concepts of FPF, TPF and the ROC curve. Indeed, it is possible think of an analogous plot of empirical (i.e., observed) power vs. empirical α, which looks like an ROC plot, with empirical α playing the role of FPF and empirical power playing the role of TPF, see below. If $\alpha = 0$, then power = 0; i.e., if Type-I errors are minimized all the way to zero, then power is zero and one makes Type-II errors all the time. On the other hand, if $\alpha = 1$ then Power = 1, and one makes Type-I errors all the time.

A little history is due at this point. The author's first FROC study, which led to his entry into this field (Chakraborty et al., 1986), was published in Radiology

in 1986 after a lot of help from a reviewer, who we (correctly) guessed was the late Prof. Charles E. Metz. Prof. Gary T. Barnes (my mentor at that time at the University of Alabama at Birmingham) and I visited Prof. Charles Metz in Chicago for a day ca. 1986, to figuratively "pick Charlie's brain". Prof. Metz referred to the concept outlined in the previous paragraph, as an *ROC within an ROC*.

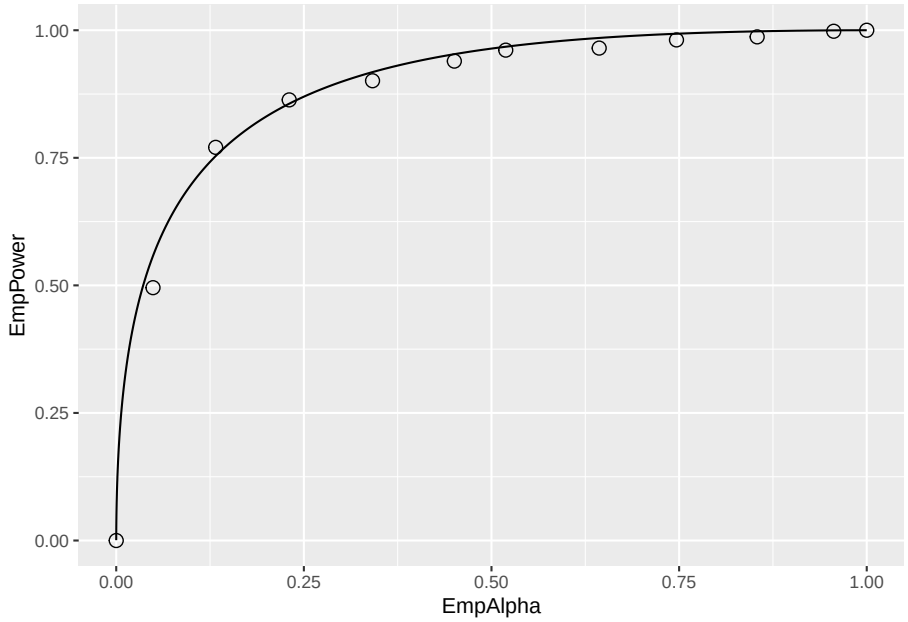
This curve does not summarize the result of a single ROC study. Rather it summarizes the probabilistic behavior of the two types of errors that occur when one conducts thousands of such studies, under both NH and AH conditions, each time with different values of α, with each trial ending in a decision to reject or not reject the null hypothesis. The long sentence is best explained with an example.

```
seed <- 1;set.seed(seed)
muNH <- 1.5;muAH <- 2.1;sigma <- 1.3;K1 <- 50;K2 <- 52# Line 6

# cheat to find the population mean and std. dev.
AUC <- array(dim = 10000) # line 8
for (i in 1:length(AUC)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = muNH, sd = sigma)
  AUC[i] <- Wilcoxon(zk1, zk2)
}
sigmaAUC <- sqrt(var(AUC));meanAUC <- mean(AUC) # Line 14

T <- 2000  # Line 16
mu <- c(muNH,muAH) # Line 17
alphaArr <- seq(0.05, 0.95, length.out = 10)
EmpAlpha <- array(dim = length(alphaArr))
EmpPower <- array(dim = length(alphaArr))
for (a in 1:length(alphaArr)) { # Line 20
  alpha <- alphaArr[a]
  reject <- array(0, dim = c(2, T))
  for (h in 1:2) {
    for (t in 1:length(reject[h,])) {
      zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu[h], sd = sigma)
      AUC <- Wilcoxon(zk1, zk2)
      obsvdZ <- (AUC - meanAUC)/sigmaAUC
      p <- 2*pnorm(-abs(obsvdZ)) # p value for individual t
      if (p < alpha) reject[h,t] = 1
    }
  }
  EmpAlpha[a] <- sum(reject[1,])/length(reject[1,])
  EmpPower[a] <- sum(reject[2,])/length(reject[2,])
}
EmpAlpha <- c(0,EmpAlpha,1); EmpPower <- c(0,EmpPower,1) # Line 19
```

```
pointData <- data.frame(EmpAlpha = EmpAlpha, EmpPower = EmpPower)
zetas <- seq(-5, 5, by = 0.01)
muRoc <- 1.8
curveData <- data.frame(EmpAlpha = pnorm(-zetas),
  EmpPower = pnorm(muRoc - zetas))
alphaPowerPlot <- ggplot(mapping = aes(x = EmpAlpha, y = EmpPower)) +
  geom_point(data = pointData, shape = 1, size = 3) +
  geom_line(data = curveData)
print(alphaPowerPlot)
```



Relevant line numbers are shown above as comments. Line 6 creates two variables, `muNH` = 1.5 (the binormal model separation parameter under the NH) and `muAH` = 2.1 (the separation parameter under the AH). Under either hypotheses, the same diseased case standard deviation `sigma` = 1.3 and 50 non-diseased and 52 diseased cases are assumed. As before, lines 8 – 14 use the "brute force" technique to determine population AUC and standard deviation of AUC under the NH condition. Line 16 defines the number of trials `T` = 2000. Line 17 creates a vector `mu` containing the NH and AH values defined at line 6. Line 18 creates `alphaArr`, a sequence of 10 equally spaced values in the range 0.05 to 0.95, which represent 10 values for α. Line 19 creates two arrays of length 10 each, named `EmpAlpha` and `EmpPower`, to hold the values of the observed Type-I error rate, i.e., empirical α, and the empirical power, respectively. The program will run `T` = 2000 NH and `T` = 2000 AH trials using as α each successive value in `alphaArr` and save the observed Type-I error rates and observed powers to

the arrays `EmpAlpha` and `EmpPower`, respectively.

Line 20 begins a for-loop in `a`, an index into `alphaArr.` Line 21 selects the appropriate value for `alpha` (0.05 on the first pass, 0.15 on the next pass, etc.). Line 22 initializes `reject[2,2000]` with zeroes, to hold the result of each trial; the first index corresponds to hypothesis `h` and the second to trial `t`. Line 23 begins a for-loop in `h`, with `h` = 1 corresponding to the NH and `h` = 2 to the AH. Line 24 begins a for-loop in `t`, the trial index. The code within this block is similar to previous examples. It simulates ratings, computes AUC, calculates the p-value, and saves a rejection of the NH as a one at the appropriate array location `reject[h,t]`. Lines 32 – 33 calculate the empirical α and empirical power for each value of α in `alphaArr`. After padding the ends with zero and ones (the trivial points), the remaining lines plot the "ROC within an ROC".

Each of the circles in the figure corresponds to a specific value of α. For example the lowest non-trivial corresponds to $\alpha = 0.05$, for which the empirical α is 0.049 and the corresponding empirical Power is 0.4955. True α increases as the operating point moves up the plot, with empirical α and empirical power increasing correspondingly. The AUC under this curve is determined by the effect size, defined as the difference between the AH and NH values of the separation parameter. If the effect size is zero, then the circles will scatter around the chance diagonal; the scatter will be consistent with the 2000 trials used to generate each coordinate of a point. As the effect size increases, the plot approaches the perfect "ROC", i.e., approaching the top-left corner. One could use AUC under this "ROC" as a measure of the incremental performance, the advantage being that it would be totally independent of α, but this would not be practical as it requires replication of the study under NH and AH conditions about 2000 times each and the entire process has to be repeated for several values of α. The purpose of this demonstration was to illustrate the concept behind Metz's profound remark.

It is time to move on to factors affecting statistical power in a single study.

2.6.1 Factors affecting statistical power

- Effect size: effect size is defined as the difference in AUC_{pop} values between the alternative hypothesis condition and the null hypothesis condition. Recall that AUC_{pop} is defined as the true or population value of the empirical ROC-AUC for the relevant hypothesis. One can use the "cheat method" to estimate it under the alternative hypothesis. The formalism is easier if one assumes it is equal to the asymptotic binormal model predicted value. The binormal model yields an estimate of the parameters, which only approach the population values in the asymptotic limit of a large number of cases. In the following, it is assumed that the parameters on the right hand side are the population values) It follows that effect size (ES) is given by (all quantities on the right hand side of Eqn. (8.13) are population values):

$$\text{AUC} = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right)$$

It follows that effect size (ES) is given by (all quantities on the right hand side of above equation are population values):

$$ES = \Phi\left(\frac{\mu_{AH}}{\sqrt{1+\sigma^2}}\right) - \Phi\left(\frac{\mu_{NH}}{\sqrt{1+\sigma^2}}\right)$$

```
EffectSize <- function (muNH, sigmaNH, muAH, sigmaAH)
{
  ES <- pnorm(muAH/sqrt(1+sigmaAH^2)) - pnorm(muNH/sqrt(1+sigmaNH^2))
  return (ES)
}

seed <- 1;set.seed(seed)
muAH <- 2.1 # NH value, defined previously, was mu = 1.5

T <- 2000
alpha <- 0.05 # size of test
reject = array(0, dim = T)
for (t in 1:length(reject)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = muAH, sd = sigma)
  AUC <- Wilcoxon(zk1, zk2)
  obsvdZ <- (AUC - meanAUC)/sigmaAUC
  p <- 2*pnorm(-abs(obsvdZ)) # p value for individual t
  if (p < alpha) reject[t] = 1
}

ObsvdTypeIErrRate <- sum(reject)/length(reject)
CI <- c(0,0);width <- -qnorm(alpha/2)
CI[1] <- ObsvdTypeIErrRate -
  width*sqrt(ObsvdTypeIErrRate*(1-ObsvdTypeIErrRate)/T)
CI[2] <- ObsvdTypeIErrRate +
  width*sqrt(ObsvdTypeIErrRate*(1-ObsvdTypeIErrRate)/T)
cat("obsvdPower = ", ObsvdTypeIErrRate, "\n")
#> obsvdPower =  0.489
cat("95% confidence interval = ", CI, "\n")
#> 95% confidence interval =  0.4670922 0.5109078
cat("Effect Size = ", EffectSize(mu, sigma, muAH, sigma), "\n")
#> Effect Size =  0.08000617 0
```

The ES for the code above is 0.08 (in AUC units). It should be obvious that if effect size is zero, then power equals α. This is because then there is no

distinction between the null and alternative hypotheses conditions. Conversely, as effect size increases, statistical power increases, the limiting value being unity, when every trial results in rejection of the null hypothesis. The reader should experiment with different values of `muAH` to be convinced of the truth of these statements.

- Sample size: increase the number of cases by a factor of two, and run the above code chunk.

```
#> pop NH mean AUC =  0.8594882 , pop NH sigma AUC =  0.02568252
#> num. non-diseased images =  100 num. diseased images =  104
#> obsvdPower =  0.313
#> 95% confidence interval =  0.2926772 0.3333228
#> Effect Size =  0.08000617 0
```

So doubling the numbers of cases (both non-diseased and diseased) results in statistical power increasing from 0.509 to 0.844. Increasing the numbers of cases decreases σ_{AUC}, the standard deviation of the empirical AUC. The new value of σ_{AUC} is 0.02947, which should be compared to the value 0.04177 for K1 = 50, K2 = 52. Recall that σ_{AUC} enters the denominator of the Z-statistic, so decreasing it will increase the probability of rejecting the null hypothesis.

- Alpha: Statistical power depends on *alpha*. The results below are for two runs of the code, the first with the original value $\alpha = 0.05$, the second with $\alpha = 0.01$:

```
#> alpha =  0.05 obsvdPower =  0.1545
#> alpha =  0.01 obsvdPower =  0.0265
```

Decreasing α results in decreased statistical power.

2.7 Comments

The Wilcoxon statistic was used to estimate the area under the ROC curve. One could have used the binormal model, introduced in Chapter 06, to obtain maximum likelihood estimates of the area under the binormal model fitted ROC curve. The reasons for choosing the simpler empirical area are as follows. (1) With continuous ratings and 102 operating points, the area under the empirical ROC curve is expected to be a close approximation to the fitted area. (2) With maximum likelihood estimation, the code would be more complex – in addition to the fitting routine one would require a binning routine and that would introduce yet another variable in the analysis, namely the number of

bins and how the bin boundaries were chosen. (3) The maximum likelihood fitting code can sometimes fail to converge, while the Wilcoxon method is always guaranteed to yield a result. The non-convergence issue is overcome by modern methods of curve fitting described in later chapters. (4) The aim was to provide an understanding of null hypothesis testing and statistical power without being bogged down in the details of curve fitting.

2.8 Why alpha is chosen as 5%

One might ask why α is traditionally chosen to be 5%. It is not a magical number, rather the result of a cost benefit tradeoff. Choosing too small a value of α would result in greater probability $(1-\alpha)$ of the NH not being rejected, even when it is false. Sometimes it is important to detect a true difference between the measured AUC and the postulated value. For example, a new eye-laser surgery procedure is invented and the number of patients is necessarily small as one does not wish to subject a large number of patients to an untried procedure. One seeks some leeway on the Type-I error probability, possibly increasing it to $\alpha = 0.1$, in order to have a reasonable chance of success in detecting an improvement in performance due to better eyesight after the surgery. If the NH is rejected and the change is in the right direction, then that is good news for the researcher. One might then consider a larger clinical trial and set α at the traditional 0.05, making up the lost statistical power by increasing the number of patients on which the surgery is tried.

If a whole branch of science hinges on the results of a study, such as discovering the Higg's Boson in particle physics, statistical significance is often expressed in multiples of the standard deviation (σ) of the normal distribution, with the significance threshold set at a much stricter level (e.g. 5σ). This corresponds to $\alpha \sim 1$ in 3.5 million (`1/pnorm(-5)` = 3.5 x 10^6, a one-sided test of significance). There is an article in Scientific American (https://blogs.scientificamerican.com/observations/five-sigmawhats-that/) on the use of $n\sigma$, where `n` is an integer, e.g. 5, to denote the significance level of a study, and some interesting anecdotes on why such high significance levels (ie., small α) are used in some fields of research.

Similar concerns apply to manufacturing where the cost of a mistake could be the very expensive recall of an entire product line. For background on Six Sigma Performance, see http://www.six-sigma-material.com/Six-Sigma.html. An article downloaded 3/30/17 from https://en.wikipedia.org/wiki/Six_Sigma is included as supplemental material to this chapter (Six Sigma.pdf). It has an explanation of why 6σ translates to one defect per 3.4 million opportunities (it has to do with short-term and long-term drifts in a process). In my opinion, looking at other fields offers a deeper understanding of this material than simply stating that by tradition one adopts alpha = 5%.

Most observer performance studies, while important in the search for better

imaging methods, are not of such "earth-shattering" importance, and it is somewhat important to detect true differences at a reasonable alpha, so alpha = 5% and beta = 20% represent a good compromise. If one adopted a 5σ criterion, the NH would never be rejected, and progress in image quality optimization would come to a grinding halt. That is not to say that a 5σ criterion cannot be used; rather if used, the number of patients needed to detect a reasonable difference (effect size) with 80% probability would be astronomically large. Truth-proven cases are a precious commodity in observer performance studies. Particle physicists working on discovering the Higg's Boson can get away with 5σ criterion because the number of independent observations and/or effect size is much larger than corresponding numbers in observer performance research.

2.9 Discussion

In most statistics books, the subject of hypothesis testing is demonstrated in different (i.e., non-ROC) contexts. That is to be expected since the ROC-analysis field is a small sub-specialty of statistics (Prof. Howard E. Rockette, private communication, ca. 2002). Since this book is about ROC analysis, I decided to use a demonstration using ROC analysis. Using a data simulator, one can "cheat" by conducting a very large number of simulations to estimate the population AUC under the null hypothesis. This permitted us to explore the related concepts of Type-I and Type-II errors within the context of ROC analysis. Ideally, both errors should be zero, but the nature of statistics leads one to two compromises. Usually one accepts a Type-I error capped at 5% and a Type-II error capped at 20%. These translate to $\alpha = 0.05$ and desired statistical power = 80%. The dependence of statistical power on α, the numbers of cases and the effect size was explored.

In TBA Chapter 11 sample-size calculations are described that allow one to estimate the numbers of readers and cases needed to detect a specified difference in inter-modality AUCs with expected statistical power = $1 - \beta$. The word "detect" in the preceding sentence is shorthand for "reject the NH with incorrect rejection probability capped at α".

This chapter also gives the first example of validation of a hypothesis testing method. Statisticians sometimes refer to this as showing a proposed test is a "5% test". What is meant is that one needs to be assured that when the NH is true the probability of NH rejection is consistent with the expected value. Since the observed NH rejection rate over 2000 simulations is a random variable, one does not expect the NH rejection rate to exactly equal 5%, rather the constructed 95% confidence interval (also a random interval variable) should include the NH value with probability 1 - α.

Comparing a single reader's performance to a specified value is not a clinically interesting problem. The next few chapters describe methods for significance testing of multiple-reader multiple-case (MRMC) ROC datasets, consisting of

interpretations by a group of readers of a common set of cases in typically two modalities. It turns out that the analyses yield variability estimates that permit sample size calculation. After all, sample size calculation is all about estimation of variability, the denominator of the z-statistic. The formulae will look more complex, as interest is not in determining the standard deviation of AUC, but in the standard deviation of the inter-modality reader-averaged AUC difference. However, the basic concepts remain the same.

2.10 Chapter References

Chapter 3

DBM method background

3.1 TBA How much finished

80%

3.2 Introduction

The term *treatment* is generic for *imaging system*, *modality* or *image processing*; *reader* is generic for *radiologist* or *algorithmic observer*, e.g., a computer aided detection (CAD) or artificial intelligence (AI) algorithm. The previous chapter described analysis of a single ROC dataset and comparing the observed area AUC under the ROC plot to a specified value. Clinically this is not an interesting problem; rather, interest is usually in comparing performance of a group of readers interpreting a common set of cases in two or more treatments. Such data is termed multiple reader multiple case (MRMC). [An argument could be made in favor of the term "multiple-treatment multiple-reader", since "multiple-case" is implicit in any ROC analysis that takes into account correct and incorrect decisions on cases. However, I will stick with existing terminology.] The basic idea is that by sampling a sufficiently large number of readers and cases one can draw conclusions that apply broadly to other readers of similar skill levels interpreting other similar case sets in the selected treatments. How one accomplishes this, termed MRMC analysis, is the subject of this chapter.

This chapter describes the first truly successful method of analyzing MRMC ROC data, namely the Dorfman-Berbaum-Metz (DBM) method (Dorfman et al., 1992). The other method, due to Obuchowski and Rockette (Obuchowski and Rockette, 1995), is the subject of Chapter 10 (TBA). Both methods have been substantially improved by Hillis (Hillis et al., 2008; Hillis, 2007, 2014). It is not an overstatement that ROC analysis came of age with the methods

described in this chapter. Prior to the techniques described here, one knew of the existence of sources of variability affecting a measured *AUC* value, as discussed in (book) Chapter 07, but then-known techniques (Swets and Pickett, 1982) for estimating the corresponding variances and correlations were impractical.

3.2.1 Historical background

The author was thrown (unprepared) into the methodology field ca. 1985 when, as a junior faculty member, he undertook comparing a prototype digital chest-imaging device (Picker International, ca. 1983) vs. an optimized analog chest-imaging device at the University of Alabama at Birmingham. At the outset a decision was made to use free-response ROC methodology instead of ROC, as the former accounted for lesion localization, and I and my mentor, Prof. Gary T. Barnes, were influenced in that decision by a publication (Bunch et al., 1977) to be described in (book) Chapter 12. Therefore, instead of ROC-AUC one had lesion-level sensitivity at a fixed number of location level false positives per case as the figure-of-merit (FOM). Details of the FOM are not relevant at this time. Suffice to state that methods described in this chapter, which had not been developed in 1983, while developed for analyzing reader-averaged inter-treatment ROC-AUC differences, *apply to any scalar FOM*. While I was successful at calculating confidence intervals (this is the heart of what is loosely termed *statistical analysis*) and publishing the work (Chakraborty et al., 1986) using techniques described in a book (Swets and Pickett, 1982) titled "Evaluation of Diagnostic Systems: Methods from Signal Detection Theory", subsequent attempts at applying these methods in a follow-up paper (Niklason et al., 1986) led to negative variance estimates (private communication, Dr. Loren Niklason, ca. 1985). With the benefit of hindsight, negative variance estimates are not that uncommon and the method to be described in this chapter has to deal with that possibility.

The methods (Swets and Pickett, 1982) described in the cited book involved estimating the different variability components – case sampling, between-reader and within-reader variability. Between-reader and within-reader variability (the two cannot be separated as discussed in (book) Chapter 07) could be estimated from the variance of the *AUC* values corresponding to the readers interpreting the cases within a treatment and then averaging the variances over all treatments. Estimating case-sampling and within-reader variability required splitting the dataset into a few smaller subsets (e.g., a case set with 60 cases might be split into 3 sub-sets of 20 cases each), analyzing each subset to get an *AUC* estimate, calculating the variance of the resulting *AUC* values (Swets and Pickett, 1982) and scaling the result to the original case size. Because it was based on few values, the estimate was inaccurate, and the already case-starved original dataset made it difficult to estimate AUCs for the subsets; moreover, the division into subsets was at the discretion of the researcher, and therefore unlikely to be

reproduced by others. Estimating within-reader variability required re-reading the entire case set, or at least a part of it. ROC studies have earned a deserved reputation for taking much time to complete, and having to re-read a case set was not a viable option. [Historical note: I recalls a barroom conversation with Dr. Thomas Mertelmeir after the conclusion of an SPIE meeting ca. 2004, where Dr. Mertelmeir commiserated mightily, over several beers, about the impracticality of some of the ROC studies required of imaging device manufacturers by the FDA.]

3.2.2 The Wagner analogy

An important objective of modality comparison studies is to estimate the variance of the difference in reader-averaged AUCs between the treatments. For two treatments one sums the reader-averaged variance in each treatment and subtracts twice the covariance (a scaled version of the correlation). Therefore, in addition to estimating variances, one needs to estimate correlations. Correlations are present due to the common case set interpreted by the readers in the different treatments. If the correlation is large, i.e., close to unity, then the individual treatment variances tend to cancel, making the constant treatment-induced difference easier to detect. The author recalls a vivid analogy used by the late Dr. Robert F. Wagner to illustrate this point at an SPIE meeting ca. 2008. To paraphrase him, *consider measuring from shore the heights of the masts on two adjacent boats in a turbulent ocean. Because of the waves, the heights, as measured from shore, are fluctuating wildly, so the variance of the individual height measurements is large. However, the difference between the two heights is likely to be relatively constant, i.e., have small variance. This is because the wave that causes one mast's height to increase also increases the height of the other mast.*

3.2.3 The shortage of numbers to analyze and a pivotal breakthrough

The basic issue was that the calculation of AUC reduces the relatively large number of ratings of a set of non-diseased and diseased cases to a single number. For example, after completion of an ROC study with 5 readers and 100 non-diseased and 100 diseased cases interpreted in two treatments, the data is reduced to just 10 numbers, i.e., five readers times two treatments. It is difficult to perform statistics with so few numbers. The author recalls a conversation with Prof. Kevin Berbaum at a Medical Image Perception Society meeting in Tucson, Arizona, ca. 1997, in which he described the basic idea that forms the subject of this chapter. Namely, using jackknife pseudovalues (to be defined below) as individual case-level figures of merit. This, of course, greatly increases the amount of data that one can work with; instead of just 10 numbers one now has 2,000 pseudovalues (2 x 5 x 200). If one assumes the pseudovalues

behave essentially as case-level data, then by assumption they are independent and identically distributed, and therefore satisfy the conditions for application of standard analysis of variance (ANOVA) techniques. [This assumption has been much criticized and is the basis for some preferring alternate approaches - but, as Hillis has stated, and I paraphrase, the pseudovalue based method "works", but lacks sufficient rigor.] The relevant paper had already been published in 1992 but other projects and lack of formal statistical training kept me from fully appreciating this work until later.

For the moment I restrict to fully paired data (i.e., each case is interpreted by all readers in all treatments). There is a long history of how this field has evolved and I cannot do justice to all methods that are currently available. Some of the methods (Toledano, 2003; Ishwaran and Gatsonis, 2000; Toledano and Gatsonis, 1996) have the advantage that they can handle explanatory variables (termed covariates) that could influence performance, e.g., years of experience, types of cases, etc. Other methods are restricted to specific choices of FOM. Specifically, the probabilistic approach (Clarkson et al., 2006; Kupinski et al., 2006; Gallas et al., 2007; Gallas, 2006) is restricted to the empirical *AUC* under the ROC curve, and is not applicable to other FOMs, e.g., parametrically fitted ROC AUCs or, more importantly, to location specific paradigm FOMs. Instead, I will focus on methods for which software is readily available (i.e., freely on websites), which have been widely used (the method that I am about to describe has been used in several hundred publications) and validated via simulations, and which apply to any scalar figure of merit, and therefore widely applicable, for example, to location specific paradigms.

3.2.4 Organization of chapter

The concepts of reader and case populations, introduced in (book) Chapter 07, are recapitulated. A distinction is made between *fixed* and *random* factors – statistical terms with which one must become familiar. Described next are three types of analysis that are possible with MRMC data, depending on which factors are regarded as random and which as fixed. The general approach to the analysis is described. Two methods of analysis are possible: the jackknife pseudovalue-based approach detailed in this chapter and an alternative approach is detailed in Chapter 10. The Dorfman-Berbaum-Metz (DBM) model for the jackknife pseudovalues is described that incorporates different sources of variability and correlations possible with MRMC data. Calculation of ANOVA-related quantities, termed mean squares, from the pseudovalues, are described followed by the significance testing procedure for testing the null hypothesis of no treatment effect. A relevant distribution used in the analysis, namely the F-distribution, is illustrated with R examples. The decision rule, i.e., whether to reject the NH, calculation of the ubiquitous p-value, confidence intervals and how to handle multiple treatments is illustrated with two datasets, one an older ROC dataset that has been widely used to demonstrate advances

in ROC analysis, and the other a recent dataset involving evaluation of digital chest tomosynthesis vs. conventional chest imaging. The approach to validation of DBM analysis is illustrated with an R example. The chapter concludes with a section on the meaning of the pseudovalues. The intent is to explain, at an intuitive level, why the DBM method "works", even though use of pseudovalues has been questioned at the conceptual level. For organizational reasons and space limitations, details of the software are relegated to Online Appendices, but they are essential reading, preferably in front of a computer running the online software that is part of this book. The author has included material here that may be obvious to statisticians, e.g., an explanation of the Satterthwaite approximation, but are expected to be helpful to others from non-statistical backgrounds.

3.3 Random and fixed factors

This paragraph introduces some analysis of variance (ANOVA) terminology. Treatment, reader and case are factors with different numbers of levels corresponding to each factor. For an ROC study with two treatments, five readers and 200 cases, there are two levels of the treatment factor, five levels of the reader factor and 200 levels of the case factor. If a factor is regarded as fixed, then the conclusions of the analysis apply only to the specific levels of the factor used in the study. If a factor is regarded as random, the levels of the factor are regarded as random samples from a parent population of the corresponding factor, and conclusions regarding specific levels are not allowed; rather, conclusions apply to the distribution from which the levels were sampled.

ROC MRMC studies require a sample of cases and interpretations by one or more readers in one or more treatments (in this book the term *multiple* includes as a special case *one*). A study is never conducted on a sample of treatments. It would be nonsensical to image patients using a "sample" of all possible treatments. Every variation of an imaging technique (e.g., different kilovoltage or kVp) or display method (e.g., window-level setting) or image processing techniques qualifies as a distinct treatment. The number of possible treatments is very large, and, from a practical point of view, most of them are uninteresting. Rather, interest is in comparing two or more (a few at most) treatments that, based on preliminary studies, are clinically interesting. One treatment may be computed tomography, the other magnetic resonance imaging, or one may be interested in comparing a standard image processing method to a newly proposed one, or one may be interested in comparing CAD to a group of readers.

This brings out an essential difference between how cases, readers and treatments have to be regarded in the variability estimation procedure. Cases and readers are usually regarded as random factors (there has to be at least one random factor – if not, there are no sources of variability and nothing to apply statistics to!), while treatments are regarded as fixed factors. The random fac-

tors contribute variability, but the fixed factors do not, rather they contribute constant shifts in performance. The terms *fixed* and *random* factors are used in this specific sense, and are derived, in turn, from ANOVA methods in statistics. With two or more treatments, there are shifts in performance of treatments relative to each other, that one seeks to assess the significance of, against a background of noise contributed by the random factors. If the shifts are sufficiently large compared to the noise, then one can state, with some certainty, that they are real. Quantifying the last statement uses the methods of hypothesis testing introduced in Chapter 2.

3.4 Reader and case populations

Consider a sample of J readers. Conceptually there is a reader-population, modeled as a normal distribution $\theta_j \sim N\left(\theta_{\bullet\{1\}}, \sigma^2_{br+wr}\right)$, describing the variation of skill-level of readers. Here θ is a generic FOM. Each reader j is characterized by a different value of θ_j, $j = 1, 2, ...J$ and one can conceptually think of a bell-shaped curve with variance σ^2_{br+wr} describing between-reader variability of the readers. A large variance implies large spread in reader skill levels.

Likewise, there is a case-population, also modeled as a normal distribution, describing the variations in difficulty levels of the patients. One actually has two unit-variance distributions, one for non-diseased and one for diseased cases, characterized by a separation parameter. The separation parameter is scaled (i.e., normalized) by the standard deviation of each distribution (assumed equal). Each distribution has unit variance. Conceptually an easy case set has a larger than usual scaled separation parameter while a difficult case set has a smaller than usual scaled separation parameter. The distribution of the scaled separation parameter can be modeled as a bell-shaped curve $\theta_{\{c\}} \sim N\left(\theta_{\{\bullet\}}, \sigma^2_{cs+wr}\right)$ with variance σ^2_{cs+wr} describing the variations in difficulty levels of different case samples. Note the need for the case-set index, introduced in (book) Chapter 07, to specify the separation parameter for a specific case-set (in principle a j-index is also needed as one cannot have an interpretation without a reader; for now it is suppressed. A small variance σ^2_{cs} implies the different case sets have similar difficulty levels while a larger variance would imply a larger spread in difficulty levels. Just as the previous paragraph described reader-varibility, this paragraph has described case-variability.

Anytime one has a common random component to two measurements, the measurements are correlated. In the Wagner analogy, the common component is the random height, as a function of time, of a wave, which contributes the same amount to both height measurements (since the boats are adjacent). Since the readers interpret a common case set in all treatments one needs to account for various types of correlations that are potentially present. These occur due to the various types of pairings that can occur with MRMC data, where each pairing implies the presence of a common component to the measurements: (a)

the same reader interpreting the *same cases* in different treatments, (b) different readers interpreting the *same cases* in the same treatment and (c) different readers interpreting the *same cases* in different treatments. These pairings are more clearly elucidated in (book) Chapter 10. The current chapter uses jackknife pseudovalue based analysis to model the variances and the correlations. Hillis has shown that the two approaches are essentially equivalent (Hillis et al., 2008).

3.5 Three types of analyses

MRMC analysis aims to draw conclusions regarding the significances of inter-treatment shifts in performance. Ideally a conclusion (i.e., a difference is significant) should generalize to the respective populations from which the random samples were obtained. In other words, the idea is to generalize from the observed samples to the underlying populations. Three types of analyses are possible depending on which factor(s) one regards as random and which as fixed: random-reader random-case (RRRC), fixed-reader random-case (FRRC) and random-reader fixed-case (RRFC). If a factor is regarded as random, then the conclusion of the study applies to the population from which the levels of the factor were sampled. If a factor is regarded as fixed, then the conclusion applies only to the specific levels of the sampled factor. For example, if reader is regarded as a random factor, the conclusion generalizes to the reader population from which the readers used in the study were obtained. If reader is regarded as a fixed factor, then the conclusion applies to the specific readers that participated in the study. Regarding a factor as fixed effectively "freezes out" the sampling variability of the population and interest then centers only on the specific levels of the factor used in the study. Likewise, treating case as a fixed factor means the conclusion of the study is specific to the case-set used in the study.

3.6 General approach

This section provides an overview of the steps involved in analysis of MRMC data. Two approaches are described in parallel: a figure of merit (FOM) derived jackknife pseudovalue based approach, detailed in this chapter and an FOM based approach, detailed in the next chapter. The analysis proceeds as follows:

1. A FOM is selected: *the selection of FOM is the single-most critical aspect of analyzing an observer performance study.* The selected FOM is denoted θ. The FOM has to be an objective scalar measure of performance with larger values characterizing better performance. [The qualifier "larger" is trivially satisfied; if the figure of merit has the opposite characteristic, a sign change is all that is needed to bring it back to compliance with this

requirement.] Examples are empirical AUC, the binormal model-based estimate A_z , other advance method based estimates of AUC, sensitivity at a predefined value of specificity, etc. An example of a FOM requiring a sign-change is FPF at a specified TPF, where smaller values signify better performance.

2. For each treatment i and reader j the figure of merit θ_{ij} is estimated from the ratings data. Repeating this over all treatments and readers yields a matrix of observed values θ_{ij}. This is averaged over all readers in each treatment yielding $\theta_{i\bullet}$. The observed effect-size ES_{obs} is defined as the difference between the reader-averaged FOMs in the two treatments, i.e., $ES_{obs} = \theta_{2\bullet} - \theta_{1\bullet}$. While extensible to more than two treatments, the explanation is more transparent by restricting to two modalities.
3. If the magnitude of ES_{obs} is "large" one has reason to suspect that there might indeed be a significant difference in AUCs between the two treatments, where *significant* is used in the sense of (book) Chapter 08. Quantification of this statement, specifically how large is "large", requires the conceptually more complex steps described next.

    - In the DBM approach, the subject of this chapter, jackknife pseudovalues are calculated as described in Chapter 08. A standard ANOVA model with uncorrelated errors is used to model the pseudovalues.
    - In the OR approach, the subject of the next chapter, the FOM is modeled directly using a custom ANOVA model with correlated errors.

4. Depending on the selected method of modeling the data (pseudovalue vs. FOM) a statistical model is used which includes parameters modeling the true values in each treatment, and expected variations due to different variability components in the model, e.g., between-reader variability, case-sampling variability, interactions (e.g., allowing for the possibility that the random effect of a given reader could be treatment dependent) and the presence of correlations (between pseudovalues or FOMs) because of the pairings inherent in the interpretations.
5. In RRRC analysis one accounts for randomness in readers and cases. In FRRC analysis one regards reader as a fixed factor. In RRFC analysis one regards the case-sample (set of cases) as a fixed factor. The statistical model depends on the type of analysis.
6. The parameters of the statistical model are estimated from the observed data.
7. The estimates are used to infer the statistical distribution of the observed effect size, ES_{obs}, regarded as a realization of a random variable, under the null hypothesis (NH) that the true effect size is zero.
8. Based on this statistical distribution, and assuming a two-sided test, the probability (this is the oft-quoted p-value) of obtaining an effect size at least as extreme as that actually observed, is calculated, as in (book) Chapter 08.

9. If the p-value is smaller than a preselected value, denoted α, one declares the treatments different at the α - significance level. The quantity α is the control (or "cap") on the probability of making a Type I error, defined as rejecting the NH when it is true. It is common to set $\alpha = 0.05$ but depending on the severity of the consequences of a Type I error, as discussed in (book) Chapter 08, one might consider choosing a different value. Notice that α is a pre-selected number while the p-value is a realization (observation) of a random variable.
10. For a valid statistical analysis, the empirical probability α_{emp} over many (typically 2000) independent NH datasets, that the p-value is smaller than α, should equal α to within statistical uncertainty.

3.7 Summary TBA

This chapter has detailed analysis of MRMC ROC data using the DBM method. A reason for the level of detail is that almost all of the material carries over to other data collection paradigms, and a thorough understanding of the relatively simple ROC paradigm data is helpful to understanding the more complex ones.

DBM has been used in several hundred ROC studies (Prof. Kevin Berbaum, private communication ca. 2010). While the method allows generalization of a study finding, e.g., rejection of the NH, to the population of readers and cases, I believe this is sometimes taken too literally. If a study is done at a single hospital, then the radiologists tend to be more homogenous as compared to sampling radiologists from different hospitals. This is because close interactions between radiologists at a hospital tend to homogenize reading styles and performance. A similar issue applies to patient characteristics, which are also expected to vary more between different geographical locations than within a given location served by the hospital. This means is that single hospital study based p-values may tend to be biased downwards, declaring differences that may not be replicable if a wider sampling "net" were used using the same sample size. The price paid for a wider sampling net is that one must use more readers and cases to achieve the same sensitivity to genuine treatment effects, i.e., statistical power (i.e., there is no "free-lunch").

A third MRMC ROC method, due to Clarkson, Kupinski and Barrett19,20, implemented in open-source JAVA software by Gallas and colleagues22,44 (http://didsr.github.io/iMRMC/) is available on the web. Clarkson et al19,20 provide a probabilistic rationale for the DBM model, provided the figure of merit is the empirical AUC. The method is elegant but it is only applicable as long as one is using the empirical AUC as the figure of merit (FOM) for quantifying observer performance. In contrast the DBM approach outlined in this chapter, and the approach outlined in the following chapter, are applicable to any scalar FOM. Broader applicability ensures that significance-testing methods described in this, and the following chapter, apply to other ROC FOMs, such as

binormal model or other fitted AUCs, and more importantly, to other observer performance paradigms, such as free-response ROC paradigm. An advantage of the Clarkson et al. approach is that it predicts truth-state dependence of the variance components. One knows from modeling ROC data that diseased cases tend to have greater variance than non-diseased ones, and there is no reason to suspect that similar differences do not exist between the variance components.

Testing validity of an analysis method via simulation testing is only as good as the simulator used to generate the datasets, and this is where current research is at a bottleneck. The simulator plays a central role in ROC analysis. In my opinion this is not widely appreciated. In contrast, simulators are taken very seriously in other disciplines, such as cosmology, high-energy physics and weather forecasting. The simulator used to validate3 DBM is that proposed by Roe and Metz39 in 1997. This simulator has several shortcomings. (a) It assumes that the ratings are distributed like an equal-variance binormal model, which is not true for most clinical datasets (recall that the b-parameter of the binormal model is usually less than one). Work extending this simulator to unequal variance has been published3. (b) It does not take into account that some lesions are not visible, which is the basis of the contaminated binormal model (CBM). A CBM model based simulator would use equal variance distributions with the difference that the distribution for diseased cases would be a mixture distribution with two peaks. The radiological search model (RSM) of free-response data, Chapter 16 &17 also implies a mixture distribution for diseased cases, and it goes farther, as it predicts some cases yield no z-samples, which means they will always be rated in the lowest bin no matter how low the reporting threshold. Both CBM and RSM account for truth dependence by accounting for the underlying perceptual process. (c) The Roe-Metz simulator is out dated; the parameter values are based on datasets then available (prior to 1997). Medical imaging technology has changed substantially in the intervening decades. (d) Finally, the methodology used to arrive at the proposed parameter values is not clearly described. Needed is a more realistic simulator, incorporating knowledge from alternative ROC models and paradigms that is calibrated, by a clearly defined method, to current datasets.

Since ROC studies in medical imaging have serious health-care related consequences, no method should be used unless it has been thoroughly validated. Much work still remains to be done in proper simulator design, on which validation is dependent.

3.8 Chapter References

Chapter 4

Significance Testing using the DBM Method

4.1 TBA How much finished

60%

4.2 The DBM sampling model

DBM = Dorfman Berbaum Metz

The figure-of-merit has three indices:

- A treatment index i, where i runs from 1 to I, where I is the total number of treatments.
- A reader index j, where j runs from 1 to J, where J is the total number of readers.
- The case-sample index $\{c\}$, where $\{1\}$ i.e., $c = 1$, denotes a set of cases, K_1 non-diseased and K_2 diseased, interpreted by all readers in all treatments, and other integer values of c correspond to other independent sets of cases that, although not in fact interpreted by the readers, could potentially be "interpreted" using resampling methods such as the bootstrap or the jackknife.

The approach (Dorfman et al., 1992) taken by DBM was to use the jackknife resampling method to calculate FOM pseudovalues ${Y'}_{ijk}$ defined by (the reason for the prime will become clear shortly):

$$Y'_{ijk} = K\theta_{ij} - (K-1)\theta_{ij(k)} \tag{4.1}$$

Here θ_{ij} is the estimate of the figure-of-merit for reader j interpreting all cases in treatment i and $\theta_{ij(k)}$ is the corresponding figure of merit with case k *deleted* from the analysis. To keep the notation compact the case-sample index $\{1\}$ on every figure of merit symbol is suppressed.

Recall from book Chapter 07 that the jackknife is a way of teasing out the case-dependence: the left hand side of Equation (4.1) has a case index k, with k running from 1 to K, where K is the total number of cases: $K = K_1 + K_2$.

Hillis et al (Hillis et al., 2008) proposed a centering transformation on the pseudovalues (he terms it "normalized" pseudovalues, but to me "centering" is a more accurate and descriptive term - *Normalize: (In mathematics) multiply (a series, function, or item of data) by a factor that makes the norm or some associated quantity such as an integral equal to a desired value (usually 1). New Oxford American Dictionary, 2016*):

$$Y_{ijk} = Y'_{ijk} + \left(\theta_{ij} - Y'_{ij\bullet}\right) \tag{4.2}$$

Note: the bullet symbol denotes an average over the corresponding index.

The effect of this transformation is that the average of the centered pseudovalues over the case index is identical to the corresponding estimate of the figure of merit:

$$Y_{ij\bullet} = Y'_{ij\bullet} + \left(\theta_{ij} - Y'_{ij\bullet}\right) = \theta_{ij} \tag{4.3}$$

This has the advantage that all confidence intervals are properly centered. The transformation is unnecessary if one uses the Wilcoxon as the figure-of-merit, as the pseudovalues calculated using the Wilcoxon as the figure of merit are "naturally" centered, i.e.,

$\theta_{ij} - Y'_{ij\bullet} = 0$

It is understood that, unless explicitly stated otherwise, all calculations from now on will use centered pseudovalues.

Consider N replications of a MRMC study, where a replication means repetition of the study with the same treatments, readers and case-set $\{C = 1\}$. For N replications per treatment-reader-case combination, the DBM model for the pseudovalues is (n is the replication index, usually $n = 1$, but kept here for now):

$$Y_{n(ijk)} = \mu + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + (\tau RC)_{ijk} + \epsilon_{n(ijk)} \tag{4.4}$$

The term μ is a constant. By definition, the treatment effect τ_i is subject to the constraint:

$$\sum_{i=1}^{I} \tau_i = 0 \Rightarrow \tau_\bullet = 0 \tag{4.5}$$

This constraint ensures that μ has the interpretation of the average of the pseudovalues over treatments, readers and cases.

The (nesting) notation for the replication index, i.e., $n(ijk)$, implies n observations for treatment-reader-case combination ijk. With no replications ($N = 1$) it is convenient to omit the n-symbol.

The parameter τ_i is estimated as follows:

$$Y_{ijk} \equiv Y_{1(ijk)}\tau_i = Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet} \tag{4.6}$$

The basic assumption of the DBM model is that the pseudovalues can be regarded as independent and identically distributed observations. That being the case, the pseudovalues can be analyzed by standard ANOVA techniques. Since pseduovalues are computed from a common dataset, this assumption is, non-intuitive. However, for the special case of Wilcoxon figure of merit, it is justified.

4.2.1 Explanation of terms in the model

The right hand side of Eqn. (4.1) consists of one fixed and 7 random effects. The current analysis assumes readers and cases as random factors (RRRC), so by definition R_j and C_k are random effects, and moreover, any term that includes a random factor is a random effect; for example, $(\tau R)_{ij}$ is a random effect because it includes the R factor. Here is a list of the random terms:

$$R_j, C_k, (\tau R)_{ij}, (\tau C)_{ik}, (RC)_{jk}, (\tau RC)_{ijk}, \epsilon_{ijk} \tag{4.7}$$

Assumption: Each of the random effects is modeled as a random sample from mutually independent zero-mean normal distributions with variances as specified below:

$$\left.\begin{array}{rcl} R_j & \sim & N\left(0,\sigma_R^2\right) \\ C_k & \sim & N\left(0,\sigma_C^2\right) \\ (\tau R)_{ij} & \sim & N\left(0,\sigma_{\tau R}^2\right) \\ (\tau C)_{ik} & \sim & N\left(0,\sigma_{\tau C}^2\right) \\ (RC)_{jk} & \sim & N\left(0,\sigma_{RC}^2\right) \\ (\tau RC)_{ijk} & \sim & N\left(0,\sigma_{\tau RC}^2\right) \\ \epsilon_{ijk} & \sim & N\left(0,\sigma_\epsilon^2\right) \end{array}\right\} \tag{4.8}$$

Equation (4.8) defines the meanings of the variance components appearing in Equation (4.7). One could have placed a Y subscript (or superscript) on each of the variances, as they describe fluctuations of the pseudovalues, not FOM values. However, this tends to clutter the notation. So here is the convention:

Unless explicitly stated otherwise, all variance symbols in this chapter refer to pseudovalues. Another convention: $(\tau R)_{ij}$ is *not* the product of the treatment and reader factors, rather it is a single factor, namely the treatment-reader factor with IJ levels, subscripted by the index ij and similarly for the other product-like terms in Equation (4.8).

4.2.2 Meanings of variance components in the DBM model (TBA this section can be improved)

The variances defined in (4.8) are collectively termed *variance components*. Specifically, they are jackknife pseudovalue variance components, to be distinguished from figure of merit (FOM) variance components to be introduced in TBA Chapter 10. They are in order: $\sigma_R^2, \sigma_C^2 \sigma_{\tau R}^2, \sigma_{\tau C}^2, \sigma_{RC}^2, \sigma_{\tau RC}^2, \sigma_\epsilon^2$. They have the following meanings.

- The term σ_R^2 is the variance of readers that is independent of treatment or case, which are modeled separately. It is not to be confused with the terms σ_{br+wr}^2 and σ_{cs+wr}^2 used in §9.3, which describe the variability of θ measured under specified conditions. [A jackknife pseudovalue is a weighted difference of FOM like quantities, TBA (4.1). Its meaning will be explored later. For now, *a pseudovalue variance is distinct from a FOM variance*.]
- The term σ_C^2 is the variance of cases that is independent of treatment or reader.
- The term $\sigma_{\tau R}^2$ is the treatment-dependent variance of readers that was excluded in the definition of σ_R^2. If one were to sample readers and treatments for the same case-set, the net variance would be $\sigma_R^2 + \sigma_{\tau R}^2 + \sigma_\epsilon^2$.

- The term $\sigma^2_{\tau C}$ is the treatment-dependent variance of cases that was excluded in the definition of σ^2_C. So, if one were to sample cases and treatments for the same readers, the net variance would be $\sigma^2_C + \sigma^2_{\tau C} + \sigma^2_\epsilon$.
- The term σ^2_{RC} is the treatment-independent variance of readers and cases that were excluded in the definitions of σ^2_R and σ^2_C. So, if one were to sample readers and cases for the same treatment, the net variance would be $\sigma^2_R + \sigma^2_C + \sigma^2_{RC} + \sigma^2_\epsilon$.
- The term $\sigma^2_{\tau RC}$ is the variance of treatments, readers and cases that were excluded in the definitions of all the preceding terms in TBA (4.1). So, if one were to sample treatments, readers and cases the net variance would be $\sigma^2_R + \sigma^2_C + \sigma^2_{\tau C} + \sigma^2_{RC} + \sigma^2_{\tau RC} + \sigma^2_\epsilon$.
- The last term, σ^2_ϵ describes the variance arising from different replications of the study using the same treatments, readers and cases. Measuring this variance requires repeating the study several (N) times with the same treatments, readers and cases, and computing the variance of $Y_{n(ijk)}$, where the additional n-index refers to true replications, $n = 1, 2, \ldots, N$.

$$\sigma^2_\epsilon = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{k}\frac{1}{N-1}\sum_{n=1}^{N}\left(Y_{n(ijk)} - Y_{\bullet(ijk)}\right)^2 \tag{4.9}$$

The right hand side of TBA (4.1) is the variance of $Y_{n(ijk)}$, for specific ijk, with respect to the replication index n, averaged over all ijk. In practice $N = 1$ (i.e., there are no replications) and this variance cannot be estimated (it would imply dividing by zero). It has the meaning of *reader inconsistency*, usually termed *within-reader* variability. As will be shown later, the presence of this inestimable term does not limit ones ability to perform significance testing on the treatment effect without having to replicate the whole study, as implied in earlier work (Obuchowski and Rockette, 1995).

An equation like TBA (4.1) is termed a *linear model* with the left hand side, the pseudovalue "observations", modeled by a sum of fixed and random terms. Specifically it is a *mixed model*, because the right hand side has both fixed and random effects. Statistical methods have been developed for analysis of such linear models. One estimates the terms on the right hand side of TBA (4.1), it being understood that for the random effects, one estimates the variances of the zero-mean normal distributions, TBA (4.1)Eqn. (9.7), from which the samples are obtained (by assumption).

Estimating the fixed effects is trivial. The term μ is estimated by averaging the left hand side of TBA (4.1)Eqn. (9.4) over all three indices (since $N = 1$): $\mu = Y_{\bullet\bullet\bullet}$

Because of the way the treatment effect is defined, TBA (4.1) Eqn. (9.5), averaging, which involves summing, over the treatment-index i, yields zero, and all of the remaining random terms yield zero upon averaging, because they are

individually sampled from zero-mean normal distributions. To estimate the treatment effect one takes the difference $\tau_i = Y_{i\bullet\bullet} - \mu$.

It can be easily seen that the reader and case averaged difference between two different treatments i and i' is estimated by $\tau_i - \tau_{i'} = Y_{i\bullet\bullet} - Y_{i'\bullet\bullet}$.

Estimating the strengths of the random terms is a little more complicated. It involves methods adapted from least squares, or maximum likelihood, and more esoteric ways. I do not feel comfortable going into these methods. Instead, results are presented and arguments are made to make them plausible. The starting point is definitions of quantities called **mean squares** and their expected values.

4.2.3 Definitions of mean-squares

Again, to be clear, one chould put a Y subscript (or superscript) on each of the following definitions, but that would make the notation unnecessarily cumbersome.

In this chapter, all mean-square quantities are calculated using pseudovalues, not figure-of-merit values. The presence of three subscripts on Y should make this clear. Also the replication index and the nesting notation are suppressed. The notation is abbreviated so MST is the mean square corresponding to the treatment effect, etc.

The definitions of the mean-squares below match those (where provided) in (Hillis and Berbaum, 2004, page 1261).

$$\left.\begin{aligned}
\text{MST} &= \frac{JK\sum_{i=1}^{I}(Y_{i\bullet\bullet}-Y_{\bullet\bullet\bullet})^2}{I-1} \\
\text{MSR} &= \frac{IK\sum_{j=1}^{J}(Y_{\bullet j\bullet}-Y_{\bullet\bullet\bullet})^2}{J-1} \\
\text{MS(C)} &= \frac{IJ\sum_{k=1}^{K}(Y_{\bullet\bullet k}-Y_{\bullet\bullet\bullet})^2}{K-1} \\
\text{MSTR} &= \frac{K\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij\bullet}-Y_{i\bullet\bullet}-Y_{\bullet j\bullet}+Y_{\bullet\bullet\bullet})^2}{(I-1)(J-1)} \\
\text{MSTC} &= \frac{J\sum_{i=1}^{I}\sum_{k=1}^{K}(Y_{i\bullet k}-Y_{i\bullet\bullet}-Y_{\bullet\bullet k}+Y_{\bullet\bullet\bullet})^2}{(I-1)(K-1)} \\
\text{MSRC} &= \frac{I\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{\bullet jk}-Y_{\bullet j\bullet}-Y_{\bullet\bullet k}+Y_{\bullet\bullet\bullet})^2}{(J-1)(K-1)} \\
\text{MSTRC} &= \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk}-Y_{ij\bullet}-Y_{i\bullet k}-Y_{\bullet jk}+Y_{i\bullet\bullet}+Y_{\bullet j\bullet}+Y_{\bullet\bullet k}-Y_{\bullet\bullet\bullet})^2}{(I-1)(J-1)K-1)}
\end{aligned}\right\} \quad (4.10)$$

Note the absence of MSE, corresponding to the ϵ term on the right hand side of (4.10). With only one observation per treatment-reader-case combination, MSE cannot be estimated; it effectively gets absorbed into the $MSTRC$ term.

4.3 Expected values of mean squares

> "In our original formulation [2], expected mean squares for the ANOVA were derived from a restricted parameterization in which mixed-factor interactions sum to zero over indexes of fixed effects. In the restricted parameterization, the mixed effects are correlated, parameters are sometimes awkward to define [17], and extension to unbalanced designs is dubious [17, 18]. In this article, we recommend the unrestricted parameterization. The restricted and unrestricted parameterizations are special cases of a general model by Scheffe [19] that allows an arbitrary covariance structure among experimental units within a level of a random factor. Tables 1 and 2 show the ANOVA tables with expected mean squares for the unrestricted formulation."
>
> — (Dorfman et al., 1995)

The *observed* mean squares defined in Equation (4.10) can be calculated directly from the *observed* pseudovalues. The next step in the analysis is to obtain expressions for their *expected* values in terms of the variances defined in (4.10). Assuming no replications, i.e., $N = 1$, the expected mean squares are as follows, Table Table 4.1; understanding how this table is derived, would lead me outside my expertise and the scope of this book; suffice to say that these are *unconstrained* estimates (as summarized in the quotation above) which are different from the *constrained* estimates appearing in the original DBM publication (Dorfman et al., 1992).

Table 4.1: Unconstrained expected values of mean-squares, as in (Dorfman et al., 1995)

Source	df	E(MS)
T	(I-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + K\sigma^2_{\tau R} + J\sigma^2_{\tau C} + JK\sigma^2_\tau$
R	(J-1)	$\sigma^2_\epsilon + I\sigma^2_{RC} + IK\sigma^2_R + K\sigma^2_{\tau R}$
C	(K-1)	$\sigma^2_\epsilon + I\sigma^2_{RC} + IJ\sigma^2_C + J\sigma^2_{\tau C}$
TR	(I-1)(J-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + K\sigma^2_{\tau R}$
TC	(I-1)(K-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + J\sigma^2_{\tau C}$
RC	(J-1)(K-1)	$\sigma^2_\epsilon + I\sigma^2_{RC}$
TRC	(I-1)(J-1)(K-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC}$
ϵ	$N - 1 = 0$	σ^2_ϵ

- In Table 4.1 the following notation is used as a shorthand:

$$\sigma^2_\tau = \frac{1}{I-1} \sum_{i=1}^{I} \left(Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet}\right)^2 \tag{4.11}$$

Since treatment is a fixed effect, the variance symbol σ_τ^2, which is used for notational consistency in Table 4.1, could cause confusion. The right hand side "looks like" a variance, indeed one that could be calculated for just two treatments but, of course, random sampling from a *distribution of treatments* is not the intent of the notation.

4.4 Random-reader random-case (RRRC) analysis

Both readers and cases are regarded as random factors. The expected mean squares in Table Table 4.1 are variance-like quantities; specifically, they are weighted linear combinations of the variances appearing in (4.8). For single factors the column headed "degrees of freedom" (df) is one less than the number of levels of the corresponding factor; estimating a variance requires first estimating the mean, which imposes a constraint, thereby decreasing df by one. For interaction terms, df is the product of the degrees of freedom for the individual factors. As an example, the term $(\tau RC)_{ijk}$ contains three individual factors, and therefore $df = (I-1)(J-1)(K-1)$. The number of degrees of freedom can be thought of as the amount of information available in estimating a mean square. As a special case, with no replications, the ϵ term has zero df as $N - 1 = 0$. With only one observation $Y_{1(ijk)}$ there is no information to estimate the variance corresponding to the ϵ term. To estimate this term one needs to replicate the study several times – each time the same readers interpret the same cases in all treatments – a very boring task for the reader and totally unnecessary from the researcher's point of view.

4.4.1 Calculation of mean squares: an example

- We choose `dataset02` to illustrate calculation of mean squares for pseudovalues. This is referred to in the book as the "VD" dataset (Van Dyke et al., 1993). It consists of 114 cases, 45 of which are diseased, interpreted in two treatments by five radiologists using the ROC paradigm.
- The first line computes the pseudovalues using the `RJafroc` function `UtilPseudoValues()`, and the second line extracts the numbers of treatments, readers and cases. The following lines calculate, using Equation (4.10) the mean-squares. After displaying the results of the calculation, the results are compared to those calculated by the `RJafroc` function `UtilMeanSquares()`.

```
Y <- UtilPseudoValues(dataset02, FOM = "Wilcoxon")$jkPseudoValues

I <- dim(Y)[1];J <- dim(Y)[2];K <- dim(Y)[3]
```

```
msT <- 0
for (i in 1:I)  {
  msT <- msT + (mean(Y[i, , ]) - mean(Y))^2
}
msT <- msT * J * K/(I - 1)

msR <- 0
for (j in 1:J) {
  msR <- msR + (mean(Y[, j, ]) - mean(Y))^2
}
msR <- msR * I * K/(J - 1)

msC <- 0
for (k in 1:K) {
  msC <- msC + (mean(Y[, , k]) - mean(Y))^2
}
msC <- msC * I * J/(K - 1)

msTR <- 0
for (i in 1:I) {
  for (j in 1:J) {
    msTR <- msTR +
      (mean(Y[i, j, ]) - mean(Y[i, , ]) - mean(Y[, j, ]) + mean(Y))^2
  }
}
msTR <- msTR * K/((I - 1) * (J - 1))

msTC <- 0
for (i in 1:I) {
  for (k in 1:K) {
    msTC <- msTC +
      (mean(Y[i, , k]) - mean(Y[i, , ]) - mean(Y[, , k]) + mean(Y))^2
  }
  msTC <- msTC * J/((I - 1) * (K - 1))
}

msTC <- 0
for (i in 1:I) {
  for (k in 1:K) { # OK
    msTC <- msTC +
      (mean(Y[i, , k]) - mean(Y[i, , ]) - mean(Y[, , k]) + mean(Y))^2
  }
}
msTC <- msTC * J/((I - 1) * (K - 1))
```

```
msRC <- 0
for (j in 1:J) {
  for (k in 1:K) {
    msRC <- msRC +
      (mean(Y[, j, k]) - mean(Y[, j, ]) - mean(Y[, , k]) + mean(Y))^2
  }
}
msRC <- msRC * I/((J - 1) * (K - 1))

msTRC <- 0
for (i in 1:I) {
  for (j in 1:J) {
    for (k in 1:K) {
      msTRC <- msTRC + (Y[i, j, k] - mean(Y[i, j, ]) -
                          mean(Y[i, , k]) - mean(Y[, j, k]) +
                          mean(Y[i, , ]) + mean(Y[, j, ]) +
                          mean(Y[, , k]) - mean(Y))^2
    }
  }
}
msTRC <- msTRC/((I - 1) * (J - 1) * (K - 1))

data.frame("msT" = msT, "msR" = msR, "msC" = msC,
           "msTR" = msTR, "msTC" = msTC,
           "msRC" = msRC, "msTRC" = msTRC)
#>         msT       msR       msC       msTR       msTC       msRC     msTRC
#> 1 0.5467634 0.4373268 0.3968699 0.06281749 0.09984808 0.06450106 0.0399716

as.data.frame(UtilMeanSquares(dataset02)[1:7])
#>         msT       msR       msC       msTR       msTC       msRC     msTRC
#> 1 0.5467634 0.4373268 0.3968699 0.06281749 0.09984808 0.06450106 0.0399716
```

4.4.2 Significance testing

If the NH of no treatment effect is true, i.e., if $\sigma_\tau^2 = 0$, then according to Table 4.1 the following holds (the last term in the row labeled T in Table 4.1 drops out):

$$E\left(MST \mid NH\right) = \sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2 + J\sigma_{\tau C}^2 \tag{4.12}$$

Also, the following linear combination is equal to $E\left(MST \mid NH\right)$:

$$\begin{aligned} & E(MSTR) + E(MSTC) - E(MSTRC) \\ & = (\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2) + (\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2) - (\sigma_\epsilon^2 + \sigma_{\tau RC}^2) \\ & = \sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2 + K\sigma_{\tau R}^2 \\ & = E(MST \mid NH) \end{aligned} \tag{4.13}$$

Therefore, under the NH, the ratio:

$$\frac{E(MST \mid NH)}{E(MSTR) + E(MSTC) - E(MSTRC)} = 1 \tag{4.14}$$

In practice, one does not know the expected values – that would require averaging each of these quantities, regarded as random variables, over their respective distributions. Therefore, one defines the following statistic, denoted F_{DBM}, using the observed values of the mean squares, calculated almost trivially as in the previous example, using their definitions in Equation (4.10):

$$F_{DBM} = \frac{MST}{MSTR + MSTC - MSTRC} \tag{4.15}$$

F_{DBM} is a realization of a random variable. A non-zero treatment effect, i.e., $\sigma_\tau^2 > 0$, will cause the ratio to be larger than one, because $E(MST)$ will be larger, see row labeled T in Table 4.1. Therefore values of $F_{DBM} > 1$ will tend to reject the NH. Drawing on a theorem from statistics (Larsen and Marx, 2005), under the NH the ratio of two independent mean squares is distributed as a (central) F-statistic with degrees of freedom corresponding to those of the mean squares forming the numerator and denominator of the ratio (Theorem 12.2.5 in "An Introduction to Mathematical Statistics and Its Applications"). To perform hypothesis testing one needs the distribution, under the NH, of the statistic defined by Eqn. (4.15). This is completely analogous to Chapter 08 where knowledge of the distribution of AUC under the NH enabled testing the null hypothesis that the observed value of AUC equals a pre-specified value.

Under the NH, $F_{DBM|NH}$ is distributed according to the F-distribution characterized by two numbers:

- A numerator degrees of freedom (ndf) – determined by the degrees of freedom of the numerator, MST, of the ratio comprising the F-statistic, i.e., I–1, and
- A denominator degrees of freedom (ddf) - determined by the degrees of freedom of the denominator, $MSTR + MSTC - MSTRC$, of the ratio comprising the F-statistic, to be described in the next section.

Summarizing,

$$\left.\begin{array}{r} F_{DBM|NH} \sim F_{\text{ndf,ddf}} \\ \text{ndf} = I - 1 \end{array}\right\} \tag{4.16}$$

The next topic is estimating ddf.

4.4.3 The Satterthwaite approximation

The denominator of the F-ratio is $MSTR + MSTC - MSTRC$. This is not a *simple* mean square (I am using terminology in the Satterthwaite papers - he means any mean square defined by equations such as in Equation (4.10)). Rather it is a *linear combination of mean squares* (with coefficients 1, 1 and -1), and the resulting value could even be negative leading to a negative $F_{DBM|NH}$, which is an illegal value for a sample from an F-distribution (a ratio of two variances). In 1941 Satterthwaite (Satterthwaite, 1941, 1946) proposed an approximate degree of freedom for a linear combination of simple mean square quantities. TBA Online Appendix 9.A explains the approximation in more detail. The end result is that the mean square quantity described in Equation (4.15) has an approximate degree of freedom defined by (this is called the *Satterthwaite's approximation*):

$$ddf_{Sat} = \frac{\left(MSTR + MSTC - MSTRC\right)^2}{\left(\frac{MSTR^2}{(I-1)(J-1)} + \frac{MSTC^2}{(I-1)(K-1)} + \frac{MSTRC^2}{(I-1)(J-1)(K-1)}\right)} \tag{4.17}$$

The subscript Sat is for Satterthwaite. From Equation (4.17) it should be fairly obvious that in general ddf_{Sat} is not an integer. To accommodate possible negative estimates of the denominator of Equation (4.17), the original DBM method (Dorfman et al., 1992) proposed, depending on the signs of $\sigma^2_{\tau R}$ and $\sigma^2_{\tau C}$, four expressions for the F-statistic and corresponding expressions for ddf. Rather than repeat them here, since they have been superseded by the method described below, the interested reader is referred to Eqn. 6 and Eqn. 7 in Reference (Hillis et al., 2008).

Instead Hillis (Hillis, 2007) proposed the following statistic for testing the null hypothesis:

$$F_{DBM} = \frac{MST}{MSTR + \max\left(MSTC - MSTRC, 0\right)} \tag{4.18}$$

Now the denominator cannot be negative. One can think of the F-statistic F_{DBM} as a signal-to-noise ratio like quantity, with the difference that both numerator and denominator are variance like quantities. If the "variance" represented by the treatment effect is larger than the variance of the noise tending to mask the treatment effect, then F_{DBM} tends to be large, which makes the observed treatment "variance" stand out more clearly compared to the noise,

and the NH is more likely to be rejected. Hillis in (Hillis et al., 2005) has shown that the left hand side of Equation (4.18) is distributed as an F-statistic with $\text{ndf} = I - 1$ and denominator degrees of freedom ddf_H defined by:

$$ddf_H = \frac{(MSTR + \max(MSTC - MSTRC, 0))^2}{\text{MSTR}^2}(I-1)(J-1) \tag{4.19}$$

Summarizing,

$$F_{DBM} \sim F_{\text{ndf},\text{ddf}_H} \text{ndf} = I - 1 \tag{4.20}$$

Instead of 4 rules, as in the original DBM method, the Hillis modification involves just one rule, summarized by Equations (4.19) through (4.20). Moreover, the F-statistic is constrained to non-negative values. Using simulation testing (Hillis et al., 2008) he has been shown that the modified DBM method has better null hypothesis behavior than the original DBM method. The latter tended to be too conservative, typically yielding Type I error rates smaller than the expected 5% for $\alpha = 0.05$.

4.4.4 Decision rules, p-value and confidence intervals

The *critical* value of the F-distribution, denoted $F_{1-\alpha,\text{ndf},\text{ddf}_H}$, is defined such that fraction $1 - \alpha$ of the distribution lies to the left of the critical value, in other words it is the $1 - \alpha$ *quantile* of the F-distribution:

$$\Pr\left(F \le F_{1-\alpha,\text{ndf},\text{ddf}_H} \mid F \sim F_{\text{ndf},\text{ddf}_H}\right) = 1 - \alpha \tag{4.21}$$

The critical value $F_{1-\alpha,\text{ndf},\text{ddf}_H}$ increases as α decreases. The value of α, generally chosen to be 0.05, termed the *nominal* α, is fixed. The decision rule is that if $F_{DBM} > F_{1-\alpha,\text{ndf},\text{ddf}_H}$ one rejects the NH and otherwise one does not. It follows, from the definition of F_{DBM}, Equation (4.18), that rejection of the NH is more likely to occur if:

- F_{DBM} is large, which occurs if MST is large, meaning the treatment effect is large
- $MSTR + \max(MSTC - MSTRC, 0)$ is small, see comments following TBA (4.1) Eqn. (9.23).
- α is large: for then $F_{1-\alpha,\text{ndf},\text{ddf}_H}$ decreases and is more likely to be exceeded by the observed value of F_{DBM}.
- ndf is large: the more the number of treatment pairings, the greater the chance that at least one pairing will reject the NH. This is one reason sample size calculations are rarely conducted for more than 2-treatments.
- ddf_H is large: this causes the critical value to decrease, see below, and is more likely to be exceeded by the observed value of F_{DBM}.

4.4.4.1 p-value of the F-test

The p-value of the test is the probability, under the NH, that an equal or larger value of the F-statistic than observed F_{DBM} could occur by chance. In other words, it is the area under the (central) F-distribution $F_{\text{ndf,ddf}}$ that lies to the right of the observed value of F_{DBM}:

$$p = \Pr\left(F > F_{DBM} \mid F \sim F_{\text{ndf,ddf}_H}\right) \tag{4.22}$$

4.4.4.2 Confidence intervals for inter-treatment FOM differences

If $p < \alpha$ then the NH that all treatments are identical is rejected at significance level α. That informs the researcher that there exists at least one treatment-pair that has a difference significantly different from zero. To identify which pair(s) are different, one calculates confidence intervals for each paired difference. Hillis in (Hillis et al., 2005) has shown that the $(1-\alpha)$ confidence interval for $Y_{i\bullet\bullet} - Y_{i'\bullet\bullet}$ is given by:

$$CI_{1-\alpha} = (Y_{i\bullet\bullet} - Y_{i'\bullet\bullet}) \pm t_{\alpha/2;\text{ddf}_H} \sqrt{\frac{2}{JK}\left(MSTR + \max\left(MSTC - MSTRC, 0\right)\right)} \tag{4.23}$$

Here $t_{\alpha/2;\text{ddf}_H}$ is that value such that $\alpha/2$ of the *central t-distribution* with ddf_H degrees of freedom is contained in the upper tail of the distribution:

$$\Pr\left(T > t_{\alpha/2;\text{ddf}_H}\right) = \alpha/2 \tag{4.24}$$

Since centered pseudovalues were used:

$$(Y_{i\bullet\bullet} - Y_{i'\bullet\bullet}) = (\theta_{i\bullet} - \theta_{i'\bullet}) \tag{4.25}$$

Therefore, Equation (4.23) can be rewritten:

$$CI_{1-\alpha} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2;\text{ddf}_H} \sqrt{\frac{2}{JK}\left(MSTR + \max\left(MSTC - MSTRC, 0\right)\right)} \tag{4.26}$$

For two treatments any of the following equivalent rules could be adopted to reject the NH:

- $F_{DBM} > F_{1-\alpha,\text{ndf,ddf}_H}$
- $p < \alpha$

- $CI_{1-alpha}$ excludes zero

For more than two treatments the first two rules are equivalent and if a significant difference is found using either of them, then one can use the confidence intervals to determine which treatment pair differences are significantly different from zero. The first F-test is called the *overall F-test* and the subsequent tests the *treatment-pair t-tests*. One only conducts treatment pair t-tests if the overall F-test yields a significant result.

4.4.4.3 Code illustrating the F-statistic, ddf and p-value for RRRC analysis, Van Dyke data

Line 1 defines α. Line 2 forms a data frame from previously calculated mean-squares. Line 3 calculates the denominator appearing in Equation (4.18). Line 4 computes the observed value of F_{DBM}, namely the ratio of the numerator and denominator in Equation (4.18). Line 5 sets ndf to $I-1$. Line 6 computes ddf_H. Line 7 computes the critical value of the F-distribution $F_{crit} \equiv F_{\text{ndf},\text{ddf}_H}$. Line 8 calculates the p-value, using the definition Equation (4.22). Line 9 prints out the just calculated quantities. The next line uses the `RJafroc` function `StSignificanceTesting()` and the 2nd last line prints out corresponding `RJafroc`-computed quantities. Note the correspondences between the values just computed and those provide by `RJafroc`. Note that the FOM difference is not significant at the 5% level of significance as $p > \alpha$. The last line shows that F_{DBM} does not exceed F_{crit}. The two rules are equivalent.

```
alpha <- 0.05
retMS <- data.frame("msT" = msT, "msR" = msR, "msC" = msC,
                    "msTR" = msTR, "msTC" = msTC,
                    "msRC" = msRC, "msTRC" = msTRC)
F_DBM_den <- retMS$msTR+max(retMS$msTC - retMS$msTRC,0)
F_DBM <- retMS$msT / F_DBM_den
ndf <- (I-1)
ddf_H <- (F_DBM_den^2/retMS$msTR^2)*(I-1)*(J-1)
FCrit <- qf(1 - alpha, ndf, ddf_H)
pValueH <- 1 - pf(F_DBM, ndf, ddf_H)
data.frame("F_DBM" = F_DBM, "ddf_H"= ddf_H, "pValueH" = pValueH) # Line 9
#>      F_DBM    ddf_H    pValueH
#> 1 4.456319 15.25967 0.05166569
retRJafroc <- StSignificanceTesting(dataset02,
                                    FOM = "Wilcoxon",
                                    method = "DBM")
data.frame("F_DBM" = retRJafroc$RRRC$FTests$FStat[1],
           "ddf_H"= retRJafroc$RRRC$FTests$DF[2],
           "pValueH" = retRJafroc$RRRC$FTests$p[1])
#>       F_DBM     ddf_H     pValueH
```

```
#> 1 4.4563187 15.259675 0.051665686
F_DBM > FCrit
#> [1] FALSE
```

4.4.4.4 Code illustrating the inter-treatment confidence interval for RRRC analysis, Van Dyke data

Line 1 computes the FOM matrix using function `UtilFigureOfMerit`. The next 9 lines compute the treatment FOM differences. The next line `nDiffs` (for "number of differences") evaluates to 1, as with two treatments, there is only one difference. The next line initializes `CI_DIFF_FOM_RRRC`, which stands for "confidence intervals, FOM differences, for RRRC analysis". The next 8 lines evaluate, using Equation (4.26), and prints the lower value, the mid-point and the upper value of the confidence interval. Finally, these values are compared to those yielded by `RJafroc`. The FOM difference is not significant, whether viewed from the point of view of the F-statistic not exceeding the critical value, the observed p-value being larger than alpha or the 95% CI for the FOM difference including zero.

```
theta <- as.matrix(UtilFigureOfMerit(dataset02, FOM = "Wilcoxon"))
theta_i_dot <- array(dim = I)
for (i in 1:I) theta_i_dot[i] <- mean(theta[i,])
trtDiff <- array(dim = c(I,I))
for (i1 in 1:(I-1)) {
  for (i2 in (i1+1):I) {
    trtDiff[i1,i2] <- theta_i_dot[i1]- theta_i_dot[i2]
  }
}
trtDiff <- trtDiff[!is.na(trtDiff)]
nDiffs <- I*(I-1)/2
CI_DIFF_FOM_RRRC <- array(dim = c(nDiffs, 3))
for (i in 1 : nDiffs) {
  CI_DIFF_FOM_RRRC[i,1] <- qt(alpha/2,df = ddf_H)*sqrt(2*F_DBM_den/J/K) + trtDiff[i]
  CI_DIFF_FOM_RRRC[i,2] <- trtDiff[i]
  CI_DIFF_FOM_RRRC[i,3] <- qt(1-alpha/2,df = ddf_H)*sqrt(2*F_DBM_den/J/K) + trtDiff[i]
  print(data.frame("Lower" = CI_DIFF_FOM_RRRC[i,1],
                   "Mid" = CI_DIFF_FOM_RRRC[i,2],
                   "Upper" = CI_DIFF_FOM_RRRC[i,3]))
}
#>          Lower          Mid         Upper
#> 1 -0.087959499 -0.043800322 0.00035885444
data.frame("Lower" = retRJafroc$RRRC$ciDiffTrt[1,"CILower"],
           "Mid" = retRJafroc$RRRC$ciDiffTrt[1,"Estimate"],
           "Upper" = retRJafroc$RRRC$ciDiffTrt[1,"CIUpper"])
```

```
#>          Lower          Mid          Upper
#> 1 -0.087959499 -0.043800322 0.00035885444
```

4.5 Sample size estimation for random-reader random-case generalization

4.5.1 The non-centrality parameter

In the significance-testing procedure just described, the relevant distribution was that of the F-statistic when the NH is true, Equation (4.20). *For sample size estimation, one needs to know the distribution of the statistic when the NH is false.* In the latter condition (i.e., the AH) the observed F-statistic, defined by Equation (4.15), is distributed as a *non-central* F-distribution $F_{\text{ndf},\text{ddf}_H,\Delta}$ with *non-centrality parameter* Δ:

$$F_{DBM|AH} \sim F_{\text{ndf},ddf_H,\Delta} \tag{4.27}$$

The non-centrality parameter Δ is defined, compare (Hillis and Berbaum, 2004) Eqn. 6, by:

$$\Delta = \frac{JK\sigma_\tau^2}{\sigma_\epsilon^2 + K\sigma_{\tau R}^2 + J\sigma_{\tau C}^2}$$

The parameters σ_τ^2, $\sigma_{\tau R}^2$ and $\sigma_{\tau C}^2$ appearing in this equation are identical to three of the six variances describing the DBM model, Equation (4.4). The estimates of $\sigma_{\tau R}^2$ and/or $\sigma_{\tau C}^2$ can turn out to be negative (if either of these parameters is close to zero, an estimate from a small pilot study can be negative). To avoid a possibly negative denominator, (Hillis and Berbaum, 2004) suggest the following modifications (see sentence following Eqn. 4 in cited paper):

$$\Delta = \frac{JK\sigma_\tau^2}{\sigma_\epsilon^2 + \max(K\sigma_{\tau R}^2, 0) + \max(J\sigma_{\tau C}^2, 0)} \tag{4.28}$$

The observed effect size d, a realization of a random variable, is defined by (the bullet represents an average over the reader index):

$$d = \theta_{1\bullet} - \theta_{2\bullet} \tag{4.29}$$

For two treatments, since the individual treatment effects must be the negatives of each other (because they sum to zero, see (4.5)), it follows that:

$$\sigma_\tau^2 = \frac{d^2}{2} \tag{4.30}$$

Therefore, for two treatments the numerator of the expression for Δ is $JKd^2/2$. Dividing numerator and denominator of Equation (4.28) by K, one gets the final expression for Δ, as coded in `RJafroc`, namely:

$$\Delta = \frac{Jd^2/2}{\max(\sigma_{\tau R}^2, 0) + (\sigma_\epsilon^2 + \max(J\sigma_{\tau C}^2, 0))/K} \tag{4.31}$$

The variances, σ_τ^2, $\sigma_{\tau R}^2$ and $\sigma_{\tau C}^2$, appearing in Equation (4.31), can be calculated from the observed mean squares using the following equations, see (Hillis and Berbaum, 2004) Eqn. 4,

$$\left.\begin{aligned} \sigma_\epsilon^2 &= \text{MSTRC}^* \\ \sigma_{\tau R}^2 &= \frac{\text{MSTR}^* - \text{MSTRC}^*}{K^*} \\ \sigma_{\tau C}^2 &= \frac{\text{MSTC}^* - \text{MSTRC}^*}{J^*} \end{aligned}\right\} \tag{4.32}$$

- Here the asterisk is used to (consistently) denote quantities, including the mean squares, pertaining to the *pilot* study.
- In particular, J^* and K^* denote the numbers of readers and cases, respectively, *in the pilot study*, while J and K, appearing elsewhere, for example in Equation (4.31), are the corresponding numbers for the *planned or pivotal study*.
- The three variances, determined from the pilot study via Equation (4.32), are assumed to apply unchanged to the pivotal study (as they are sample-size independent parameters of the DBM model).

4.5.2 The denominator degrees of freedom

- (The numerator degrees of freedom of the non-central F distribution is always unity.) It remains to calculate the appropriate denominator degrees of freedom for the pivotal study. This is denoted df_2, to distinguish it from ddf_H, where the latter applies to the pilot study as in Equation (4.19).
- The starting point is Equation (4.19) with the left hand side replaced by df_2, and with the emphasis that *all quantities appearing in it apply to the pivotal study.*
- The mean squares appearing in Equation (4.19) can be related to the variances by an equation analogous to Equation (4.32), except that, again, all quantities in it apply to the *pivotal* study (note the absence of asterisks):

$$\left.\begin{aligned} \sigma_\epsilon^2 &= MSTRC \\ \sigma_{\tau R}^2 &= \frac{MSTR - MSTRC}{K} \\ \sigma_{\tau C}^2 &= \frac{MSTC - MSTRC}{J} \end{aligned}\right\} \tag{4.33}$$

Substituting from Equation (4.33) into Equation (4.19) with the left hand side replaced by df_2, and dividing numerator and denominator by K^2, one has the final expression as coded in `RJafroc`:

$$df_2 = \frac{(\max(\sigma_{\tau R}^2, 0) + (\max(J\sigma_{\tau C}^2, 0) + \sigma_\epsilon^2)/K)^2}{(\max(\sigma_{\tau R}^2, 0) + \sigma_\epsilon^2/K)^2}(J-1) \tag{4.34}$$

4.5.3 Example of sample size estimation, RRRC generalization

The Van Dyke dataset is regarded as a pilot study. In the first block of code function `StSignificanceTesting()` is used to get the DBM variances (i.e., `VarTR` $= \sigma_{\tau R}^2$, etc.) and the effect size d.

```
rocData <- dataset02 # select Van Dyke dataset
retDbm <- StSignificanceTesting(dataset = rocData,
                                FOM = "Wilcoxon",
                                method = "DBM")
VarTR <- retDbm$ANOVA$VarCom["VarTR","Estimates"]
VarTC <- retDbm$ANOVA$VarCom["VarTC","Estimates"]
VarErr <- retDbm$ANOVA$VarCom["VarErr","Estimates"]
d <- retDbm$FOMs$trtMeanDiffs["trt0-trt1","Estimate"]
```

The observed effect size is -0.04380032. The sign is negative as the reader-averaged second modality has greater FOM than the first. The next code block shows implementation of the RRRC formulae just presented. The values of J and K were preselected to achieve 80% power, as verified from the final line of the output.

```
#RRRC
J <- 10; K <- 163
den <- max(VarTR, 0) + (VarErr + J * max(VarTC, 0)) / K
deltaRRRC <- (d^2 * J/2) / den
df2 <- den^2 * (J - 1) / (max(VarTR, 0) + VarErr / K)^2
fvalueRRRC <- qf(1 - alpha, 1, df2)
Power <- 1-pf(fvalueRRRC, 1, df2, ncp = deltaRRRC)
```

```
data.frame("J"= J,  "K" = K, "fvalueRRRC" = fvalueRRRC, "df2" = df2, "deltaRRRC" = delt
#>    J   K fvalueRRRC       df2 deltaRRRC  PowerRRRC
#> 1 10 163  3.9930236 63.137871 8.1269825 0.80156249
```

4.6 Significance testing and sample size estimation for fixed-reader random-case generalization

The extension to FRRC generalization is as follows. One sets $\sigma_R^2 = 0$ and $\sigma_{\tau R}^2 = 0$ in the DBM model (4.4). The F-statistic for testing the NH and its distribution under the NH is:

$$F = \frac{\text{MST}}{\text{MSTC}} \sim F_{I-1,(I-1)(K-1)} \tag{4.35}$$

The NH is rejected if the observed value of F exceeds the critical value defined by $F_{\alpha,I-1,(I-1)(K-1)}$. For two modalities the denominator degrees of freedom is $df_2 = K - 1$. The expression for the non-centrality parameter follows from (4.31) upon setting $\sigma_{\tau R}^2 = 0$.

$$\Delta = \frac{Jd^2/2}{(\sigma_\epsilon^2 + \max(J\sigma_{\tau C}^2, 0))/K} \tag{4.36}$$

These equations are coded in the following code-chunk:

```
#FRRC
# set VarTC = 0 in RRRC formulae
J <- 10; K <- 133
den <- (VarErr + J * max(VarTC, 0)) / K
deltaFRRC <- (d^2 * J/2) / den
df2FRRC <- K - 1
fvalueFRRC <- qf(1 - alpha, 1, df2FRRC)
powerFRRC <- pf(fvalueFRRC, 1, df2FRRC, ncp = deltaFRRC, FALSE)
data.frame("J"= J,  "K" = K, "fvalueFRRC" = fvalueFRRC, "df2" = df2FRRC, "deltaFRRC" =
#>    J   K fvalueFRRC df2 deltaFRRC  powerFRRC
#> 1 10 133   3.912875 132 7.9873835 0.80111671
```

4.7 Significance testing and sample size estimation for random-reader fixed-case generalization

The extension to RRFC generalization is as follows. One sets $\sigma_C^2 = 0$ and $\sigma_{\tau C}^2 = 0$ in the DBM model (4.4). The F-statistic for testing the NH and its distribution under the NH is:

$$F = \frac{\text{MST}}{\text{MSTR}} \sim F_{I-1,(I-1)(J-1)} \tag{4.37}$$

The NH is rejected if the observed value of F exceeds the critical value defined by $F_{\alpha,I-1,(I-1)(J-1)}$. For two modalities the denominator degrees of freedom is $df_2 = J-1$. The expression for the non-centrality parameter follows from (4.31) upon setting $\sigma_{\tau C}^2 = 0$.

$$\Delta = \frac{Jd^2/2}{\max(\sigma_{\tau R}^2, 0) + \sigma_\epsilon^2/K} \tag{4.38}$$

These equations are coded in the following code-chunk:

```
#RRFC
# set VarTR = 0 in RRRC formulae
J <- 10; K <- 53
den <- max(VarTR, 0) + VarErr/K
deltaRRFC <- (d^2 * J/2) / den
df2RRFC <- J - 1
fvalueRRFC <- qf(1 - alpha, 1, df2RRFC)
powerRRFC <- pf(fvalueRRFC, 1, df2RRFC, ncp = deltaRRFC, FALSE)
data.frame("J"= J,  "K" = K, "fvalueRRFC" = fvalueRRFC, "df2" = df2RRFC, "deltaRRFC" = deltaRRFC,
#>    J  K fvalueRRFC df2 deltaRRFC  powerRRFC
#> 1 10 53   5.117355   9 10.048716 0.80496663
```

It is evident that for this dataset, for 10 readers, the numbers of cases needed for 80% power is largest (163) for RRRC and least for RRFC (53). For all three analyses, the expectation of 80% power is met - the numbers of cases and readers were deliberately chosen to achieve close to 80% statistical power.

4.8 Summary TBA

This chapter has detailed analysis of MRMC ROC data using the DBM method. A reason for the level of detail is that almost all of the material carries over to

other data collection paradigms, and a thorough understanding of the relatively simple ROC paradigm data is helpful to understanding the more complex ones.

DBM has been used in several hundred ROC studies (Prof. Kevin Berbaum, private communication ca. 2010). While the method allows generalization of a study finding, e.g., rejection of the NH, to the population of readers and cases, I believe this is sometimes taken too literally. If a study is done at a single hospital, then the radiologists tend to be more homogenous as compared to sampling radiologists from different hospitals. This is because close interactions between radiologists at a hospital tend to homogenize reading styles and performance. A similar issue applies to patient characteristics, which are also expected to vary more between different geographical locations than within a given location served by the hospital. This means is that single hospital study based p-values may tend to be biased downwards, declaring differences that may not be replicable if a wider sampling "net" were used using the same sample size. The price paid for a wider sampling net is that one must use more readers and cases to achieve the same sensitivity to genuine treatment effects, i.e., statistical power (i.e., there is no "free-lunch").

A third MRMC ROC method, due to Clarkson, Kupinski and Barrett19,20, implemented in open-source JAVA software by Gallas and colleagues22,44 (http://didsr.github.io/iMRMC/) is available on the web. Clarkson et al19,20 provide a probabilistic rationale for the DBM model, provided the figure of merit is the empirical AUC. The method is elegant but it is only applicable as long as one is using the empirical AUC as the figure of merit (FOM) for quantifying observer performance. In contrast the DBM approach outlined in this chapter, and the approach outlined in the following chapter, are applicable to any scalar FOM. Broader applicability ensures that significance-testing methods described in this, and the following chapter, apply to other ROC FOMs, such as binormal model or other fitted AUCs, and more importantly, to other observer performance paradigms, such as free-response ROC paradigm. An advantage of the Clarkson et al. approach is that it predicts truth-state dependence of the variance components. One knows from modeling ROC data that diseased cases tend to have greater variance than non-diseased ones, and there is no reason to suspect that similar differences do not exist between the variance components.

Testing validity of an analysis method via simulation testing is only as good as the simulator used to generate the datasets, and this is where current research is at a bottleneck. The simulator plays a central role in ROC analysis. In my opinion this is not widely appreciated. In contrast, simulators are taken very seriously in other disciplines, such as cosmology, high-energy physics and weather forecasting. The simulator used to validate3 DBM is that proposed by Roe and Metz39 in 1997. This simulator has several shortcomings. (a) It assumes that the ratings are distributed like an equal-variance binormal model, which is not true for most clinical datasets (recall that the b-parameter of the binormal model is usually less than one). Work extending this simulator to unequal variance has been published3. (b) It does not take into account that

some lesions are not visible, which is the basis of the contaminated binormal model (CBM). A CBM model based simulator would use equal variance distributions with the difference that the distribution for diseased cases would be a mixture distribution with two peaks. The radiological search model (RSM) of free-response data, Chapter 16 &17 also implies a mixture distribution for diseased cases, and it goes farther, as it predicts some cases yield no z-samples, which means they will always be rated in the lowest bin no matter how low the reporting threshold. Both CBM and RSM account for truth dependence by accounting for the underlying perceptual process. (c) The Roe-Metz simulator is out dated; the parameter values are based on datasets then available (prior to 1997). Medical imaging technology has changed substantially in the intervening decades. d Finally, the methodology used to arrive at the proposed parameter values is not clearly described. Needed is a more realistic simulator, incorporating knowledge from alternative ROC models and paradigms that is calibrated, by a clearly defined method, to current datasets.

Since ROC studies in medical imaging have serious health-care related consequences, no method should be used unless it has been thoroughly validated. Much work still remains to be done in proper simulator design, on which validation is dependent.

4.9 Things for me to think about

4.9.1 Expected values of mean squares

Assuming no replications the expected mean squares are as follows, Table Table 4.1; understanding how this table is derived, would lead me outside my expertise and the scope of this book; suffice to say that these are *unconstrained* estimates (as summarized in the quotation above) which are different from the *constrained* estimates appearing in the original DBM publication (Dorfman et al., 1992), Table 9.2; the differences between these two types of estimates is summarized in (Dorfman et al., 1995). For reference, Table 9.3 is the table published in the most recent paper that I am aware of (Hillis, 2014). All three tables are different! **In this chapter I will stick to Table Table 4.1 for the subsequent development.**

Table 4.2: Table 9.1 Unconstrained expected values of mean-squares, as in (Dorfman et al., 1995)

Source	df	E(MS)
T	(I-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + K\sigma^2_{\tau R} + J\sigma^2_{\tau C} + JK\sigma^2_\tau$
R	(J-1)	$\sigma^2_\epsilon + I\sigma^2_{RC} + IK\sigma^2_R + K\sigma^2_{\tau R}$
C	(K-1)	$\sigma^2_\epsilon + I\sigma^2_{RC} + IJ\sigma^2_C + J\sigma^2_{\tau C}$
TR	(I-1)(J-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + K\sigma^2_{\tau R}$

Source	df	E(MS)
TC	(I-1)(K-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + J\sigma^2_{\tau C}$
RC	(J-1)(K-1)	$\sigma^2_\epsilon + I\sigma^2_{RC}$
TRC	(I-1)(J-1)(K-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC}$
ϵ	$N-1=0$	σ^2_ϵ

Table 4.3: Table 9.2 Constrained expected values of mean-squares, as in (Dorfman et al., 1992)

Source	df	E(MS)
T	(I-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + K\sigma^2_{\tau R} + J\sigma^2_{\tau C} + JK\sigma^2_\tau$
R	(J-1)	$\sigma^2_\epsilon + I\sigma^2_{RC} + IK\sigma^2_R$
C	(K-1)	$\sigma^2_\epsilon + I\sigma^2_{RC} + IJ\sigma^2_C$
TR	(I-1)(J-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + K\sigma^2_{\tau R}$
TC	(I-1)(K-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + J\sigma^2_{\tau C}$
RC	(J-1)(K-1)	$\sigma^2_\epsilon + I\sigma^2_{RC}$
TRC	(I-1)(J-1)(K-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC}$
ϵ	0	σ^2_ϵ

Table 4.4: Table 9.3 As in Hillis "marginal-means ANOVA paper" (Hillis, 2014)

Source	df	E(MS)
T	(I-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + K\sigma^2_{\tau R} + J\sigma^2_{\tau C} + JK\sigma^2_\tau$
R	(J-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + I\sigma^2_{RC} + IK\sigma^2_R + K\sigma^2_{\tau R}$
C	(K-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + I\sigma^2_{RC} + IJ\sigma^2_C + J\sigma^2_{\tau C}$
TR	(I-1)(J-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + K\sigma^2_{\tau R}$
TC	(I-1)(K-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + J\sigma^2_{\tau C}$
RC	(J-1)(K-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC} + I\sigma^2_{RC}$
TRC	(I-1)(J-1)(K-1)	$\sigma^2_\epsilon + \sigma^2_{\tau RC}$
ϵ	0	σ^2_ϵ

4.10 Chapter References

Chapter 5

DBM method special cases

Special cases of DBM analysis are described here, namely fixed-reader random-case (FRRC), sub-special case of which is Single-reader multiple-treatment analysis, and random-reader fixed-case (RRFC).

5.1 TBA How much finished

30%

5.2 Fixed-reader random-case (FRRC) analysis

The model is the same as in Eqn. (4.4) except one sets $\sigma_R^2 = \sigma_{\tau R}^2 = 0$ in Table 4.1. The appropriate test statistic is:

$$\frac{E\left(MST\right)}{E\left(MSTC\right)} = \frac{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2 + JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2} \tag{5.1}$$

Under the null hypothesis $\sigma_\tau^2 = 0$:

$$\frac{E\left(MST\right)}{E\left(MSTC\right)} = 1 \tag{5.2}$$

The F-statistic is (replacing *expected* with *observed* values):

$$F_{DBM|R} = \frac{MST}{MSTC} \tag{5.3}$$

The observed value $F_{DBM|R}$ (the Roe-Metz notation (Roe and Metz, 1997a) is used which indicates that the factor appearing to the right of the vertical bar is regarded as fixed) is distributed as an F-statistic with ndf = I–1 and $ddf = (I-1)(K-1)$; the degrees of freedom follow from the rows labeled T and TC in TBA Table Table 4.1. Therefore, the distribution of the observed value is (no Satterthwaite approximation needed this time as both numerator and denominator are simple mean-squares):

$$F_{DBM|R} \sim F_{I-1,(I-1)(K-1)} \tag{5.4}$$

The null hypothesis is rejected if the observed value of the F- statistic exceeds the critical value:

$$F_{DBM|R} > F_{1-\alpha,I-1,(I-1)(K-1)} \tag{5.5}$$

The p-value of the test is the probability that a random sample from the F-distribution TBA (4.1) Eqn. (9.39), exceeds the observed value:

$$p = \Pr\left(F > F_{DBM|R} \mid F \sim F_{I-1,(I-1)(K-1)}\right) \tag{5.6}$$

The $(1-\alpha)$ confidence interval for the inter-treatment reader-averaged difference FOM is given by:

$$CI_{1-\alpha} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,(I-1)(K-1)}\sqrt{2\frac{MST}{JK}} \tag{5.7}$$

5.2.1 Single-reader multiple-treatment analysis

With a single reader interpreting cases in two or more treatments, the reader factor must necessarily be regarded as fixed. The preceding analysis is applicable. One simply puts $J = 1$ in the equations above.

5.2.1.1 Example 5: Code illustrating p-values for FRRC analysis, Van Dyke data

```
alpha <- 0.05
retMS <- UtilMeanSquares(dataset02)
I <- length(dataset02$ratings$NL[,1,1,1])
J <- length(dataset02$ratings$NL[1,,1,1])
K <- length(dataset02$ratings$NL[1,1,,1])
FDbmFR <- retMS$msT / retMS$msTC
```

```
ndf <- (I-1); ddf <- (I-1)*(K-1)
pValue <- 1 - pf(FDbmFR, ndf, ddf)

theta <- as.matrix(UtilFigureOfMerit(dataset02, FOM = "Wilcoxon"))
theta_i_dot <- array(dim = I)
for (i in 1:I) theta_i_dot[i] <- mean(theta[i,])

trtDiff <- array(dim = c(I,I))
for (i1 in 1:(I-1)) {
  for (i2 in (i1+1):I) {
    trtDiff[i1,i2] <- theta_i_dot[i1]- theta_i_dot[i2]
  }
}
trtDiff <- trtDiff[!is.na(trtDiff)]
nDiffs <- I*(I-1)/2

std_DIFF_FOM_FRRC <- sqrt(2*retMS$msTC/J/K)
nDiffs <- I*(I-1)/2
CI_DIFF_FOM_FRRC <- array(dim = c(nDiffs, 3))
for (i in 1 : nDiffs) {
  CI_DIFF_FOM_FRRC[i,1] <- qt(alpha/2,df = ddf)*std_DIFF_FOM_FRRC + trtDiff[i]
  CI_DIFF_FOM_FRRC[i,2] <- trtDiff[i]
  CI_DIFF_FOM_FRRC[i,3] <- qt(1-alpha/2,df = ddf)*std_DIFF_FOM_FRRC + trtDiff[i]
  print(data.frame("pValue" = pValue,
                   "Lower" = CI_DIFF_FOM_FRRC[i,1],
                   "Mid" = CI_DIFF_FOM_FRRC[i,2],
                   "Upper" = CI_DIFF_FOM_FRRC[i,3]))
}
#>       pValue       Lower         Mid        Upper
#> 1 0.02103497 -0.08088303 -0.04380032 -0.006717613

retRJafroc <- StSignificanceTesting(dataset02, FOM = "Wilcoxon", method = "DBM")

data.frame("pValue" = retRJafroc$FRRC$FTests$p[1],
           "Lower" = retRJafroc$FRRC$ciDiffTrt[1,"CILower"],
           "Mid" = retRJafroc$FRRC$ciDiffTrt[1,"Estimate"],
           "Upper" = retRJafroc$FRRC$ciDiffTrt[1,"CIUpper"])
#>        pValue        Lower          Mid         Upper
#> 1 0.021034969 -0.080883031 -0.043800322 -0.0067176131
```

As one might expect, if one "freezes" reader variability, the FOM difference becomes significant, whether viewed from the point of view of the F-statistic exceeding the critical value, the observed p-value being smaller than alpha or the 95% CI for the difference FOM not including zero.

5.3 Random-reader fixed-case (RRFC) analysis

The model is the same as in TBA (4.1) Eqn. (9.4) except one puts $\sigma_C^2 = \sigma_{\tau C}^2 = 0$ in Table Table 4.1. It follows that:

$$\frac{E(MST)}{E(MSTR)} = \frac{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2 + JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2} \tag{5.8}$$

Under the null hypothesis $\sigma_\tau^2 = 0$:

$$\frac{E(MST)}{E(MSTR)} = 1 \tag{5.9}$$

Therefore, one defines the F-statistic (replacing expected values with observed values) by:

$$F_{DBM|C} \sim \frac{MST}{MSTR} \tag{5.10}$$

The observed value $F_{DBM|C}$ is distributed as an F-statistic with $ndf = I$–1 and $ddf = (I-1)(J-1)$, see rows labeled T and TR in Table Table 4.1.

$$F_{DBM|C} \sim F_{I-1,(I-1)(J-1))} \tag{5.11}$$

The null hypothesis is rejected if the observed value of the F statistic exceeds the critical value:

$$F_{DBM|C} > F_{1-\alpha,I-1,(I-1)(J-1))} \tag{5.12}$$

The p-value of the test is the probability that a random sample from the distribution exceeds the observed value:

$$p = \Pr\left(F > F_{DBM|C} \mid F \sim F_{I-1,(I-1)(J-1)}\right) \tag{5.13}$$

The confidence interval for inter-treatment differences is given by (TBA check this):

$$CI_{1-\alpha} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,(I-1)(J-1)}\sqrt{2\frac{MSTR}{JK}} \tag{5.14}$$

5.3.0.1 Example 6: Code illustrating analysis for RRFC analysis, Van Dyke data

```
FDbmFC <- retMS$msT / retMS$msTR
ndf <- (I-1)
ddf <- (I-1)*(J-1)
pValue <- 1 - pf(FDbmFC, ndf, ddf)

nDiffs <- I*(I-1)/2
CI_DIFF_FOM_RRFC <- array(dim = c(nDiffs, 3))
for (i in 1 : nDiffs) {
  CI_DIFF_FOM_RRFC[i,1] <- qt(alpha/2,df = ddf)*sqrt(2*retMS$msTR/J/K) + trtDiff[i]
  CI_DIFF_FOM_RRFC[i,2] <- trtDiff[i]
  CI_DIFF_FOM_RRFC[i,3] <- qt(1-alpha/2,df = ddf)*sqrt(2*retMS$msTR/J/K) + trtDiff[i]
  print(data.frame("pValue" = pValue,
                   "Lower" = CI_DIFF_FOM_RRFC[i,1],
                   "Mid" = CI_DIFF_FOM_RRFC[i,2],
                   "Upper" = CI_DIFF_FOM_RRFC[i,3]))
}
#>        pValue        Lower          Mid         Upper
#> 1 0.041958752 -0.085020224 -0.043800322 -0.0025804202
data.frame("pValue" = retRJafroc$RRFC$FTests$p[1],
           "Lower" = retRJafroc$RRFC$ciDiffTrt[1,"CILower"],
           "Mid" = retRJafroc$RRFC$ciDiffTrt[1,"Estimate"],
           "Upper" = retRJafroc$RRFC$ciDiffTrt[1,"CIUpper"])
#>        pValue        Lower          Mid         Upper
#> 1 0.041958752 -0.085020224 -0.043800322 -0.0025804202
```

5.4 Chapter References

Chapter 6

Introduction to the Obuchowski-Rockette method

6.1 TBA How much finished

70%

6.2 Locations of helper functions

```
source(here("R/CH10-OR/Wilcoxon.R"))
source(here("R/CH10-OR/VarCov1FomInput.R"))
source(here("R/CH10-OR/VarCov1Bs.R"))
source(here("R/CH10-OR/VarCov1Jk.R"))
source(here("R/CH10-OR/VarCovMtrxDLStr.R"))
source(here("R/CH10-OR/VarCovs.R"))
```

6.3 Introduction

- This chapter starts with a gentle introduction to the Obuchowski and Rockette method. The reason is that the method was rather opaque to me, and I suspect most non-statistician users. Part of the problem, in my opinion, is the notation, namely lack of the *case-set* index $\{c\}$. While this

may seem like a trivial point to statisticians, it did present a conceptual problem for me.

- A key difference of the Obuchowski and Rockette method from DBM is in how the error term is modeled by a non-diagonal covariance matrix. Therefore, the structure of the covariance matrix is examined in some detail.

- To illustrate the covariance matrix, a single reader interpreting a case-set in multiple treatments is analyzed and the results compared to that using DBM fixed-reader analysis described in previous chapters.

6.4 Single-reader multiple-treatment

6.4.1 Overview

Consider a single-reader interpreting a common case-set $\{c\}$ in multiple-treatments i ($i = 1, 2, ..., I$).

In the OR method one models the figure-of-merit, not the pseudovalues; indeed this is a key differences from the DBM method. The figure of merit θ is modeled as:

$$\theta_{i\{c\}} = \mu + \tau_i + \epsilon_{i\{c\}} \tag{6.1}$$

Eqn. (6.1) models the observed figure-of-merit $\theta_{i\{c\}}$ as a constant term μ, a treatment dependent term τ_i (the treatment-effect), and a random term $\epsilon_{i\{c\}}$. The term τ_i has the constraint:

$$\sum_{i=1}^{I} \tau_i = 0 \tag{6.2}$$

The left hand side of Eqn. (6.1) is the figure-of-merit $\theta_i\{c\}$ for treatment i and case-set index $\{c\}$, where $c = 1, 2, ..., C$ denotes different independent case-sets sampled from the population, i.e., different *collections* of K_1 non-diseased and K_2 diseased cases.

The case-set index is essential for clarity. Without it θ_i is a fixed quantity - the figure of merit estimate for treatment i - lacking an index allowing for sampling related variability. Obuchowski and Rockette define a *k-index*, the:

> k^{th} repetition of the study involving the same diagnostic test, reader and patient (sic)".

Needed is a *case-set* index rather than a *repetition* index. Repeating a study with the same treatment, reader and cases yields *within-reader* variability, when what is needed, for significance testing, is *case-sampling plus within-reader* variability.

It is shown below that usage of the case-set index interpretation yields the same results using the DBM or the OR methods (for empirical AUC).

Eqn. (6.1) has an additive random error term $\epsilon_{i\{c\}}$ whose sampling behavior is described by a multivariate normal distribution with an I-dimensional zero mean vector and an $I \times I$ dimensional covariance matrix Σ:

$$\epsilon_{i\{c\}} \sim N_I\left(\vec{0}, \Sigma\right) \tag{6.3}$$

Here N_I is the I-variate normal distribution (i.e., each sample yields I random numbers). For the single-reader model Eqn. (6.1), the covariance matrix has the following structure :

$$\Sigma_{ii'} = Cov\left(\epsilon_{i\{c\}}, \epsilon_{i'\{c\}}\right) = \begin{cases} \text{Var} & (i = i') \\ Cov_1 & (i \neq i') \end{cases} \tag{6.4}$$

The reason for the subscript "1" in Cov_1 will become clear when we extend this model to multiple- treatments and multiple-readers. The $I \times I$ covariance matrix Σ is:

$$\Sigma = \begin{pmatrix} \text{Var} & Cov_1 & ... & Cov_1 & Cov_1 \\ Cov_1 & \text{Var} & ... & Cov_1 & Cov_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Cov_1 & Cov_1 & ... & \text{Var} & Cov_1 \\ Cov_1 & Cov_1 & ... & Cov_1 & \text{Var} \end{pmatrix} \tag{6.5}$$

If $I = 2$ then Σ is a symmetric 2 x 2 matrix, whose diagonal terms are the common variances in the two treatments (each assumed equal to Var) and whose off-diagonal terms (each assumed equal to Cov_1) are the co-variances. With $I = 3$ one has a 3 x 3 symmetric matrix with all diagonal elements equal to Var and all off-diagonal terms are equal to Cov_1, etc.

An important aspect of the Obuchowski and Rockette model is that the variances and co-variances are assumed to be treatment independent. This implies that Var estimates need to be averaged over all treatments. Likewise, Cov_1 estimates need to be averaged over all distinct treatment-treatment pairings.

[1]

Some elementary statistical results are presented in the Appendix.

[1] A more complex model, with more parameters and therefore more difficult to work with, would allow the variances to be treatment dependent, and the covariances to depend on the specific treatment pairings. For obvious reasons ("Occam's Razor" or the law of parsimony) one wishes to start with the simplest model that, one hopes, captures essential characteristics of the data.

6.4.2 Significance testing

The covariance matrix is needed for significance testing. Define the mean square corresponding to the treatment effect, denoted $MS(T)$, by:

$$MS(T) = \frac{1}{I-1}\sum_{i=1}^{I}(\theta_i - \theta_\bullet)^2 \tag{6.6}$$

Unlike the previous DBM related chapters, all mean square quantities in this chapter are based on FOMs, not pseudovalues.

It can be shown that under the null hypothesis that all treatments have identical performances, the test statistic χ_{1R} defined below (the $1R$ subscript denotes single-reader analysis) is distributed approximately as a χ^2 distribution with $I-1$ degrees of freedom, i.e.,

$$\chi_{1\text{R}} \equiv \frac{(I-1)MS(T)}{\text{Var} - \text{Cov1}} \sim \chi^2_{I-1} \tag{6.7}$$

Eqn. (6.7) is from §5.4 in (Hillis, 2007) with two covariance terms "zeroed out" because they are multiplied by $J-1=0$ (since we are restricting to $J=1$).

Or equivalently, in terms of the F-distribution (Hillis et al., 2005):

$$F_{1\text{R}} \equiv \frac{MS(T)}{\text{Var} - \text{Cov1}} \sim F_{I-1,\infty} \tag{6.8}$$

6.4.3 p-value and confidence interval

The p-value is the probability that a sample from the $F_{I-1,\infty}$ distribution is greater than the observed value of the test statistic, namely:

$$p \equiv \Pr(f > F_{1R} \mid f \sim F_{I-1,\infty}) \tag{6.9}$$

The $(1-\alpha)$ confidence interval for the inter-treatment FOM difference is given by:

$$CI_{1-\alpha,1R} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,\infty}\sqrt{2(\text{Var} - \text{Cov1})} \tag{6.10}$$

Comparing Eqn. (6.10) to Eqn. (6.27) shows that the term $\sqrt{2(\text{Var} - \text{Cov1})}$ is the standard error of the inter-treatment FOM difference, whose square root is the standard deviation. The term $t_{\alpha/2,\infty}$ is -1.96. Therefore, the confidence interval is constructed by adding and subtracting 1.96 times the standard deviation of the difference from the central value. [One has probably encountered the rule that a 95% confidence interval is plus or minus two standard deviations from the central value. The "2" comes from rounding up 1.96.]

6.4.4 Null hypothesis validation

It is important to validate the significance testing method just outlined above. If the testing procedure is valid, then, when the NH is true, the procedure should reject it with probability α. In the following, as is usual, we set $\alpha = 0.05$.

```
set.seed(seed = 201)
mu <- 0.8
vc <- UtilORVarComponentsFactorial(dataset02, FOM = "Wilcoxon")
trueVar <- vc$IndividualRdr$varEachRdr[1]
trueCov1 <- vc$IndividualRdr$cov1EachRdr[1]
sigma <- matrix(c(trueVar,
                  trueCov1,
                  trueCov1,
                  trueVar),
                  ncol = 2)
I <- 2
S <-  2000
# simulate foms for two modalities, S times
# using the sampling model
theta_i <- t(rmvnorm(n=S, mean=c(0,0), sigma=sigma) + mu)
# estimated variance covariances
vc <- VarCov1_FomInput(theta_i)
Var <- vc$Var
Cov1 <- vc$Cov1

# conduct NH testing
reject <- 0
for(s in 1:S) {

  MS_T <- 0
  for (i in 1:I) {
    MS_T <- MS_T + (theta_i[i,s]-mean(theta_i[,s]))^2
  }
  MS_T <- MS_T/(I-1)

  F_1R <- MS_T/(Var - Cov1)
  pValue <- 1 - pf(F_1R, I-1, Inf)
  if (pValue < 0.05) reject <- reject + 1
}
alphaObs <- reject/S
```

```
## True, estimated diagonal elements =  0.000699,   0.000695
```

```
## True, estimated off-diagonal elements =  0.000373,   0.000351
```

```
## NH rejection fraction =    0.0515
```

The `seed` variable, set to 201 at line 1, is equivalent to the case sample index c in Eqn. (6.1). Different values of `seed` correspond to different case samples.

Line 2 sets the value of μ to 0.8, the average figure of merit, appearing in Eqn. (6.1).

Lines 3-4 set the values of true Var and true Cov_1 to values characterizing `dataset02` for reader one, as determined by function `UtilORVarComponentsFactorial`.

Lines 5-9 initializes the covariance matrix Σ. The diagonal contains the variance and the off-diagonal contains Cov_1. These are the *true* values.

Lines 10-11 initializes $I = 2$, the number of treatments, and $S = 2000$, the number of simulations.

Line 14 generates 2000 samples from the two dimensional multivariate normal distribution with zero mean vector (**this is the null hypothesis**) and covariance equal to Σ.

Lines 16-18 computes the *estimates* of the means and covariances. The helper function used `VarCov1_FomInput` (the name stands for Var and Cov_1 using FOM input) is included in the distribution. The locations of helper functions are shown in Section 6.2.

Lines 21-33 performs the NH testing. It starts by setting the counter variable `reject` to zero. A for-loop is set up to repeat 2000 times. For each iteration line 24-28 computes the treatment mean-square `MS_T`. Note the use, at line 25, of the two values of $theta_i$ corresponding to the s-th sample from the multivariate norrmal distribution (at line 14). Line 30 computes the F-statistic - compare to Eqn. (6.8). Line 31 computes the p-values and, if the p-value is less than $\alpha = 0.05$, line 32 increments `reject` by one. The observed NH rejection rate, `alphaObs`, is the final value of `reject` divided by 2000, line 34. For a valid test it is expected to be in the range (0.04, 0.06). The actual value, for the chosen value of `seed`, is 0.0515.

6.4.5 Application 1

Here is an application of the method to an ROC dataset, `dataset02`, which consists of two treatments and five readers.

```
1 ds <- DfExtractDataset(dataset02, rdrs = 1)
2 fom <- as.vector(unlist(UtilFigureOfMerit(ds, FOM = "Wilcoxon")), mode = "numeric")
3 vc <- UtilORVarComponentsFactorial(ds, FOM = "Wilcoxon")
4 Cov1 <- vc$IndividualRdr$cov1EachRdr
5 Var <- vc$IndividualRdr$varEachRdr
6 msT <- vc$IndividualRdr$msTEachRdr
```

```
I <- length(ds$ratings$NL[,1,1,1])
chiObs <- (I-1)*msT/(Var-Cov1)
pval <- pchisq(chiObs,I-1,lower.tail = F)
ci <- array(dim = 2)
ci[1] <- (fom[1] - fom[2]) + qt(0.025, Inf, lower.tail = F) * sqrt(2*(Var - Cov1))
ci[2] <- (fom[1] - fom[2]) - qt(0.025, Inf, lower.tail = F) * sqrt(2*(Var - Cov1))
```

```
## fom =  0.9196457 0.9478261
```

```
## fom diff =  -0.02818035
```

```
## pval =  0.2693389
```

```
## ci =  0.02182251 -0.07818322
```

We extract the data for reader 1 only, line 1, resulting in a 2-treatment single-reader dataset `ds`. Lines 2-3 compute the Wilcoxon figures of merit for each treatment as a row vector. Lines 4-7 computes OR treatment mean square `msT`, the OR variance components `Var` and `Cov1`: function`UtilORVarComponentsFactorial` is used with the Wilcoxon figure of merit specified. Line 8 obtains the number of treatments, $I = 2$ in this example. Line 9 computes the observed chisquare statistic, `chiObs`. Line 10 computes the p-value, `pValue`, i.e., the probability that a sample from a chisquare distribution with I-1 degrees of freedom exceeds the observed value. Lines 11-13 compute the 95% confidence interval for the inter-treatment FOM difference. For this reader the two treatments are not significantly different.

6.4.6 Application 2

Here is an application of the method to an FROC dataset, `dataset04`, which consists of five treatments and four readers.

```
ds <- DfExtractDataset(dataset04, rdrs = 1, trts = c(4,5))
fom <- as.vector(unlist(UtilFigureOfMerit(ds, FOM = "wAFROC")), mode = "numeric")
vc <- UtilORVarComponentsFactorial(ds, FOM = "wAFROC")
Cov1 <- vc$IndividualRdr$cov1EachRdr
Var <- vc$IndividualRdr$varEachRdr
msT <- vc$IndividualRdr$msTEachRdr
I <- length(ds$ratings$NL[,1,1,1])
chiObs <- (I-1)*msT/(Var-Cov1)
pval <- pchisq(chiObs,I-1,lower.tail = F)
ci <- array(dim = 2)
```

```
11 ci[1] <- (fom[1] - fom[2]) +
12   qt(0.025, Inf, lower.tail = F) * sqrt(2*(Var - Cov1))
13 ci[2] <- (fom[1] - fom[2]) -
14   qt(0.025, Inf, lower.tail = F) * sqrt(2*(Var - Cov1))
```

```
## fom =  0.8101333 0.7488

## fom diff =  0.06133333

## pval =  0.03189534

## ci =  0.117357 0.005309652
```

We extract the data for reader 1 only, for treatments 4 and 5, line 1, resulting in a 2-treatment single-reader dataset `ds`. Lines 2-3 compute the wAFROC figures of merit for each treatment as a row vector. Lines 4-7 computes OR treatment mean square `msT`, the OR variance components `Var` and `Cov1`: function`UtilORVarComponentsFactorial` is used with the wAFROC figure of merit specified. Line 8 obtains the number of treatments, $I = 2$ in this example. Line 9 computes the observed chisquare statistic, `chiObs`. Line 10 computes the p-value, `pValue`, i.e., the probability that a sample from a chisquare distribution with I-1 degrees of freedom exceeds the observed value. Lines 11-13 compute the 95% confidence interval for the inter-treatment FOM difference. For this reader the two treatments are significantly different.

6.5 Single-treatment multiple-reader

6.5.1 Overview

Consider multiple readers j ($j = 1, 2, ..., J$) interpreting a common case-set $\{c\}$ in a single treatment. The OR sampling model is:

$$\theta_{j\{c\}} = \mu + R_j + \epsilon_{j\{c\}} \tag{6.11}$$

The error term $\epsilon_{j\{c\}}$ has sampling behavior described by a multivariate normal distribution with a J-dimensional zero mean vector and a $J \times J$ dimensional covariance matrix Σ:

$$\epsilon_{j\{c\}} \sim N_J\left(\vec{0}, \Sigma\right) \tag{6.12}$$

The covariance matrix has the following structure:

$$\Sigma_{jj'} = Cov\left(\epsilon_{j\{c\}}, \epsilon_{j'\{c\}}\right) = \begin{cases} \text{Var} & (j = j') \\ Cov_2 & (j \neq j') \end{cases} \tag{6.13}$$

The reason for the subscript "2" in Cov_2 will become clear when one extends this model to multiple- treatments and multiple-readers. The $J \times J$ covariance matrix Σ is:

$$\Sigma = \begin{pmatrix} \text{Var} & Cov_2 & ... & Cov_2 & Cov_2 \\ Cov_2 & \text{Var} & ... & Cov_2 & Cov_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Cov_2 & Cov_2 & ... & \text{Var} & Cov_2 \\ Cov_2 & Cov_2 & ... & Cov_2 & \text{Var} \end{pmatrix} \tag{6.14}$$

The covariance matrix is estimated, as usual, by either a resampling method (jackknife of bootstrap) or, for the special case of Wilcoxon figure of merit, by the DeLong method.

6.5.2 Significance testing

Unlike the seemingly analogous single-reader multiple-treatment case addressed in Section 6.4.2, the single-treatment multiple-reader case is fundamentally different. This is because reader is a *random* factor while treatment, in Section 6.4.2, was a *fixed* factor. This makes it impossible to define a null hypothesis analogous to that with the treatment factor, e.g., $R_1 = R_2$, since reader is modeled as a random sample from a distribution, i.e., $R \sim N(0, \sigma_R^2)$.

6.5.3 Special case

If reader is regarded as a *fixed* factor significance testing between readers can be performed. The analysis presented in Section 6.4.2 is applicable, with the treatment factor replaced by the reader factor. This is appropriate, for example, when comparing two AI (artificial intelligence) algorithms. The two algorithms, each of which qualifies as a reader, are not random samples from a population of AI readers: rather they are two fixed algorithms, in the literal sense.

6.6 Multiple-reader multiple-treatment

The previous sections introduced Obuchowski and Rockette method using single reader and single treatment examples. This section extends it to multiple-readers interpreting a common case-set in multiple-treatments (MRMC). The extension is, in principle, fairly straightforward. Compared to Eqn. (6.1), one needs an additional j index to denote reader dependence of the figure of merit,

and additional terms to model reader and treatment-reader variability, and the error term needs to be modified to account for the additional random reader factor.

The Obuchowski and Rockette model for fully paired multiple-reader multiple-treatment interpretations is:

$$\theta_{ij\{c\}} = \mu + \tau_i + R_j + (\tau R)_{ij} + \epsilon_{ij\{c\}} \tag{6.15}$$

- The fixed treatment effect τ_i is subject to the usual constraint, Eqn. (6.2).
- The first two terms on the right hand side of Eqn. (6.15) have their usual meanings: a constant term μ representing performance averaged over treatments and readers, and a treatment effect τ_i (i = 1,2, ..., I).
- The next two terms are, by assumption, mutually independent random samples specified as follows:
  - R_j denotes the random treatment-independent figure-of-merit contribution of reader j (j = 1,2, ..., J), modeled by a zero-mean normal distribution with variance σ_R^2;
  - $(\tau R)_{ij}$ denotes the treatment-dependent random contribution of reader j in treatment i, modeled as a sample from a zero-mean normal distribution with variance $\sigma_{\tau R}^2$.
- Summarizing:

$$\left.\begin{array}{c} R_j \sim N(0, \sigma_R^2) \\ \tau R \sim N(0, \sigma_{\tau R}^2) \end{array}\right\} \tag{6.16}$$

For a single dataset $c = 1$. An estimate of μ follows from averaging over the i and j indices (the averages over the random terms are zeroes):

$$\mu = \theta_{\bullet\bullet\{1\}} \tag{6.17}$$

Averaging over the j index and performing a subtraction yields an estimate of τ_i:

$$\tau_i = \theta_{i\bullet\{1\}} - \theta_{\bullet\bullet\{1\}} \tag{6.18}$$

The τ_i estimates obey the constraint Eqn. (6.2). For example, with two treatments, the values of τ_i must be the negatives of each other: $\tau_1 = -\tau_2$.

The error term on the right hand side of Eqn. (6.15) is more complex than the corresponding DBM model error term. Obuchowski and Rockette model

this term with a multivariate normal distribution with a length (IJ) zero-mean vector and a $(IJ \times IJ)$ dimensional covariance matrix Σ. In other words,

$$\epsilon_{ij\{c\}} \sim N_{IJ}(\vec{0}, \Sigma) \tag{6.19}$$

Here N_{IJ} is the IJ-variate normal distribution, $\vec{0}$ is the zero-vector with length IJ, denoting the vector-mean of the distribution. The counterpart of the variance, namely the covariance matrix Σ of the distribution, is defined by 4 parameters, $\text{Var}, \text{Cov1}, \text{Cov2}, \text{Cov3}$, defined as follows:

$$Cov(\epsilon_{ij\{c\}}, \epsilon_{i'j'\{c\}}) = \left\{ \begin{array}{l} \text{Var } (i = i', j = j') \\ \text{Cov1 } (i \neq i', j = j') \\ \text{Cov2 } (i = i', j \neq j') \\ \text{Cov3 } (i \neq i', j \neq j') \end{array} \right\} \tag{6.20}$$

Apart from fixed effects, the model implied by Eqn. (6.15) and Eqn. (6.20) contains 6 parameters:

$$\sigma^2_R, \sigma^2_{\tau R}, \text{Var}, \text{Cov1}, \text{Cov2}, \text{Cov3}$$

This is the same number of variance component parameters as in the DBM model, which should not be a surprise since one is modeling the data with equivalent models. The Obuchowski and Rockette model Eqn. (6.15) "looks" simpler because four covariance terms are encapsulated in the ϵ term. As with the singe-reader multiple-treatment model, the covariance matrix is assumed to be independent of treatment or reader.

It is implicit in the Obuchowski-Rockette model that the Var, Cov1, Cov_2, and Cov_3 estimates are averaged over all applicable treatment-reader combinations.

6.6.1 Structure of the covariance matrix

To understand the structure of this matrix, recall that the diagonal elements of a square covariance matrix are the variances and the off-diagonal elements are covariances. With two indices ij one can still imagine a square matrix where the position along each dimension is labeled by a pair of indices ij. One ij pair corresponds to the horizontal direction, and the other ij pair corresponds to the vertical direction. To visualize this let consider the simpler situation of two treatments $(I = 2)$ and three readers $(J = 3)$. The resulting 6×6 covariance matrix would look like this:

$$\Sigma = \begin{bmatrix} (11,11) & (12,11) & (13,11) & (21,11) & (22,11) & (23,11) \\ & (12,12) & (13,12) & (21,12) & (22,12) & (23,12) \\ & & (13,13) & (21,13) & (22,13) & (23,13) \\ & & & (21,21) & (22,21) & (23,21) \\ & & & & (22,22) & (23,22) \\ & & & & & (23,23) \end{bmatrix}$$

Shown in each cell of the matrix is a pair of ij-values, serving as column indices, followed by a pair of ij-values serving as row indices, and a comma separates the pairs. For example, the first column is labeled by (11,xx), where xx depends on the row. The second column is labeled (12,xx), the third column is labeled (13,xx), and the remaining columns are successively labeled (21,xx), (22,xx) and (23,xx). Likewise, the first row is labeled by (yy,11), where yy depends on the column. The following rows are labeled (yy,12), (yy,13), (yy,21), (yy,22)and (yy,23). Note that the reader index increments faster than the treatment index.

The diagonal elements are evidently those cells where the row and column index-pairs are equal. These are (11,11), (12,12), (13,13), (21,21), (22,22) and (23,23). According to Eqn. (6.20) these cells represent Var.

$$\Sigma = \begin{bmatrix} Var & (12,11) & (13,11) & (21,11) & (22,11) & (23,11) \\ & Var & (13,12) & (21,12) & (22,12) & (23,12) \\ & & Var & (21,13) & (22,13) & (23,13) \\ & & & Var & (22,21) & (23,21) \\ & & & & Var & (23,22) \\ & & & & & Var \end{bmatrix}$$

According to Eqn. (6.20) cells with different treatment indices but identical reader indices represent Cov1. As an example, cell (21,11) has the same reader indices, namely reader 1, but different treatment indices, namely 2 and 1, so it is Cov1:

$$\Sigma = \begin{bmatrix} \text{Var} & (12,11) & (13,11) & \text{Cov1} & (22,11) & (23,11) \\ & \text{Var} & (13,12) & (21,12) & \text{Cov1} & (23,12) \\ & & \text{Var} & (21,13) & (22,13) & \text{Cov1} \\ & & & \text{Var} & (22,21) & (23,21) \\ & & & & \text{Var} & (23,22) \\ & & & & & \text{Var} \end{bmatrix}$$

Similarly, cells with identical treatment indices but different reader indices represent Cov_2:

$$\Sigma = \begin{bmatrix} Var & Cov_2 & Cov_2 & \text{Cov1} & (22,11) & (23,11) \\ & \text{Var} & Cov_2 & (21,12) & \text{Cov1} & (23,12) \\ & & \text{Var} & (21,13) & (22,13) & \text{Cov1} \\ & & & \text{Var} & Cov_2 & Cov_2 \\ & & & & \text{Var} & Cov_2 \\ & & & & & \text{Var} \end{bmatrix}$$

Finally, cells with different treatment indices and different reader indices represent Cov_3:

$$\Sigma = \begin{bmatrix} \text{Var} & Cov_2 & Cov_2 & \text{Cov1} & Cov_3 & Cov_3 \\ & \text{Var} & Cov_2 & Cov_3 & \text{Cov1} & Cov_3 \\ & & \text{Var} & Cov_3 & Cov_3 & \text{Cov1} \\ & & & \text{Var} & Cov_2 & Cov_2 \\ & & & & \text{Var} & Cov_2 \\ & & & & & \text{Var} \end{bmatrix}$$

To understand these terms consider how they might be estimated. Suppose one had the luxury of repeating the study with different case-sets, c = 1, 2, ..., C. Then the variance Var is estimated as follows:

$$\text{Var} = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})^2 \right\rangle_{ij} \quad \epsilon_{ij\{c\}} \sim N_{IJ}(\vec{0}, \Sigma) \tag{6.21}$$

Of course, in practice one would use the bootstrap or the jackknife as a stand-in for the c-index (with the understanding that if the jackknife is used, then a variance inflation factor has to be included on the right hand side of Eqn. (6.21). Notice that the left-hand-side of Eqn. (6.21) lacks treatment or reader indices. This is because implicit in the notation is averaging the observed variances over all treatments and readers, as implied by $\langle\rangle_{ij}$. Likewise, the covariance terms are estimated as follows:

$$Cov = \begin{cases} \text{Cov1} = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})(\theta_{i'j\{c\}} - \theta_{i'j\{\bullet\}}) \right\rangle_{ii',jj} \\ Cov_2 = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})(\theta_{ij'\{c\}} - \theta_{ij'\{\bullet\}}) \right\rangle_{ii,jj'} \\ Cov_3 = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})(\theta_{i'j'\{c\}} - \theta_{i'j'\{\bullet\}}) \right\rangle_{ii',jj'} \end{cases} \tag{6.22}$$

In Eqn. (6.22) *the convention is that primed and unprimed variables are always different.*

Since there are no treatment and reader dependencies on the left-hand-sides of the above equations, one averages the estimates as follows:

- For Cov1 one averages over all combinations of *different treatments and same readers*, as denoted by $\langle\rangle_{ii',jj}$.
- For Cov_2 one averages over all combinations of *same treatment and different readers*, as denoted by $\langle\rangle_{ii,jj'}$.
- For Cov_3 one averages over all combinations of *different treatments and different readers*, as denoted by $\langle\rangle_{ii',jj'}$.

6.6.2 Physical meanings of the covariance terms

The meanings of the different terms follow a similar description to that given in Eqn. 6.6.1. The diagonal term Var is the variance of the figures-of-merit when reader j interprets different case-sets $\{c\}$ in treatment i. Each case-set yields a number $\theta_{ij\{c\}}$ and the variance of the C numbers, averaged over the $I \times J$ treatments and readers, is Var. It captures the total variability due to varying difficulty levels of the case-sets, inter-reader and within-reader variability.

It is easier to see the physical meanings of Cov1, Cov_2, Cov_3 if one starts with the correlations.

- $\rho_{1;ii'jj}$ is the correlation of the figures-of-merit when reader j interprets case-sets in different treatments i, i'. Each case-set, starting with $c = 1$, yields two numbers $\theta_{ij\{1\}}$ and $\theta_{i'j\{1\}}$. The correlation of the two pairs of C-length arrays, averaged over all pairings of different treatments and same readers, is ρ_1. The correlation exists due to the common contribution of the shared reader. When the common variation is large, the two arrays become more correlated and ρ_1 approaches unity. If there is no common variation, the two arrays become independent, and ρ_1 equals zero. Converting from correlation to covariance, see Eqn. (6.28), one has Cov1 < Var.

- $\rho_{2;iijj'}$ is the correlation of the figures-of-merit values when different readers j, j' interpret the same case-sets in the same treatment i. As before this yields two C-length arrays, whose correlation, upon averaging over all distinct treatment pairings and same readers, yields ρ_2. If one assumes that common variation between different-reader same-treatment FOMs is smaller than the common variation between same-reader different-treatment FOMs, then ρ_2 will be smaller than ρ_1. This is equivalent to stating that readers agree more with themselves in different treatments than they do with other readers in the same treatment. Translating to covariances, one has $Cov_2 <$ Cov1 < Var.

- $\rho_{3;ii'jj'}$ is the correlation of the figure-of-merit values when different readers j, j' interpret the same case set in different treatments i, i', etc., yielding ρ_3. This is expected to yield the least correlation.

Summarizing, one expects the following ordering for the terms in the covariance matrix:

$$Cov_3 \leq Cov_2 \leq \text{Cov1} \leq \text{Var} \tag{6.23}$$

6.7 Summary

6.8 Discussion

6.9 Appendix: Covariance and correlation

Some elementary statistical results are reviewed here.

6.9.1 Relation: chisquare and F with infinite ddf

Define $D_{1-\alpha}$, the $(1-\alpha)$ quantile of distribution D, such that the probability of observing a random sample d less than or equal to $D_{1-\alpha}$ is $(1-\alpha)$:

$$\Pr(d \leq D_{1-\alpha} \mid d \sim D) = 1-\alpha \tag{6.24}$$

With definition Eqn. (6.24), the $(1-\alpha)$ quantile of the χ^2_{I-1} distribution, i.e., $\chi^2_{1-\alpha,I-1}$, is related to the $(1-\alpha)$ quantile of the $F_{I-1,\infty}$ distribution, i.e., $F_{1-\alpha,I-1,\infty}$, as follows (see Hillis et al., 2005, Eq. 22):

$$\frac{\chi^2_{1-\alpha,I-1}}{I-1} = F_{1-\alpha,I-1,\infty} \tag{6.25}$$

Eqn. (6.25) implies that the $(1-\alpha)$ quantile of the F-distribution with $ndf = (I-1)$ and $ddf = \infty$ equals the $(1-\alpha)$ quantile of the χ^2_{I-1} distribution *divided by* $(I-1)$.

Here is an `R` illustration of this theorem for $I-1=4$ and $\alpha = 0.05$:

```
qf(0.05, 4, Inf)
```

```
## [1] 0.1776808
```

```
qchisq(0.05,4)/4
```

```
## [1] 0.1776808
```

6.9.2 Definitions of covariance and correlation

The covariance of two scalar random variables X and Y is defined by:

$$Cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i - x_\bullet)(y_i - y_\bullet)}{N-1} = E(XY) - E(X) - E(Y) \qquad (6.26)$$

Here $E(X)$ is the expectation value of the random variable X, i.e., the integral of x multiplied by its pdf over the range of x:

$$E(X) = \int \text{pdf(x)} x dx$$

The covariance can be thought of as the *common* part of the variance of two random variables. The variance, a special case of covariance, of X is defined by:

$$\text{Var}(X,X) = Cov(X,X) = E(X^2) - (E(X))^2 = \sigma_x^2$$

It can be shown, this is the Cauchy–Schwarz inequality, that:

$$\mid Cov(X,Y) \mid^2 \leq \text{Var}(X)\text{Var}(Y)$$

A related quantity, namely the correlation ρ is defined by (the σs are standard deviations):

$$\rho_{XY} \equiv Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

It has the property:

$$\mid \rho_{XY} \mid \leq 1$$

6.9.3 Special case when variables have equal variances

Assuming X and Y have the same variance:

$$\text{Var}(X) = \text{Var}(Y) \equiv \text{Var} \equiv \sigma^2$$

A useful theorem applicable to the OR single-reader multiple-treatment model is:

$$\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) - 2Cov(X,Y) = 2(\text{Var} - Cov) \tag{6.27}$$

The right hand side specializes to the OR single-reader multiple-treatment model where the variances (for different treatments) are equal and likewise the covariances in Eqn. (6.5) are equal) The correlation ρ_1 is defined by (the reason for the subscript 1 on ρ is the same as the reason for the subscript 1 on Cov1, which will be explained later):

$$\rho_1 = \frac{\text{Cov1}}{\text{Var}}$$

The I x I covariance matrix Σ can be written alternatively as (shown below is the matrix for I = 5; as the matrix is symmetric, only elements at and above the diagonal are shown):

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & & \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & & & \sigma^2 & \rho_1\sigma^2 \\ & & & & \sigma^2 \end{bmatrix} \tag{6.28}$$

6.9.4 Estimating the variance-covariance matrix

An unbiased estimate of the covariance matrix Eqn. (6.4) follows from:

$$\Sigma_{ii'} \mid_{ps} = \frac{1}{C-1} \sum_{c=1}^{C} \left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right) \left(\theta_{i'\{c\}} - \theta_{i'\{\bullet\}}\right) \tag{6.29}$$

The subscript *ps* denotes population sampling. As a special case, when $i = i'$, this equation yields the population sampling based variance.

$$\text{Var}_i \mid_{ps} = \frac{1}{C-1} \sum_{c=1}^{C} \left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)^2 \tag{6.30}$$

The I-values when averaged yield the population sampling based estimate of Var.

Sampling different case-sets, as required by Eqn. (6.29), is unrealistic. In reality one has $C = 1$, i.e., a single dataset. Therefore, direct application of this formula is impossible. However, as seen when this situation was encountered before in (book) Chapter 07, one uses resampling methods to realize, for example, different bootstrap samples, which are resampling-based "stand-ins" for actual

case-sets. If B is the total number of bootstraps, then the estimation formula is:

$$\Sigma_{ii'} \mid_{bs} = \frac{1}{B-1} \sum_{b=1}^{B} \left(\theta_{i\{b\}} - \theta_{i\{\bullet\}}\right) \left(\theta_{i'\{b\}} - \theta_{i'\{\bullet\}}\right) \tag{6.31}$$

Eqn. (6.31), the bootstrap method of estimating the covariance matrix, is a direct translation of Eqn. (6.29). Alternatively, one could have used the jackknife FOM values $\theta_{i(k)}$, i.e., the figure of merit with a case k removed, repeated for all k, to estimate the covariance matrix:

$$\Sigma_{ii'} \mid_{jk} = \frac{(K-1)^2}{K} \left[\frac{1}{K-1} \sum_{k=1}^{K} \left(\theta_{i(k)} - \theta_{i(\bullet)}\right) \left(\theta_{i'(k)} - \theta_{i'(\bullet)}\right) \right] \tag{6.32}$$

[For either bootstrap or jackknife, if $i = i'$, the equations yield the corresponding variance estimates.]

Note the subtle difference in usage of ellipses and parentheses between Eqn. (6.29) and Eqn. (6.32). In the former, the subscript $\{c\}$ denotes a set of K cases while in the latter, (k) denotes the original case set with case k removed, leaving $K-1$ cases. There is a similar subtle difference in usage of ellipses and parentheses between Eqn. (6.31) and Eqn. (6.32). The subscript enclosed in parenthesis, i.e., (k), denotes the FOM with case k removed, while in the bootstrap equation one uses the ellipses (curly brackets) $\{b\}$ to denote the b^{th} bootstrap *case-set*, i.e., a whole set of K_1 non-diseased and K_2 diseased cases, sampled with replacement from the original dataset.

The index k ranges from 1 to K, where the first K_1 values represent non-diseased cases and the following K_2 values represent diseased cases. Jackknife figure of merit values, such as $\theta_{i(k)}$, are not to be confused with jackknife pseudovalues used in the DBM chapters. The jackknife FOM corresponding to a particular case is the FOM with the particular case removed while the pseudovalue is K times the FOM with all cases include minus $(K-1)$ times the jackknife FOM. Unlike pseudovalues, jackknife FOM values cannot be regarded as independent and identically distributed, even when using the empirical AUC as FOM.

6.9.5 The variance inflation factor

In Eqn. (6.32), the expression for the jackknife covariance estimate contains a *variance inflation factor*:

$$\frac{(K-1)^2}{K} \tag{6.33}$$

This factor multiplies the traditional expression for the covariance, shown in square brackets in Eqn. (6.32). It is only needed for the jackknife estimate. The bootstrap and the DeLong estimate, see next, do not require this factor.

A third method of estimating the covariance (DeLong et al., 1988), only applicable to the empirical AUC, is not discussed here; however, it is implemented in the software.

6.9.6 Meaning of the covariance matrix

With reference to Eqn. (6.5), suppose one has the luxury of repeatedly sampling case-sets, each consisting of K cases from the population. A single radiologist interprets these cases in I treatments. Therefore, each case-set $\{c\}$ yields I figures of merit. The final numbers at ones disposal are $\theta_{i\{c\}}$, where $i = 1,2,\ldots,I$ and $c = 1,2,\ldots,C$. Considering treatment i, the variance of the FOM-values for the different case-sets $c = 1,2,\ldots,C$, is an estimate of Var_i for this treatment:

$$\sigma_i^2 \equiv Var_i = \frac{1}{C-1}\sum_{c=1}^{C}\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right) \tag{6.34}$$

The process is repeated for all treatments and the I-variance values are averaged. This is the final estimate of Var appearing in Eqn. (6.3).

To estimate the covariance matrix one considers pairs of FOM values for the same case-set $\{c\}$ but different treatments, i.e., $\theta_{i\{c\}}$ and $\theta_{i'\{c\}}$; *by definition primed and un-primed indices are different*. The process is repeated for different case-sets. The covariance is calculated as follows:

$$\mathrm{Cov}_{1;ii'} = \frac{1}{C-1}\sum_{c=1}^{C}\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i'\{c\}} - \theta_{i'\{\bullet\}}\right) \tag{6.35}$$

The process is repeated for all combinations of different-treatment pairings and the resulting $I(I-1)/2$ values are averaged yielding the final estimate of Cov_1. [Recall that the Obuchowski-Rockette model does not allow treatment-dependent parameters in the covariance matrix - hence the need to average over all treatment pairings.]

Since they are derived from the same case-set, one expects the $\theta_{i\{c\}}$ and $\theta_{i'\{c\}}$ values to be correlated. As an example, for a particularly easy *case-set* one expects $\theta_{i\{c\}}$ and $\theta_{i'\{c\}}$ to be both higher than usual. The correlation $\rho_{1;ii'}$ is defined by:

$$\rho_{1;ii'} = \frac{1}{C-1}\sum_{c=1}^{C}\frac{\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i'\{c\}} - \theta_{i'\{\bullet\}}\right)}{\sigma_i\sigma_{i'}} \tag{6.36}$$

Averaging over all different-treatment pairings yields the final estimate of the correlation ρ_1. Since the covariance is smaller than the variance, the magnitude of the correlation is smaller than 1. In most situations one expects ρ_1 to be positive. There is a scenario that could lead to negative correlation. With "complementary" treatments, e.g., CT vs. MRI, where one treatment is good for bone imaging and the other for soft-tissue imaging, an easy case-set in one treatment could correspond to a difficult case-set in the other treatment, leading to negative correlation.

To summarize, the covariance matrix can be estimated using the jackknife or the bootstrap, or, in the special case of the empirical AUC figure of merit, the DeLong method can be used. In (book) Chapter 07, these three methods were described in the context of estimating the *variance* of AUC. Eqn. (6.31) and Eqn. (6.32) extend the jackknife and the bootstrap methods, respectively, to estimating the *covariance* of AUC (whose diagonal elements are the variances estimated in the earlier chapter).

6.9.7 Code illustrating the covariance matrix

To minimize clutter, the `R` functions (for estimating `Var` and `Cov1` using bootstrap, jackknife, and the DeLong methods) are not shown, but they are compiled. To display them `clone` or 'fork' the book repository and look at the `Rmd` file corresponding to this output and the sourced `R` files listed below:

The following code chunk extracts (using the `DfExtractDataset` function) a single-reader multiple-treatment ROC dataset corresponding to the first reader from `dataset02`, which is the Van Dyke dataset.

```
rocData1R <- DfExtractDataset(dataset02, rdrs = 1) #select the 1st reader to be analyz
zik1 <- rocData1R$ratings$NL[,1,,1];K <- dim(zik1)[2];I <- dim(zik1)[1]
zik2 <- rocData1R$ratings$LL[,1,,1];K2 <- dim(zik2)[2];K1 <- K-K2;zik1 <- zik1[,1:K1]
```

The following notation is used in the code below:

- jk = jackknife method
- bs = bootstrap method, with B = number of bootstraps and `seed` = value.
- dl = DeLong method
- rj_jk = RJafroc, `covEstMethod` = "jackknife"
- rj_bs = RJafroc, `covEstMethod` = "bootstrap"

For example, `Cov1_jk` is the jackknife estimate of `Cov1`. Shown below are the results of the jackknife method, first using the code in this repository and next, as a cross-check, using `RJafroc` function `UtilORVarComponentsFactorial`:

```
ret1 <- VarCov1_Jk(zik1, zik2)
Var <- ret1$Var
Cov1 <- ret1$Cov1 # use these (i.e., jackknife) as default values in subsequent code
data.frame ("Cov1_jk" = Cov1, "Var_jk" = Var)
```

```
##        Cov1_jk       Var_jk
## 1 0.0003734661 0.0006989006
```

```
ret4 <- UtilORVarComponentsFactorial(
  rocData1R, FOM = "Wilcoxon") # the functions default `covEstMethod` is jackknife
data.frame ("Cov1_rj_jk" = ret4$VarCom["Cov1", "Estimates"],
            "Var_rj_jk" = ret4$VarCom["Var", "Estimates"])
```

```
##     Cov1_rj_jk    Var_rj_jk
## 1 0.0003734661 0.0006989006
```

Note that the estimates are identical and that the Cov1 estimate is smaller than the Var estimate (their ratio is the correlation $\rho_1 = \text{Cov1}/\text{Var} = 0.5343623$).

Shown next are bootstrap method estimates with increasing number of bootstraps (200, 2000 and 20,000):

```
ret2 <- VarCov1_Bs(zik1, zik2, 200, seed = 100)
data.frame ("Cov_bs" = ret2$Cov1, "Var_bs" = ret2$Var)
```

```
##        Cov_bs       Var_bs
## 1 0.000283905 0.0005845354
```

```
ret2 <- VarCov1_Bs(zik1, zik2, 2000, seed = 100)
data.frame ("Cov_bs" = ret2$Cov1, "Var_bs" = ret2$Var)
```

```
##         Cov_bs       Var_bs
## 1 0.0003466804 0.0006738506
```

```
ret2 <- VarCov1_Bs(zik1, zik2, 20000, seed = 100)
data.frame ("Cov_bs" = ret2$Cov1, "Var_bs" = ret2$Var)
```

```
##         Cov_bs       Var_bs
## 1 0.0003680714 0.0006862668
```

With increasing number of bootstraps the values approach the jackknife estimates.

Following, as a cross check, are results of bootstrap method as calculated by the `RJafroc` function `UtilORVarComponentsFactorial`:

```
ret5 <- UtilORVarComponentsFactorial(
  rocData1R, FOM = "Wilcoxon",
  covEstMethod = "bootstrap", nBoots = 2000, seed = 100)
data.frame ("Cov_rj_bs" = ret5$VarCom["Cov1", "Estimates"],
            "Var_rj_bs" = ret5$VarCom["Var", "Estimates"])
```

```
##      Cov_rj_bs    Var_rj_bs
## 1 0.0003466804 0.0006738506
```

Note that the two estimates shown above for $B = 2000$ are identical. This is because *the seeds are identical*. With different seeds on expect sampling related fluctuations.

Following are results of the DeLong covariance estimation method, the first output is using this repository code and the second using the `RJafroc` function `UtilORVarComponentsFactorial` with appropriate arguments:

```
mtrxDLStr <- VarCovMtrxDLStr(rocData1R)
ret3 <- VarCovs(mtrxDLStr)
data.frame ("Cov_dl" = ret3$cov1, "Var_dl" = ret3$var)
```

```
##         Cov_dl       Var_dl
## 1 0.0003684357 0.0006900766
```

```
ret5 <- UtilORVarComponentsFactorial(
  rocData1R, FOM = "Wilcoxon", covEstMethod = "DeLong")
data.frame ("Cov_rj_dl" = ret5$VarCom["Cov1", "Estimates"],
            "Var_rj_dl" = ret5$VarCom["Var", "Estimates"])
```

```
##      Cov_rj_dl    Var_rj_dl
## 1 0.0003684357 0.0006900766
```

Note that the two estimates are identical and that the DeLong estimate are close to the bootstrap estimates using 20,000 bootstraps. The just demonstrated close correspondence is only expected when using the Wilcoxon figure of merit, i.e., the empirical AUC.

6.9.8 Comparing DBM to Obuchowski and Rockette for single-reader multiple-treatments

We have shown two methods for analyzing a single reader in multiple treatments: the DBM method, involving jackknife derived pseudovalues and the Obuchowski and Rockette method that does not have to use the jackknife, since it could use the bootstrap, or the DeLong method, if one restricts to the Wilcoxon statistic for the figure of merit, to get the covariance matrix. Since one is dealing with a single reader in multiple treatments, for DBM one needs the fixed-reader random-case analysis described in TBA §9.8 of the previous chapter (it should be obvious that with one reader the conclusions apply to the specific reader only, so reader must be regarded as a fixed factor).

Shown below are results obtained using `RJafroc` function `StSignificanceTesting` with `analysisOption = "FRRC"` for `method` = "DBM" (which uses the jackknife), and for OR using 3 different ways of estimating the covariance matrix for the one-reader analysis (i.e., Cov1 and Var).

```
ret1 <- StSignificanceTesting(
  rocData1R,FOM = "Wilcoxon", method = "DBM", analysisOption = "FRRC")
data.frame("DBM:F" = ret1$FRRC$FTests["Treatment", "FStat"],
           "DBM:ddf" = ret1$FRRC$FTests["Treatment", "DF"],
           "DBM:P-val" = ret1$FRRC$FTests["Treatment", "p"])
```

```
##       DBM.F DBM.ddf  DBM.P.val
## 1 1.2201111       1 0.27168532
```

```
ret2 <- StSignificanceTesting(
  rocData1R,FOM = "Wilcoxon", method = "OR", analysisOption = "FRRC")
data.frame("ORJack:Chisq" = ret2$FRRC$FTests["Treatment", "Chisq"],
           "ORJack:ddf" = ret2$FRRC$FTests["Treatment", "DF"],
           "ORJack:P-val" = ret2$FRRC$FTests["Treatment", "p"])
```

```
##   ORJack.Chisq ORJack.ddf ORJack.P.val
## 1    1.2201111          1   0.26933885
```

```
ret3 <- StSignificanceTesting(
  rocData1R,FOM = "Wilcoxon", method = "OR", analysisOption = "FRRC",
                              covEstMethod = "DeLong")
data.frame("ORDeLong:Chisq" = ret3$FRRC$FTests["Treatment", "Chisq"],
           "ORDeLong:ddf" = ret3$FRRC$FTests["Treatment", "DF"],
           "ORDeLong:P-val" = ret3$FRRC$FTests["Treatment", "p"])
```

```
##   ORDeLong.Chisq ORDeLong.ddf ORDeLong.P.val
## 1      1.2345017            1     0.26653335
```

```
ret4 <- StSignificanceTesting(
  rocData1R,FOM = "Wilcoxon", method = "OR", analysisOption = "FRRC",
                              covEstMethod = "bootstrap")
data.frame("ORBoot:Chisq" = ret4$FRRC$FTests["Treatment", "Chisq"],
           "ORBoot:ddf" = ret4$FRRC$FTests["Treatment", "DF"],
           "ORBoot:P-val" = ret4$FRRC$FTests["Treatment", "p"])
```

```
##   ORBoot.Chisq ORBoot.ddf ORBoot.P.val
## 1    1.2914763          1   0.25577607
```

The DBM and OR-jackknife methods yield identical F-statistics, but the denominator degrees of freedom are different, $(I-1)(K-1) = 113$ for DBM and ∞ for OR. The F-statistics for OR-bootstrap and OR-DeLong are different.

Shown below is a first-principles implementation of OR significance testing for the one-reader case.

```
alpha <- 0.05
theta_i <- c(0,0);for (i in 1:I) theta_i[i] <- Wilcoxon(zik1[i,], zik2[i,])

MS_T <- 0
for (i in 1:I) {
  MS_T <- MS_T + (theta_i[i]-mean(theta_i))^2
}
MS_T <- MS_T/(I-1)

F_1R <- MS_T/(Var - Cov1)
pValue <- 1 - pf(F_1R, I-1, Inf)

trtDiff <- array(dim = c(I,I))
for (i1 in 1:(I-1)) {
  for (i2 in (i1+1):I) {
    trtDiff[i1,i2] <- theta_i[i1]- theta_i[i2]
  }
}
trtDiff <- trtDiff[!is.na(trtDiff)]
nDiffs <- I*(I-1)/2
CI_DIFF_FOM_1RMT <- array(dim = c(nDiffs, 3))
for (i in 1 : nDiffs) {
  CI_DIFF_FOM_1RMT[i,1] <- trtDiff[i] + qt(alpha/2,  df = Inf)*sqrt(2*(Var - Cov1))
  CI_DIFF_FOM_1RMT[i,2] <- trtDiff[i]
  CI_DIFF_FOM_1RMT[i,3] <- trtDiff[i] + qt(1-alpha/2,df = Inf)*sqrt(2*(Var - Cov1))
  print(data.frame("theta_1" = theta_i[1],
                   "theta_2" = theta_i[2],
                   "Var" = Var,
```

```
                    "Cov1" = Cov1,
                    "MS_T" = MS_T,
                    "F_1R" = F_1R,
                    "pValue" = pValue,
                    "Lower" = CI_DIFF_FOM_1RMT[i,1],
                    "Mid" = CI_DIFF_FOM_1RMT[i,2],
                    "Upper" = CI_DIFF_FOM_1RMT[i,3]))
}
```

```
##      theta_1    theta_2           Var          Cov1           MS_T      F_1R
## 1 0.91964573 0.94782609 0.00069890056 0.0003734661 0.00039706618 1.2201111
##       pValue        Lower          Mid       Upper
## 1 0.26933885 -0.078183215 -0.028180354 0.021822507
```

The following shows the corresponding output of `RJafroc`.

```
ret_rj <- StSignificanceTesting(
  rocData1R, FOM = "Wilcoxon", method = "OR", analysisOption = "FRRC")
print(data.frame("theta_1" = ret_rj$FOMs$foms[1,1],
                 "theta_2" = ret_rj$FOMs$foms[2,1],
                 "Var" = ret_rj$ANOVA$VarCom["Var", "Estimates"],
                 "Cov1" = ret_rj$ANOVA$VarCom["Cov1", "Estimates"],
                 "MS_T" = ret_rj$ANOVA$TRanova[1,3],
                 "Chisq_1R" = ret_rj$FRRC$FTests["Treatment","Chisq"],
                 "pValue" = ret_rj$FRRC$FTests["Treatment","p"],
                 "Lower" = ret_rj$FRRC$ciDiffTrt[1,"CILower"],
                 "Mid" = ret_rj$FRRC$ciDiffTrt[1,"Estimate"],
                 "Upper" = ret_rj$FRRC$ciDiffTrt[1,"CIUpper"]))
```

```
##      theta_1    theta_2           Var          Cov1           MS_T  Chisq_1R
## 1 0.91964573 0.94782609 0.00069890056 0.0003734661 0.00039706618 1.2201111
##       pValue        Lower          Mid       Upper
## 1 0.26933885 -0.078183215 -0.028180354 0.021822507
```

The first-principles and the `RJafroc` values agree exactly with each other [for $I = 2$, the F and chisquare statistics are identical]. This above code also shows how to extract the different estimates (Var, Cov1, etc.) from the object `ret_rj` returned by `RJafroc`. Specifically,

- Var: ret_rj$ANOVA$VarCom[“Var”, “Estimates”]
- Cov1: ret_rj$ANOVA$VarCom[“Cov1”, “Estimates”]
- Chisquare-statistic: ret_rj$FRRC$FTests[“Treatment”,“Chisq”]
- df: ret_rj$FRRC$FTests[1,“DF”]

- p-value: ret_rj$FRRC$FTests[“Treatment”,“p”]
- CI Lower: ret_rj$FRRC$ciDiffTrt[1,“CILower”]
- Mid Value: ret_rj$FRRC$ciDiffTrt[1,“Estimate”]
- CI Upper: ret_rj$FRRC$ciDiffTrt[1,“CIUpper”]

6.9.8.1 Jumping ahead

If RRRC analysis were conducted, the values are [one needs to analyze a dataset like `dataset02` having more than one treatments and readers and use `analysisOption` = “RRRC”]:

- msR: ret_rj$ANOVA$TRanova[“R”, “MS”]
- msT: ret_rj$ANOVA$TRanova[“T”, “MS”]
- msTR: ret_rj$ANOVA$TRanova[“TR”, “MS”]
- Var: ret_rj$ANOVA$VarCom[“Var”, “Estimates”]
- Cov1: ret_rj$ANOVA$VarCom[“Cov1”, “Estimates”]
- Cov2: ret_rj$ANOVA$VarCom[“Cov2”, “Estimates”]
- Cov3: ret_rj$ANOVA$VarCom[“Cov3”, “Estimates”]
- varR: ret_rj$ANOVA$VarCom[“VarR”, “Estimates”]
- varTR: ret_rj$ANOVA$VarCom[“VarTR”, “Estimates”]
- F-statistic: ret_rj$RRRC$FTests[“Treatment”, “FStat”]
- ddf: ret_rj$RRRC$FTests[“Error”, “DF”]
- p-value: ret_rj$RRRC$FTests[“Treatment”, “p”]
- CI Lower: ret_rj$RRRC$ciDiffTrt[“trt0-trt1”,“CILower”]
- Mid Value: ret_rj$RRRC$ciDiffTrt[“trt0-trt1”,“Estimate”]
- CI Upper: ret_rj$RRRC$ciDiffTrt[“trt0-trt1”,“CIUpper”]

For `RRFC` analysis, one replaces `RRRC` with `RRFC`, etc. I should note that the auto-prompt feature of `RStudio` makes it unnecessary to enter the complex string names shown above - `RStudio` will suggest them.

6.10 Chapter References

Chapter 7

Obuchowski Rockette (OR) Analysis

7.1 TBA How much finished

80%

7.2 Introduction

In previous chapters the DBM significance testing procedure (Dorfman et al., 1992) for analyzing MRMC ROC data, along with improvements (Hillis, 2014), has been described. Because the method assumes that jackknife pseudovalues can be regarded as independent and identically distributed case-level figures of merit, it has been rightly criticized by Hillis and others (Zhou et al., 2009). Hillis states that the method "works" but lacks firm statistical foundations (Hillis et al., 2005; Hillis, 2007; Hillis et al., 2008). I would add that it "works" as long as one restricts to the empirical AUC figure of merit. In my book I gave a justification for why the method "works". Specifically, the *empirical AUC pseudovalues qualify as case-level FOMs* - this property has also been noted by (Hajian-Tilaki et al., 1997). However, this property applies *only* to the empirical AUC, so an alternate approach that applies to any figure of merit is highly desirable.

Hillis' has proposed that a method based on an earlier publication (Obuchowski and Rockette, 1995), which does not depend on pseudovalues, is preferable from both conceptual and practical points of view. This chapter is named "OR Analysis", where OR stands for Obuchowski and Rockette. The OR method has advantages in being able to handle more complex study designs (Hillis, 2014)

that are addressed in subsequent chapters, and applications to other FOMs (e.g., the FROC paradigm uses a rather different FOM from empirical ROC-AUC) are best performed with the OR method.

This chapter delves into the significance testing procedure employed in OR analysis.

Multiple readers interpreting a case-set in multiple treatments is analyzed and the results, DBM vs. OR, are compared for the same dataset. The special cases of fixed-reader and fixed-case analyses are described. Single treatment analysis, where interest is in comparing average performance of readers to a fixed value, is described. Three methods of estimating the covariance matrix are described.

Before proceeding, it is understood that datasets analyzed in this chapter follow a *factorial* design, sometimes call fully-factorial or fully-crossed design. Basically, the data structure is symmetric, e.g., all readers interpret all cases in all modalities. The next chapter will describe the analysis of *split-plot* datasets, where, for example, some readers interpret all cases in one modality, while the remaining readers interpret all cases in the other modality.

7.3 Random-reader random-case

In conventional ANOVA models, such as used in DBM, the covariance matrix of the error term is diagonal with all diagonal elements equal to a common variance, represented in the DBM model by the scalar ϵ term. Because of the correlated structure of the error term, in OR analysis, a customized ANOVA is needed. The null hypothesis (NH) is that the true figures-of-merit of all treatments are identical, i.e.,

$$NH : \tau_i = 0 \;\; (i = 1, 2, ..., I) \tag{7.1}$$

The analysis described next considers both readers and cases as random effects. The F-statistic is denoted F_{ORH}, defined by:

$$F_{ORH} = \frac{MS(T)}{MS(TR) + J \max(\text{Cov2} - \text{Cov3}, 0)} \tag{7.2}$$

Eqn. (7.2) incorporates Hillis' modification of the original OR F-statistic. The modification ensures that the constraint Eqn. (6.23) is always obeyed and also avoids a possibly negative (and hence illegal) F-statistic. The relevant mean squares are defined by (note that these are calculated using *FOM* values, not *pseudovalues*):

$$\left.\begin{aligned} MS(T) &= \tfrac{J}{I-1}\sum_{i=1}^{I}(\theta_{i\bullet}-\theta_{\bullet\bullet})^2 \\ MS(R) &= \tfrac{I}{J-1}\sum_{j=1}^{J}(\theta_{\bullet j}-\theta_{\bullet\bullet})^2 \\ MS(TR) &= \tfrac{1}{(I-1)(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}(\theta_{ij}-\theta_{i\bullet}-\theta_{\bullet j}+\theta_{\bullet\bullet}) \end{aligned}\right\} \tag{7.3}$$

The original paper (Obuchowski and Rockette, 1995) actually proposed a different test statistic F_{OR}:

$$F_{OR} = \frac{MS(T)}{MS(TR) + J(\mathrm{Cov2} - \mathrm{Cov3})} \tag{7.4}$$

Note that Eqn. (7.4) lacks the constraint, subsequently proposed by Hillis, which ensures that the denominator cannot be negative. The following distribution was proposed for the test statistic.

$$F_{OR} \sim F_{\mathrm{ndf,ddf}} \tag{7.5}$$

The original degrees of freedom were defined by:

$$\begin{aligned} \mathrm{ndf} &= I - 1 \\ \mathrm{ddf} &= (I-1)\times(J-1) \end{aligned} \tag{7.6}$$

It turns out that the Obuchowski-Rockette test statistic is very conservative, meaning it is highly biased against rejecting the null hypothesis (the data simulator used in the validation described in their publication did not detect this behavior). Because of the conservative behavior, the predicted sample sizes tended to be quite large (if the test statistic does not reject the NH as often as it should, one way to overcome this tendency is to use a larger sample size). In this connection I have two informative anecdotes.

7.3.1 Two anecdotes

- The late Dr. Robert F. Wagner once stated to me (ca. 2001) that the sample-size tables published by Obuchowski (Obuchowski, 1998, 2000), using the version of Eqn. (7.2) with the *ddf* as originally suggested by Obuchowski and Rockette, predicted such high number of readers and cases that he was doubtful about the chances of anyone conducting a practical ROC study!

- The second story is that I once conducted NH simulations and analyses using a Roe-Metz simulator (Roe and Metz, 1997b) and the significance testing described in the Obuchowski-Rockette paper: the method did not reject the null hypothesis even once in 2000 trials! Recall that with $\alpha = 0.05$ a valid test should reject the null hypothesis about 100 ± 20 times in 2000 trials. I recalls (ca. 2004) telling Dr. Steve Hillis about this issue, and he suggested a different denominator degrees of freedom *ddf*, see next, substitution of which magically solved the problem, i.e., the simulations rejected the null hypothesis 5% of the time.

7.3.2 Hillis ddf

Hillis' proposed new *ddf* is shown below (*ndf* is unchanged), with the subscript H denoting the Hillis modification:

$$\text{ddf}_H = \frac{[MS(TR) + J \max(\text{Cov2} - \text{Cov3}, 0)]^2}{\frac{[MS(TR)]^2}{(I-1)(J-1)}} \tag{7.7}$$

From the previous chapter, the ordering of the covariances is as follows:

$$\text{Cov3} \leq \text{Cov2} \leq \text{Cov1} \leq \text{Var}$$

If Cov2 $<$ Cov3 (which is the *exact opposite* of the expected ordering), ddf_H reduces to $(I-1) \times (J-1)$, the value originally proposed by Obuchowski and Rockette. With Hillis' proposed changes, under the null hypothesis the observed statistic F_{ORH}, defined in Eqn. (7.2), is distributed as an F-statistic with ndf $= I-1$ and *ddf* $= \text{ddf}_H$ degrees of freedom (Hillis et al., 2005; Hillis, 2007; Hillis et al., 2008):

$$F_{ORH} \sim F_{\text{ndf},\text{ddf}_H} \tag{7.8}$$

If the expected ordering is true, i.e., Cov2 $>$ Cov3 , which is the more likely situation, then ddf_H is *larger* than $(I-1) \times (J-1)$, i.e., the Obuchowski-Rockette *ddf*, and the p-value decreases and there is a larger probability of rejecting the NH. The modified OR method is more likely to have the correct NH behavior, i.e, it will reject the NH 5% of the time when alpha is set to 0.05 (statisticians refer to this as "passing the 5% test"). The correct NH behavior has been confirmed in simulation testing using the Roe-Metz simulator (Hillis et al. (2008)).

7.3.3 Decision rule, p-value and confidence interval

The critical value of the F-statistic for rejection of the null hypothesis is $F_{1-\alpha,\text{ndf},\text{ddf}_H}$, i.e., that value such that fraction $(1-\alpha)$ of the area under the distribution lies to the left of the critical value. From Eqn. (7.2):

- Rejection of the NH is more likely if $MS(T)$ increases, meaning the treatment effect is larger;
- $MS(TR)$ is smaller, meaning there is less contamination of the treatment effect by treatment-reader variability;
- The greater of Cov2 or Cov3, which is usually Cov2, decreases, meaning there is less "noise" in the measurement due to between-reader variability. Recall that Cov2 involves different-reader same-treatment pairings.

- α increases, meaning one is allowing a greater probability of Type I errors;
- ndf increases, as this lowers the critical value of the F-statistic. With more treatment pairings, the chance that at least one paired-difference will reject the NH is larger.
- ddf_H increases, as this lowers the critical value of the F-statistic.

The p-value of the test is the probability, under the NH, that an equal or larger value of the F-statistic than F_{ORH} could be observed by chance. In other words, it is the area under the F-distribution $F_{\text{ndf},\text{ddf}_H}$ that lies above the observed value F_{ORH}:

$$p = \Pr(F > F_{ORH} \mid F \sim F_{\text{ndf},\text{ddf}_H}) \tag{7.9}$$

The $(1-\alpha)$ confidence interval for $\theta_{i\bullet} - \theta_{i'\bullet}$ is given by:

$$\begin{aligned} CI_{1-\alpha,RRRC,\theta_{i\bullet}-\theta_{i'\bullet}} =& \theta_{i\bullet} - \theta_{i'\bullet} \\ & \pm t_{\alpha/2,\text{ddf}_H} \sqrt{\tfrac{2}{J}(MS(TR) + J\max(\text{Cov2} - \text{Cov3}, 0))} \end{aligned} \tag{7.10}$$

Define df_i, the degrees of freedom for modality i:

$$\text{df}_i = (\text{MS(R)}_i + J\max(\text{Cov2}_i, 0))^2 / \text{MS(R)}_i^2 * (J-1) \tag{7.11}$$

Here MS(R)_i is the reader mean-square for modality i, and Cov2_i is Cov2 for modality i. Note that all quantities with an i index are calculated using data from modality i only.

The $(1-\alpha)$ confidence interval for $\theta_{i\bullet}$, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}}$, is given by:

$$CI_{1-\alpha,RRRC,\theta_{i\bullet}} = \theta_{i\bullet} \pm t_{\alpha/2,\mathrm{df}_i} \sqrt{\tfrac{1}{J}(\mathrm{MS(R)}_i + J\max(\mathrm{Cov2}_i, 0))} \tag{7.12}$$

7.4 Fixed-reader random-case

Using the vertical bar notation $\mid R$ to denote that reader is regarded as a fixed effect (Roe and Metz, 1997a), the F -statistic for testing the null hypothesis $NH : \tau_i = 0$ $(i = 1, 1, 2, ...I)$ is (Hillis, 2007):

$$F_{ORH|R} = \frac{MS(T)}{\mathrm{Var} - \mathrm{Cov1} + (J-1)\max(\mathrm{Cov2} - \mathrm{Cov3}, 0)} \tag{7.13}$$

[For $J = 1$, Eqn. (7.13) reduces to Eqn. (6.8), i.e., the single-reader analysis described in the previous chapter.]

$F_{ORH|R}$ is distributed as an F-statistic with ndf $= I - 1$ and ddf $= \infty$:

$$F_{ORH|R} \sim F_{I-1,\infty} \tag{7.14}$$

One can get rid of the infinite denominator degrees of freedom by recognizing, as in the previous chapter, that $(I-1)F_{I-1,\infty}$ is distributed as a χ^2 distribution with $I-1$ degrees of freedom, i.e., as χ^2_{I-1}. Therefore, one has, analogous to Eqn. (6.7),

$$\chi^2_{ORH|R} \equiv (I-1)F_{ORH|R} \sim \chi^2_{I-1} \tag{7.15}$$

The critical value of the χ^2 statistic is $\chi^2_{1-\alpha,I-1}$, which is that value such that fraction $(1-\alpha)$ of the area under the χ^2_{I-1} distribution lies to the left of the critical value. The null hypothesis is rejected if the observed value of the χ^2 statistic exceeds the critical value, i.e.,

$$\chi^2_{ORH|R} > \chi^2_{1-\alpha,I-1}$$

The p-value of the test is the probability that a random sample from the chi-square distribution χ^2_{I-1} exceeds the observed value of the test statistic $\chi^2_{ORH|R}$ statistic defined in Eqn. (7.15):

$$p = \Pr(\chi^2 > \chi^2_{ORH|R} \mid \chi^2 \sim \chi^2_{I-1}) \tag{7.16}$$

The $(1-\alpha)$ (symmetric) confidence interval for the difference figure of merit is given by:

$$CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,\infty}\sqrt{\tfrac{2}{J}(\text{Var} - \text{Cov1} + (J-1)\max(\text{Cov2} - \text{Cov3}, 0))} \tag{7.17}$$

The NH is rejected if any of the following equivalent conditions is met (these statements are also true for `RRRC` analysis, and `RRFC` analysis to be described next):

- The observed value of the χ^2 statistic exceeds the critical value $\chi^2_{1-\alpha,I-1}$.
- The p-value is less than α.
- The $(1-\alpha)$ confidence interval for at least one treatment-pairing does not include zero.

Additional confidence intervals are stated below:

- The confidence interval for the reader-averaged FOM for each treatment, denoted $CI_{1-\alpha,FRRC,\theta_{i\bullet}}$.
- The confidence interval for treatment FOM differences for each reader, denoted $CI_{1-\alpha,FRRC,\theta_{ij}-\theta_{i'j}}$.

$$CI_{1-\alpha,FRRC,\theta_{i\bullet}} = \theta_{i\bullet} \pm z_{\alpha/2}\sqrt{\tfrac{1}{J}(\text{Var}_i + (J-1)\max(\text{Cov2}_i, 0)} \tag{7.18}$$

$$CI_{1-\alpha,FRRC,\theta_{ij}-\theta_{i'j}} = (\theta_{ij} - \theta_{i'j}) \pm z_{\alpha/2}\sqrt{2(\text{Var}_j - \text{Cov1}_j)} \tag{7.19}$$

In these equations Var_i and Cov2_i are computed using the data for treatment i only, and Var_j and Cov1_j are computed using the data for reader j only.

7.5 Random-reader fixed-case

When case is treated as a fixed factor, the appropriate F-statistic for testing the null hypothesis $NH : \tau_i = 0$ $(i = 1, 1, 2, ...I)$ is:

$$F_{ORH|C} = \frac{MS(T)}{MS(TR)} \tag{7.20}$$

$F_{ORH|C}$ is distributed as an F-statistic with $ndf = I-1$ and $ddf = (I-1)(J-1)$:

$$\left.\begin{array}{rcl} \text{ndf} & = & I-1 \\ \text{ddf} & = & (I-1)(J-1) \\ F_{ORH|C} & \sim & F_{\text{ndf,ddf}} \end{array}\right\} \tag{7.21}$$

Here is a situation where the degrees of freedom agree with those originally proposed by Obuchowski-Rockette. The critical value of the statistic is $F_{1-\alpha,I-1,(I-1)(J-1)}$, which is that value such that fraction $(1-\alpha)$ of the distribution lies to the left of the critical value. The null hypothesis is rejected if the observed value of the F statistic exceeds the critical value:

$$F_{ORH|C} > F_{1-\alpha,I-1,(I-1)(J-1)}$$

The p-value of the test is the probability that a random sample from the distribution exceeds the observed value:

$$p = \Pr(F > F_{ORH|C} \mid F \sim F_{1-\alpha,I-1,(I-1)(J-1)})$$

The $(1-\alpha)$ confidence interval for the reader-averaged difference FOM, $CI_{1-\alpha,RRFC,\theta_{i\bullet}-\theta_{i'\bullet}}$, is given by:

$$CI_{1-\alpha,RRFC,\theta_{i\bullet}-\theta_{i'\bullet}} = (\theta_{i\bullet}-\theta_{i'\bullet}) \pm t_{\alpha/2,(I-1)(J-1)}\sqrt{\tfrac{2}{J}MS(TR)} \tag{7.22}$$

The $(1-\alpha)$ confidence interval for the reader-averaged FOM for each treatment, $CI_{1-\alpha,RRFC,\theta_{i\bullet}}$, is given by:

$$CI_{1-\alpha,RRFC,\theta_{i\bullet}} = \theta_{i\bullet} \pm t_{\alpha/2,J-1}\sqrt{\tfrac{1}{J}\text{MS(R)}_i} \tag{7.23}$$

Here MS(R)_i is the reader mean-square for modality i.

7.6 Single treatment analysis

TBA ## Summary{#or-analysis-st-summary} ## Discussion{#or-analysis-st-discussion} ## Chapter References {#or-analysis-st-references}

Chapter 8

Obuchowski Rockette Applications

8.1 TBA How much finished

80%

8.2 Introduction

This chapter illustrates Obuchowski-Rockette analysis with several examples. The first example is a full-blown "hand-calculation" for `dataset02`, showing explicit implementations of formulae presented in the previous chapter. The second example shows application of the `RJafroc` package function `StSignificanceTesting()` to the same dataset: this function encapsulates all formulae and accomplishes all analyses with one function call. The third example shows application of the `StSignificanceTesting()` function to an ROC dataset derived from the Federica Zanca dataset (Zanca et al., 2009), which has five modalities and four readers. This illustrates multiple treatment pairings (in contrast, `dataset02` has only one treatment pairing). The fourth example shows application of `StSignificanceTesting()` to `dataset04`, which is an **FROC** dataset (in contrast to the previous examples, which employed **ROC** datasets). It illustrates the key difference involved in FROC analysis, namely the choice of figure of merit. The final example again uses `dataset04`, i.e., FROC data, *but this time we use DBM analysis*. Since DBM analysis is pseudovalue based, and the figure of merit is not the empirical AUC under the ROC, one may expect to see differences from the previously presented OR analysis on the same dataset.

Each analysis involves the following steps:

- Calculate the figure of merit;
- Calculate the variance-covariance matrix and mean-squares;
- Calculate the NH statistic, p-value and confidence interval(s).
- For each analysis, three sub-analyses are shown:
  - random-reader random-case (RRRC),
  - fixed-reader random-case (FRRC), and
  - random-reader fixed-case (RRFC).

8.3 Hand calculation

Dataset `dataset02` is well-know in the literature (Van Dyke et al., 1993) as it has been widely used to illustrate advances in ROC methodology. The following code extract the numbers of modalities, readers and cases for `dataset02` and defines strings `modalityID`, `readerID` and `diffTRName` that are needed for the hand-calculations.

```
I <- length(dataset02$ratings$NL[,1,1,1])
J <- length(dataset02$ratings$NL[1,,1,1])
K <- length(dataset02$ratings$NL[1,1,,1])
modalityID <- dataset02$descriptions$modalityID
readerID <- dataset02$descriptions$readerID
diffTRName <- array(dim = choose(I, 2))
ii <- 1
for (i in 1:I) {
  if (i == I)
    break
  for (ip in (i + 1):I) {
    diffTRName[ii] <-
      paste0("trt", modalityID[i],
             sep = "-", "trt", modalityID[ip])
    ii <- ii + 1
  }
}
```

The dataset consists of I = 2 treatments, J = 5 readers and K = 114 cases.

8.3.1 Random-Reader Random-Case (RRRC) analysis

- The first step is to calculate the figures of merit using `UtilFigureOfMerit()`.
- Note that the `FOM` argument has to be explicitly specified as there is no default.

```
foms <- UtilFigureOfMerit(dataset02, FOM = "Wilcoxon")
print(foms, digits = 4)
#>        rdr0   rdr1   rdr2   rdr3   rdr4
#> trt0 0.9196 0.8588 0.9039 0.9731 0.8298
#> trt1 0.9478 0.9053 0.9217 0.9994 0.9300
```

- For example, for the first treatment, "trt0", the second reader "rdr1" figure of merit is 0.8587762.
- The next step is to calculate the variance-covariance matrix and the mean-squares.
- The function UtilORVarComponentsFactorial() returns these quantities, which are saved to vc.
- The Factorial in the function name is because this code applies to the factorial design. A different function is used for a split-plot design.

```
vc <- UtilORVarComponentsFactorial(
  dataset02, FOM = "Wilcoxon", covEstMethod = "jackknife")
print(vc, digits = 4)
#> $TRanova
#>          SS DF       MS
#> T  0.004796  1 0.004796
#> R  0.015345  4 0.003836
#> TR 0.002204  4 0.000551
#>
#> $VarCom
#>       Estimates   Rhos
#> VarR  0.0015350     NA
#> VarTR 0.0002004     NA
#> Cov1  0.0003466 0.4320
#> Cov2  0.0003441 0.4289
#> Cov3  0.0002390 0.2979
#> Var   0.0008023     NA
#>
#> $IndividualTrt
#>      DF msREachTrt varEachTrt cov2EachTrt
#> trt0  4   0.003083  0.0010141   0.0004840
#> trt1  4   0.001305  0.0005905   0.0002042
#>
#> $IndividualRdr
#>      DF msTEachRdr varEachRdr cov1EachRdr
#> rdr0  1  0.0003971  0.0006989   3.735e-04
#> rdr1  1  0.0010829  0.0011061   7.602e-04
#> rdr2  1  0.0001597  0.0008423   3.553e-04
#> rdr3  1  0.0003445  0.0001506   1.083e-06
#> rdr4  1  0.0050161  0.0012136   2.430e-04
```

- The next step is the calculate the NH testing statistic.
- The relevant equation is Eqn. (7.2).
- `vc` contains the values needed in this equation, as follows:
  - MS(T) is in `vc$TRanova["T", "MS"]`, whose value is 0.0047962.
  - MS(TR) is in `vc$TRanova["TR", "MS"]`, whose value is 5.5103062×10^{-4}.
  - `Cov2` is in `vc$VarCom["Cov2", "Estimates"]`, whose value is 3.4407483×10^{-4}.
  - `Cov3` is in `vc$VarCom["Cov3", "Estimates"]`, whose value is 2.3902837×10^{-4}.

Applying Eqn. (7.2) one gets (`den` is the denominator on the right hand side of the referenced equation) and F_ORH_RRRC is the value of the F-statistic:

```
den <- vc$TRanova["TR", "MS"] +
  J* max(vc$VarCom["Cov2", "Estimates"] -
           vc$VarCom["Cov3", "Estimates"],0)
F_ORH_RRRC <- vc$TRanova["T", "MS"]/den
print(F_ORH_RRRC, digits = 4)
#> [1] 4.456
```

- The F-statistic has numerator degrees of freedom ndf $= I - 1$ and denominator degrees of freedom, `ddf`, to be calculated next.
- From the previous chapter, `ddf` is calculated using Eqn. (7.7)). The numerator of `ddf` is identical to `den^2`, where `den` was calculated in the preceding code block. The implementation follows:

```
ddf <- den^2*(I-1)*(J-1)/(vc$TRanova["TR", "MS"])^2
print(ddf, digits = 4)
#> [1] 15.26
```

- The next step is calculation of the p-value for rejecting the NH
- The relevant equation is Eqn. (7.9) whose implementation follows:

```
p <- 1 - pf(F_ORH_RRRC, I - 1, ddf)
print(p, digits = 4)
#> [1] 0.05167
```

- The difference is not significant at $\alpha = 0.05$.
- The next step is to calculate confidence intervals.
- Since `I` $= 2$, their is only one paired difference in reader-averaged FOMs, namely, the first treatment minus the second.

```
trtMeans <- rowMeans(foms)
trtMeanDiffs <- trtMeans[1] - trtMeans[2]
names(trtMeanDiffs) <- "trt0-trt1"
print(trtMeans, digits = 4)
#>   trt0   trt1
#> 0.8970 0.9408
print(trtMeanDiffs, digits = 4)
#> trt0-trt1
#>   -0.0438
```

- `trtMeans`contains the reader-averaged figures of merit for each treatment.
- `trtMeanDiffs`contains the reader-averaged difference figure of merit.
- From the previous chapter, the $(1-\alpha)$ confidence interval for $\theta_{1\bullet} - \theta_{2\bullet}$ is given by Eqn. (7.10), in which equation the expression inside the square-root symbol is `2/J*den`.
- α, the significance level of the test, is set to 0.05.
- The implementation follows:

```
alpha <- 0.05
stdErr <- sqrt(2/J*den)
t_crit <- abs(qt(alpha/2, ddf))
CI_RRRC <- c(trtMeanDiffs - t_crit*stdErr,
             trtMeanDiffs + t_crit*stdErr)
names(CI_RRRC) <- c("Lower", "Upper")
print(CI_RRRC, digits = 4)
#>      Lower      Upper
#> -0.0879595  0.0003589
```

The confidence interval includes zero, which confirms the F-statistic finding that the reader-averaged FOM difference between treatments is not significant.

Calculated next is the confidence interval for the reader-averaged FOM for each treatment, i.e. $CI_{1-\alpha,RRRC,\theta_{i\bullet}}$. The relevant equations are Eqn. (7.11) and Eqn. (7.12). The implementation follows:

```
df_i <- array(dim = I)
den_i <- array(dim = I)
stdErr_i <- array(dim = I)
ci <- array(dim = c(I, 2))
CI_RRRC_IndvlTrt <- data.frame()
for (i in 1:I) {
  den_i[i] <- vc$IndividualTrt[i, "msREachTrt"] +
    J * max(vc$IndividualTrt[i, "cov2EachTrt"], 0)
  df_i[i] <-
    (den_i[i])^2/(vc$IndividualTrt[i, "msREachTrt"])^2 * (J - 1)
```

```
  stdErr_i[i] <- sqrt(den_i[i]/J)
  ci[i,] <-
    c(trtMeans[i] + qt(alpha/2, df_i[i]) * stdErr_i[i],
      trtMeans[i] + qt(1-alpha/2, df_i[i]) * stdErr_i[i])
  rowName <- paste0("trt", modalityID[i])
  CI_RRRC_IndvlTrt <- rbind(
    CI_RRRC_IndvlTrt,
    data.frame(Estimate = trtMeans[i],
               StdErr = stdErr_i[i],
               DFi = df_i[i],
               CILower = ci[i,1],
               CIUpper = ci[i,2],
               Cov2i = vc$IndividualTrt[i,"cov2EachTrt"],
               row.names = rowName,
               stringsAsFactors = FALSE))
}
print(CI_RRRC_IndvlTrt, digits = 4)
#>      Estimate  StdErr   DFi CILower CIUpper     Cov2i
#> trt0   0.8970 0.03317 12.74  0.8252  0.9689 0.0004840
#> trt1   0.9408 0.02157 12.71  0.8941  0.9875 0.0002042
```

8.3.2 Fixed-Reader Random-Case (FRRC) analysis

- The chi-square statistic is calculated using Eqn. (7.13) and Eqn. (7.15).
- The needed quantities are in `vc`.
- For example, MS(T) is in vc$TRanova["T", "MS"], see above. Likewise for `Cov2` and `Cov3`.
- The remaining needed quantities are:
- `Var` is in `vc$VarCom["Var", "Estimates"]`, whose value is 8.0228827×10^{-4}.
- `Cov1` is in `vc$VarCom["Cov1", "Estimates"]`, whose value is 3.4661371×10^{-4}.
- The degree of freedom is $I - 1$.
- The implementation follows:

```
den_FRRC <- vc$VarCom["Var","Estimates"] -
  vc$VarCom["Cov1","Estimates"] +
  (J - 1) * max(vc$VarCom["Cov2","Estimates"] -
                  vc$VarCom["Cov3","Estimates"] ,0)
chisqVal <- (I-1)*vc$TRanova["T","MS"]/den_FRRC
p <- 1 - pchisq(chisqVal, I - 1)
FTests <- data.frame(MS = c(vc$TRanova["T", "MS"], den_FRRC),
                     Chisq = c(chisqVal,NA),
                     DF = c(I - 1, NA),
```

```
                    p = c(p,NA),
                    row.names = c("Treatment", "Error"),
                    stringsAsFactors = FALSE)
print(FTests, digits = 4)
#>                   MS Chisq DF       p
#> Treatment 0.0047962 5.476  1 0.01928
#> Error     0.0008759    NA NA      NA
```

- Since p < 0.05, one has a significant finding.
- Freezing reader variability shows a significant difference between the treatments.
- The downside is that the conclusion applies only to the readers used in the study.
- The next step is to calculate the confidence interval for the reader-averaged FOM difference, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- The relevant equation is Eqn. (7.17), whose implementation follows.

```
stdErr <- sqrt(2 * den_FRRC/J)
zStat <- vector()
PrGTz <- vector()
CI <- array(dim = c(choose(I,2),2))
for (i in 1:choose(I,2)) {
  zStat[i] <- trtMeanDiffs[i]/stdErr
  PrGTz[i] <- 2 * pnorm(abs(zStat[i]), lower.tail = FALSE)
  CI[i, ] <- c(trtMeanDiffs[i] + qnorm(alpha/2) * stdErr,
               trtMeanDiffs[i] + qnorm(1-alpha/2) * stdErr)
}
ciDiffTrtFRRC <- data.frame(Estimate = trtMeanDiffs,
                            StdErr = rep(stdErr, choose(I, 2)),
                            z = zStat,
                            PrGTz = PrGTz,
                            CILower = CI[,1],
                            CIUpper = CI[,2],
                            row.names = diffTRName,
                            stringsAsFactors = FALSE)
print(ciDiffTrtFRRC, digits = 4)
#>           Estimate  StdErr     z   PrGTz  CILower   CIUpper
#> trt0-trt1  -0.0438 0.01872 -2.34 0.01928 -0.08049 -0.007115
```

- Consistent with the chi-square statistic significant finding, one finds that the treatment difference confidence interval does not include zero.
- The next step is to calculate the confidence interval for the reader-averaged figures of merit for each treatment, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}}$.
- The relevant formula is in Eqn. (7.18), whose implementation follows:

```
stdErr <- vector()
df <- vector()
CI <- array(dim = c(I,2))
ciAvgRdrEachTrt <- data.frame()
for (i in 1:I) {
  df[i] <- K - 1
  stdErr[i] <-
    sqrt((vc$IndividualTrt[i,"varEachTrt"] +
            (J-1)*max(vc$IndividualTrt[i,"cov2EachTrt"],0))/J)
  CI[i, ] <- c(trtMeans[i] + qnorm(alpha/2) * stdErr[i],
               trtMeans[i] + qnorm(1-alpha/2) * stdErr[i])
  rowName <- paste0("trt", modalityID[i])
  ciAvgRdrEachTrt <-
    rbind(ciAvgRdrEachTrt,
          data.frame(Estimate = trtMeans[i],
                     StdErr = stdErr[i],
                     DF = df[i],
                     CILower = CI[i,1],
                     CIUpper = CI[i,2],
                     row.names = rowName,
                     stringsAsFactors = FALSE))
}
print(ciAvgRdrEachTrt, digits = 4)
#>      Estimate  StdErr  DF CILower CIUpper
#> trt0   0.8970 0.02429 113  0.8494  0.9446
#> trt1   0.9408 0.01678 113  0.9080  0.9737
```

- Finally, one calculates confidence intervals for the FOM differences for individual readers, i.e., $CI_{1-\alpha,FRRC,\theta_{ij}-\theta_{i'j}}$.
- The relevant formula is in Eqn. (7.19), whose implementation follows:

```
trtMeanDiffs1 <- array(dim = c(J, choose(I, 2)))
Reader <- array(dim = c(J, choose(I, 2)))
stdErr <- array(dim = c(J, choose(I, 2)))
zStat <- array(dim = c(J, choose(I, 2)))
trDiffNames <- array(dim = c(J, choose(I, 2)))
PrGTz <- array(dim = c(J, choose(I, 2)))
CIReader <- array(dim = c(J, choose(I, 2),2))
ciDiffTrtEachRdr <- data.frame()
for (j in 1:J) {
  Reader[j,] <- rep(readerID[j], choose(I, 2))
  stdErr[j,] <-
    sqrt(
      2 *
```

```
        (vc$IndividualRdr[j,"varEachRdr"] -
           vc$IndividualRdr[j,"cov1EachRdr"]))
  pair <- 1
  for (i in 1:I) {
    if (i == I) break
    for (ip in (i + 1):I) {
      trtMeanDiffs1[j, pair] <- foms[i, j] - foms[ip, j]
      trDiffNames[j,pair] <- diffTRName[pair]
      zStat[j,pair] <- trtMeanDiffs1[j,pair]/stdErr[j,pair]
      PrGTz[j,pair] <-
        2 * pnorm(abs(zStat[j,pair]), lower.tail = FALSE)
      CIReader[j, pair,] <-
        c(trtMeanDiffs1[j,pair] +
            qnorm(alpha/2) * stdErr[j,pair],
          trtMeanDiffs1[j,pair] +
            qnorm(1-alpha/2) * stdErr[j,pair])
      rowName <-
        paste0("rdr", Reader[j,pair], "::", trDiffNames[j, pair])
      ciDiffTrtEachRdr <- rbind(
        ciDiffTrtEachRdr,
        data.frame(Estimate = trtMeanDiffs1[j, pair],
                   StdErr = stdErr[j,pair],
                   z = zStat[j, pair],
                   PrGTz = PrGTz[j, pair],
                   CILower = CIReader[j, pair,1],
                   CIUpper = CIReader[j, pair,2],
                   row.names = rowName,
                   stringsAsFactors = FALSE))
      pair <- pair + 1
    }
  }
}
print(ciDiffTrtEachRdr, digits = 3)
#>                 Estimate StdErr      z  PrGTz CILower  CIUpper
#> rdr0::trt0-trt1  -0.0282 0.0255 -1.105 0.2693 -0.0782  0.02182
#> rdr1::trt0-trt1  -0.0465 0.0263 -1.769 0.0768 -0.0981  0.00501
#> rdr2::trt0-trt1  -0.0179 0.0312 -0.573 0.5668 -0.0790  0.04330
#> rdr3::trt0-trt1  -0.0262 0.0173 -1.518 0.1290 -0.0601  0.00764
#> rdr4::trt0-trt1  -0.1002 0.0441 -2.273 0.0230 -0.1865 -0.01381
```

The notation in the first column shows the reader and the treatment pairing. For example, `rdr1::trt0-trt1` means the FOM difference for reader `rdr1`. Only the fifth reader, i.e., `rdr4`, shows a significant difference between the treatments: the p-value is 0.023001 and the confidence interval also does not include zero. The large FOM difference for this reader, -0.100161, was enough to result in a

significant finding for FRRC analysis. The FOM differences for the other readers are about a factor of 2.1522491 or more smaller than that for this reader.

8.3.3 Random-Reader Fixed-Case (RRFC) analysis

The F-statistic is shown in Eqn. (7.20). This time ndf $= I - 1$ and ddf $= (I-1) \times (J-1)$, the values proposed in the Obuchowski-Rockette paper. The implementation follows:

```
den <- vc$TRanova["TR","MS"]
f <- vc$TRanova["T","MS"]/den
ddf <- ((I - 1) * (J - 1))
p <- 1 - pf(f, I - 1, ddf)
FTests_RRFC <-
  data.frame(DF = c(I-1,(I-1)*(J-1)),
             MS = c(vc$TRanova["T","MS"],vc$TRanova["TR","MS"]),
             F = c(f,NA),  p = c(p,NA),
             row.names = c("T","TR"),
             stringsAsFactors = FALSE)
print(FTests_RRFC, digits = 4)
#>    DF       MS     F       p
#> T   1 0.004796 8.704 0.04196
#> TR  4 0.000551    NA      NA
```

Freezing case variability also results in a significant finding, but the conclusion is only applicable to the specific case set used in the study. Next one calculates confidence intervals for the reader-averaged FOM differences, the relevant formula is in Eqn. (7.22), whose implementation follows.

```
stdErr <- sqrt(2 * den/J)
tStat <- vector()
PrGTt <- vector()
CI <- array(dim = c(choose(I,2), 2))
for (i in 1:choose(I,2)) {
  tStat[i] <- trtMeanDiffs[i]/stdErr
  PrGTt[i] <- 2 *
    pt(abs(tStat[i]), ddf, lower.tail = FALSE)
  CI[i, ] <- c(trtMeanDiffs[i] + qt(alpha/2, ddf) * stdErr,
               trtMeanDiffs[i] + qt(1-alpha/2, ddf) * stdErr)
}
ciDiffTrt_RRFC <-
  data.frame(Estimate = trtMeanDiffs,
             StdErr = rep(stdErr, choose(I, 2)),
             DF = rep(ddf, choose(I, 2)),
```

```
              t = tStat,
              PrGTt = PrGTt,
              CILower = CI[,1],
              CIUpper = CI[,2],
              row.names = diffTRName,
              stringsAsFactors = FALSE)

print(ciDiffTrt_RRFC, digits = 4)
#>           Estimate  StdErr DF     t   PrGTt  CILower  CIUpper
#> trt0-trt1  -0.0438 0.01485  4 -2.95 0.04196 -0.08502 -0.00258
```

- As expected because the overall F-test showed significance, the confidence interval does not include zero (the p-value is identical to that found by the F-test).
- This completes the hand calculations.

8.4 RJafroc: dataset02

The second example shows application of the `RJafroc` package function `StSignificanceTesting()` to `dataset02`. This function encapsulates all formulae discussed previously and accomplishes the analyses with a single function call. It returns an object, denoted `st1` below, that contains all results of the analysis. It is a `list` with the following components:

- `FOMs`, this in turn is a `list` containing the following data frames:
  - `foms`, the individual treatment-reader figures of merit, i.e., θ_{ij},
  - `trtMeans`, the treatment figures of merit averaged over readers, i.e., $\theta_{i\bullet}$,
  - `trtMeanDiffs`, the inter-treatment figures of merit differences averaged over readers, i.e., $\theta_{i\bullet} - \theta_{i'\bullet}$.
- `ANOVA`, a `list` containing the following data frames:
  - `TRanova`, the treatment-reader ANOVA table,
  - `VarCom`, Obuchowski-Rockette variance-covariances and correlations,
  - `IndividualTrt`, the mean-squares, `Var` and `Cov2` calculated over individual treatments,
  - `IndividualRdr`, the mean-squares, `Var` and `Cov1` calculated over individual readers.
- `RRRC`, a `list` containing the following data frames:
  - `FTests`, the results of the F-test,

  - `ciDiffTrt`, the confidence intervals for inter-treatment FOM differences, averaged over readers, denoted $CI_{1-\alpha,RRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$ in the previous chapter,
  - `ciAvgRdrEachTrt`, the confidence intervals for individual treatment FOMs, averaged over readers, denoted $CI_{1-\alpha,RRRC,\theta_{i\bullet}}$ in the previous chapter.

- `FRRC`, a `list` containing the following data frames:

  - `FTests`, the results of the F-tests, which in this case specializes to chi-square tests,
  - `ciDiffTrt`, the confidence intervals for inter-treatment FOM differences, averaged over readers, denoted $CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$ in the previous chapter,
  - `ciAvgRdrEachTrt`, the confidence intervals for individual treatment FOMs, averaged over readers, denoted $CI_{1-\alpha,FRRC,\theta_{i\bullet}}$ in the previous chapter,
  - `ciDiffTrtEachRdr`, the confidence intervals for inter-treatment FOM differences for individual readers, denoted $CI_{1-\alpha,FRRC,\theta_{ij}-\theta_{i'j}}$ in the previous chapter,
  - `IndividualRdrVarCov1`, the individual reader variance-covariances and means squares.

- `RRFC`, a `list` containing the following data frames:

  - `FTests`, the results of the F-tests, which in this case specializes to chi-square tests,
  - `ciDiffTrt`, the confidence intervals for inter-treatment FOM differences, averaged over readers, denoted $CI_{1-\alpha,RRFC,\theta_{i\bullet}-\theta_{i'\bullet}}$ in the previous chapter,
  - `ciAvgRdrEachTrt`, the confidence intervals for indvidual treatment FOMs, averaged over readers, denoted $CI_{1-\alpha,RRFC,\theta_{i\bullet}}$ in the previous chapter.

In the interest of clarity, in the first example using the `RJafroc` package the components of the returned object `st1` are listed separately and described explicitly. In the interest of brevity, in subsequent examples the object is listed in its entirety.

Online help on the `StSignificanceTesting()` function is available:

```
?`StSignificanceTesting`
```

The lower right `RStudio` panel contains the online description. Click on the small up-and-right pointing arrow icon to expand this to a new window.

8.4.1 Random-Reader Random-Case (RRRC) analysis

- Since `analysisOption` is not explicitly specified in the following code, the function `StSignificanceTesting` performs all three analyses: `RRRC`, `FRRC` and `RRFC`.
- Likewise, the significance level of the test, also an argument, `alpha`, defaults to 0.05.
- The code below applies `StSignificanceTesting()` and saves the returned object to `st1`.
- The first member of this object, a `list` named `FOMs`, is then displayed.
- `FOMs` contains three data frames:
  - `FOMS$foms`, the figures of merit for each treatment and reader,
  - `FOMS$trtMeans`, the figures of merit for each treatment averaged over readers, and
  - `FOMS$trtMeanDiffs`, the inter-treatment difference figures of merit averaged over readers. The difference is always the first treatment minus the second, etc., in this example, `trt0` minus `trt1`.

```
st1 <- StSignificanceTesting(dataset02, FOM = "Wilcoxon", method = "OR")
print(st1$FOMs, digits = 4)
#> $foms
#>        rdr0   rdr1   rdr2   rdr3   rdr4
#> trt0 0.9196 0.8588 0.9039 0.9731 0.8298
#> trt1 0.9478 0.9053 0.9217 0.9994 0.9300
#>
#> $trtMeans
#>      Estimate
#> trt0   0.8970
#> trt1   0.9408
#>
#> $trtMeanDiffs
#>           Estimate
#> trt0-trt1  -0.0438
```

- Displayed next are the variance components and mean-squares contained in the `ANOVA list`.
  - `ANOVA$TRanova` contains the treatment-reader ANOVA table, i.e. the sum of squares, the degrees of freedom and the mean-squares, for treatment, reader and treatment-reader factors, i.e., `T`, `R` and `TR`.
  - `ANOVA$VarCom` contains the OR variance components and the correlations.
  - `ANOVA$IndividualTrt` contains the quantities necessary for individual treatment analyses.
  - `ANOVA$IndividualRdr` contains the quantities necessary for individual reader analyses.

```
print(st1$ANOVA, digits = 4)
#> $TRanova
#>          SS DF       MS
#> T  0.004796  1 0.004796
#> R  0.015345  4 0.003836
#> TR 0.002204  4 0.000551
#>
#> $VarCom
#>       Estimates   Rhos
#> VarR  0.0015350     NA
#> VarTR 0.0002004     NA
#> Cov1  0.0003466 0.4320
#> Cov2  0.0003441 0.4289
#> Cov3  0.0002390 0.2979
#> Var   0.0008023     NA
#>
#> $IndividualTrt
#>      DF msREachTrt varEachTrt cov2EachTrt
#> trt0  4   0.003083  0.0010141   0.0004840
#> trt1  4   0.001305  0.0005905   0.0002042
#>
#> $IndividualRdr
#>      DF msTEachRdr varEachRdr cov1EachRdr
#> rdr0  1  0.0003971  0.0006989   3.735e-04
#> rdr1  1  0.0010829  0.0011061   7.602e-04
#> rdr2  1  0.0001597  0.0008423   3.553e-04
#> rdr3  1  0.0003445  0.0001506   1.083e-06
#> rdr4  1  0.0050161  0.0012136   2.430e-04
```

- Displayed next are the results of the RRRC significance test, contained in `st1$RRRC`.

```
print(st1$RRRC$FTests, digits = 4)
#>              DF       MS FStat       p
#> Treatment  1.00 0.004796 4.456 0.05167
#> Error     15.26 0.001076    NA      NA
```

- `st1$RRRC$FTests` contains the results of the F-tests: the degrees of freedom, the mean-squares, the observed value of the F-statistic and the p-value for rejecting the NH, listed separately, where applicable, for the treatment and error terms.
- For example, the treatment mean squares is `st1$RRRC$FTests["Treatment", "MS"]` whose value is 0.00479617.

```
print(st1$RRRC$ciDiffTrt, digits = 3)
#>           Estimate StdErr   DF     t  PrGTt CILower  CIUpper
#> trt0-trt1  -0.0438 0.0207 15.3 -2.11 0.0517  -0.088 0.000359
```

- `st1$RRRC$ciDiffTrt` contains the results of the confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.

```
print(st1$RRRC$ciAvgRdrEachTrt, digits = 4)
#>      Estimate  StdErr    DF CILower CIUpper      Cov2
#> trt0   0.8970 0.03317 12.74  0.8252  0.9689 0.0004840
#> trt1   0.9408 0.02157 12.71  0.8941  0.9875 0.0002042
```

- `st1$RRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}}$.

8.4.2 Fixed-Reader Random-Case (FRRC) analysis

- Displayed next are the results of FRRC analysis, contained in `st1$FRRC`.
- `st1$FRRC$FTests` contains the results of the F-tests: the degrees of freedom, the mean-squares, the observed value of the F-statistic and the p-value for rejecting the NH, listed separately, where applicable, for the treatment and error terms.
- For example, the treatment mean squares is `st1$FRRC$FTests["Treatment", "MS"]` whose value is 0.00479617.

```
print(st1$FRRC$FTests, digits = 4)
#>                   MS Chisq DF       p
#> Treatment 0.0047962 5.476  1 0.01928
#> Error     0.0008759    NA NA      NA
```

- Note that this time the output lists a chi-square distribution observed value, 5.47595324, with degree of freedom $df = I - 1 = 1$.
- The listed mean-squares and the p-value agree with the previously performed hand calculations.
- For FRRC analysis the value of the chi-square statistic is significant and the p-value is smaller than α.

```
print(st1$FRRC$ciDiffTrt, digits = 4)
#>           Estimate  StdErr     z   PrGTz  CILower   CIUpper
#> trt0-trt1  -0.0438 0.01872 -2.34 0.01928 -0.08049 -0.007115
```

- `st1$FRRC$ciDiffTrt` contains confidence intervals for inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- The confidence interval excludes zero, and the p-value, listed under `PrGTz` (for probability greater than `z`) is smaller than 0.05.
- One could be using the t-distribution with infinite degrees of freedom, but this is identical to the normal distribution. Hence the listed value is a `z` statistic, i.e., `z = -0.043800322/0.018717483` = -2.34007543.

```
print(st1$FRRC$ciAvgRdrEachTrt, digits = 4)
#>      Estimate  StdErr  DF CILower CIUpper
#> trt0   0.8970 0.02429 113  0.8494  0.9446
#> trt1   0.9408 0.01678 113  0.9080  0.9737
```

- `st1$FRRC$st1$FRRC$ciAvgRdrEachTrt` contains confidence intervals for individual treatment FOMs, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}}$.

```
print(st1$FRRC$ciDiffTrtEachRdr, digits = 3)
#>                 Estimate StdErr      z  PrGTz CILower  CIUpper
#> rdr0::trt0-trt1  -0.0282 0.0255 -1.105 0.2693 -0.0782  0.02182
#> rdr1::trt0-trt1  -0.0465 0.0263 -1.769 0.0768 -0.0981  0.00501
#> rdr2::trt0-trt1  -0.0179 0.0312 -0.573 0.5668 -0.0790  0.04330
#> rdr3::trt0-trt1  -0.0262 0.0173 -1.518 0.1290 -0.0601  0.00764
#> rdr4::trt0-trt1  -0.1002 0.0441 -2.273 0.0230 -0.1865 -0.01381
```

- `st1$FRRC$st1$FRRC$ciDiffTrtEachRdr` contains confidence intervals for inter-treatment difference FOMs, for each reader, i.e., $CI_{1-\alpha,FRRC,\theta_{ij}-\theta_{i'j}}$.

8.4.3 Random-Reader Fixed-Case (RRFC) analysis

```
print(st1$RRFC$FTests, digits = 4)
#>    DF       MS     F       p
#> T   1 0.004796 8.704 0.04196
#> TR  4 0.000551    NA      NA
```

- `st1$RRFC$FTests` contains results of the F-test: the degrees of freedom, the mean-squares, the observed value of the F-statistic and the p-value for rejecting the NH, listed separately, where applicable, for the treatment and treatment-reader terms. The latter is also termed the "error term".
- For example, the treatment-reader mean squares is `st1$RRFC$FTests["TR", "MS"]` whose value is $5.51030622 \times 10^{-4}$.

```
print(st1$RRFC$ciDiffTrt, digits = 4)
#>           Estimate  StdErr DF     t   PrGTt  CILower  CIUpper
#> trt0-trt1  -0.0438 0.01485  4 -2.95 0.04196 -0.08502 -0.00258
```

- `st1$RRFC$ciDiffTrt` contains confidence intervals for the inter-treatment paired difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}-\theta_{i'\bullet}}$.

```
print(st1$RRFC$ciAvgRdrEachTrt, digits = 4)
#>      Estimate  StdErr DF CILower CIUpper
#> Trt0   0.8970 0.02483  4  0.8281  0.9660
#> Trt1   0.9408 0.01615  4  0.8960  0.9857
```

- `st1$RRFC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}}$.

8.5 RJafroc: dataset04

- The third example uses the Federica Zanca dataset (Zanca et al., 2009), i.e., `dataset04`, which has five modalities and four readers.
- It illustrates the situation when multiple treatment pairings are involved. In contrast, the previous example had only one treatment pairing.
- Since this is an FROC dataset, in order to keep it comparable with the previous example, one converts it to an inferred-ROC dataset.
- The function `DfFroc2Roc(dataset04)` converts, using the highest-rating, the FROC dataset to an inferred-ROC dataset.
- The results are contained in `st2`.
- As noted earlier, this time the object is listed in its entirety.

```
ds <- DfFroc2Roc(dataset04) # convert to ROC
I <- length(ds$ratings$NL[,1,1,1])
J <- length(ds$ratings$NL[1,,1,1])
cat("I = ", I, ", J = ", J, "\n")
#> I =  5 , J =  4
st2 <- StSignificanceTesting(ds, FOM = "Wilcoxon", method = "OR")
print(st2, digits = 3)
#> $FOMs
#> $FOMs$foms
#>       rdr1  rdr2  rdr3  rdr4
#> trt1 0.904 0.798 0.812 0.866
#> trt2 0.864 0.845 0.821 0.872
#> trt3 0.813 0.816 0.753 0.857
#> trt4 0.902 0.832 0.789 0.880
```

```
#> trt5 0.841 0.773 0.771 0.848
#>
#> $FOMs$trtMeans
#>      Estimate
#> trt1    0.845
#> trt2    0.850
#> trt3    0.810
#> trt4    0.851
#> trt5    0.808
#>
#> $FOMs$trtMeanDiffs
#>            Estimate
#> trt1-trt2 -0.005100
#> trt1-trt3  0.035325
#> trt1-trt4 -0.005412
#> trt1-trt5  0.036775
#> trt2-trt3  0.040425
#> trt2-trt4 -0.000312
#> trt2-trt5  0.041875
#> trt3-trt4 -0.040737
#> trt3-trt5  0.001450
#> trt4-trt5  0.042187
#>
#>
#> $ANOVA
#> $ANOVA$TRanova
#>         SS DF       MS
#> T  0.00759  4 0.001897
#> R  0.02188  3 0.007294
#> TR 0.00555 12 0.000462
#>
#> $ANOVA$VarCom
#>       Estimates  Rhos
#> VarR   1.28e-03    NA
#> VarTR -1.09e-05    NA
#> Cov1   2.95e-04 0.374
#> Cov2   2.33e-04 0.296
#> Cov3   2.12e-04 0.269
#> Var    7.89e-04    NA
#>
#> $ANOVA$IndividualTrt
#>      DF msREachTrt varEachTrt cov2EachTrt
#> trt1  3   0.002422   0.000711    0.000211
#> trt2  3   0.000523   0.000751    0.000266
#> trt3  3   0.001855   0.000876    0.000246
```

```
#> trt4  3   0.002578   0.000727    0.000220
#> trt5  3   0.001766   0.000882    0.000222
#>
#> $ANOVA$IndividualRdr
#>      DF msTEachRdr varEachRdr cov1EachRdr
#> rdr1  4   0.001551   0.000689    0.000215
#> rdr2  4   0.000794   0.000824    0.000346
#> rdr3  4   0.000786   0.001009    0.000354
#> rdr4  4   0.000153   0.000635    0.000265
#>
#>
#> $RRRC
#> $RRRC$FTests
#>             DF       MS FStat      p
#> Treatment  4.0 0.001897  3.47 0.0305
#> Error     16.8 0.000547    NA     NA
#>
#> $RRRC$ciDiffTrt
#>            Estimate StdErr   DF       t  PrGTt   CILower  CIUpper
#> trt1-trt2 -0.005100 0.0165 16.8 -0.3084 0.7616 -0.040021  0.02982
#> trt1-trt3  0.035325 0.0165 16.8  2.1361 0.0477  0.000404  0.07025
#> trt1-trt4 -0.005412 0.0165 16.8 -0.3273 0.7475 -0.040334  0.02951
#> trt1-trt5  0.036775 0.0165 16.8  2.2238 0.0402  0.001854  0.07170
#> trt2-trt3  0.040425 0.0165 16.8  2.4445 0.0258  0.005504  0.07535
#> trt2-trt4 -0.000312 0.0165 16.8 -0.0189 0.9851 -0.035234  0.03461
#> trt2-trt5  0.041875 0.0165 16.8  2.5322 0.0216  0.006954  0.07680
#> trt3-trt4 -0.040737 0.0165 16.8 -2.4634 0.0249 -0.075659 -0.00582
#> trt3-trt5  0.001450 0.0165 16.8  0.0877 0.9312 -0.033471  0.03637
#> trt4-trt5  0.042187 0.0165 16.8  2.5511 0.0208  0.007266  0.07711
#>
#> $RRRC$ciAvgRdrEachTrt
#>      Estimate StdErr    DF CILower CIUpper     Cov2
#> trt1    0.845 0.0286  5.46   0.774   0.917 0.000211
#> trt2    0.850 0.0199 27.72   0.809   0.891 0.000266
#> trt3    0.810 0.0266  7.04   0.747   0.873 0.000246
#> trt4    0.851 0.0294  5.40   0.777   0.925 0.000220
#> trt5    0.808 0.0258  6.78   0.747   0.870 0.000222
#>
#>
#> $FRRC
#> $FRRC$FTests
#>                  MS Chisq DF       p
#> Treatment 0.001897  13.6  4 0.00868
#> Error     0.000558    NA NA      NA
#>
```

```
#> $FRRC$ciDiffTrt
#>            Estimate StdErr       z  PrGTz  CILower CIUpper
#> trt1-trt2 -0.005100 0.0167 -0.3054 0.7601 -0.03783  0.0276
#> trt1-trt3  0.035325 0.0167  2.1151 0.0344  0.00259  0.0681
#> trt1-trt4 -0.005412 0.0167 -0.3241 0.7459 -0.03815  0.0273
#> trt1-trt5  0.036775 0.0167  2.2019 0.0277  0.00404  0.0695
#> trt2-trt3  0.040425 0.0167  2.4204 0.0155  0.00769  0.0732
#> trt2-trt4 -0.000312 0.0167 -0.0187 0.9851 -0.03305  0.0324
#> trt2-trt5  0.041875 0.0167  2.5073 0.0122  0.00914  0.0746
#> trt3-trt4 -0.040737 0.0167 -2.4392 0.0147 -0.07347 -0.0080
#> trt3-trt5  0.001450 0.0167  0.0868 0.9308 -0.03128  0.0342
#> trt4-trt5  0.042187 0.0167  2.5260 0.0115  0.00945  0.0749
#>
#> $FRRC$ciAvgRdrEachTrt
#>      Estimate StdErr  DF CILower CIUpper
#> trt1    0.845 0.0183 199   0.809   0.881
#> trt2    0.850 0.0197 199   0.812   0.889
#> trt3    0.810 0.0201 199   0.770   0.849
#> trt4    0.851 0.0186 199   0.814   0.887
#> trt5    0.808 0.0197 199   0.770   0.847
#>
#> $FRRC$ciDiffTrtEachRdr
#>                  Estimate StdErr       z   PrGTz  CILower CIUpper
#> rdr1::trt1-trt2  0.04000 0.0308  1.2989 0.19400 -0.02036  0.1004
#> rdr1::trt1-trt3  0.09130 0.0308  2.9646 0.00303  0.03094  0.1517
#> rdr1::trt1-trt4  0.00190 0.0308  0.0617 0.95081 -0.05846  0.0623
#> rdr1::trt1-trt5  0.06285 0.0308  2.0408 0.04127  0.00249  0.1232
#> rdr1::trt2-trt3  0.05130 0.0308  1.6658 0.09576 -0.00906  0.1117
#> rdr1::trt2-trt4 -0.03810 0.0308 -1.2372 0.21603 -0.09846  0.0223
#> rdr1::trt2-trt5  0.02285 0.0308  0.7420 0.45811 -0.03751  0.0832
#> rdr1::trt3-trt4 -0.08940 0.0308 -2.9029 0.00370 -0.14976 -0.0290
#> rdr1::trt3-trt5 -0.02845 0.0308 -0.9238 0.35559 -0.08881  0.0319
#> rdr1::trt4-trt5  0.06095 0.0308  1.9791 0.04780  0.00059  0.1213
#> rdr2::trt1-trt2 -0.04650 0.0309 -1.5039 0.13260 -0.10710  0.0141
#> rdr2::trt1-trt3 -0.01815 0.0309 -0.5870 0.55719 -0.07875  0.0424
#> rdr2::trt1-trt4 -0.03330 0.0309 -1.0770 0.28147 -0.09390  0.0273
#> rdr2::trt1-trt5  0.02520 0.0309  0.8150 0.41505 -0.03540  0.0858
#> rdr2::trt2-trt3  0.02835 0.0309  0.9169 0.35918 -0.03225  0.0889
#> rdr2::trt2-trt4  0.01320 0.0309  0.4269 0.66943 -0.04740  0.0738
#> rdr2::trt2-trt5  0.07170 0.0309  2.3190 0.02040  0.01110  0.1323
#> rdr2::trt3-trt4 -0.01515 0.0309 -0.4900 0.62414 -0.07575  0.0454
#> rdr2::trt3-trt5  0.04335 0.0309  1.4021 0.16090 -0.01725  0.1039
#> rdr2::trt4-trt5  0.05850 0.0309  1.8921 0.05848 -0.00210  0.1191
#> rdr3::trt1-trt2 -0.00875 0.0362 -0.2418 0.80896 -0.07969  0.0622
#> rdr3::trt1-trt3  0.05900 0.0362  1.6302 0.10307 -0.01194  0.1299
```

```
#> rdr3::trt1-trt4  0.02310 0.0362  0.6383 0.52331 -0.04784  0.0940
#> rdr3::trt1-trt5  0.04060 0.0362  1.1218 0.26196 -0.03034  0.1115
#> rdr3::trt2-trt3  0.06775 0.0362  1.8719 0.06122 -0.00319  0.1387
#> rdr3::trt2-trt4  0.03185 0.0362  0.8800 0.37885 -0.03909  0.1028
#> rdr3::trt2-trt5  0.04935 0.0362  1.3635 0.17271 -0.02159  0.1203
#> rdr3::trt3-trt4 -0.03590 0.0362 -0.9919 0.32124 -0.10684  0.0350
#> rdr3::trt3-trt5 -0.01840 0.0362 -0.5084 0.61118 -0.08934  0.0525
#> rdr3::trt4-trt5  0.01750 0.0362  0.4835 0.62872 -0.05344  0.0884
#> rdr4::trt1-trt2 -0.00515 0.0272 -0.1893 0.84987 -0.05848  0.0482
#> rdr4::trt1-trt3  0.00915 0.0272  0.3363 0.73664 -0.04418  0.0625
#> rdr4::trt1-trt4 -0.01335 0.0272 -0.4907 0.62366 -0.06668  0.0400
#> rdr4::trt1-trt5  0.01845 0.0272  0.6781 0.49770 -0.03488  0.0718
#> rdr4::trt2-trt3  0.01430 0.0272  0.5256 0.59918 -0.03903  0.0676
#> rdr4::trt2-trt4 -0.00820 0.0272 -0.3014 0.76312 -0.06153  0.0451
#> rdr4::trt2-trt5  0.02360 0.0272  0.8674 0.38572 -0.02973  0.0769
#> rdr4::trt3-trt4 -0.02250 0.0272 -0.8270 0.40825 -0.07583  0.0308
#> rdr4::trt3-trt5  0.00930 0.0272  0.3418 0.73249 -0.04403  0.0626
#> rdr4::trt4-trt5  0.03180 0.0272  1.1688 0.24249 -0.02153  0.0851
#>
#> $FRRC$IndividualRdrVarCov1
#>      varEachRdr cov1EachRdr
#> rdr1   0.000689    0.000215
#> rdr2   0.000824    0.000346
#> rdr3   0.001009    0.000354
#> rdr4   0.000635    0.000265
#>
#>
#> $RRFC
#> $RRFC$FTests
#>    DF       MS   F      p
#> T   4 0.001897 4.1 0.0253
#> TR 12 0.000462  NA     NA
#>
#> $RRFC$ciDiffTrt
#>            Estimate StdErr DF       t  PrGTt  CILower  CIUpper
#> trt1-trt2 -0.005100 0.0152 12 -0.3355 0.7431 -0.03822  0.02802
#> trt1-trt3  0.035325 0.0152 12  2.3237 0.0385  0.00220  0.06845
#> trt1-trt4 -0.005412 0.0152 12 -0.3560 0.7280 -0.03854  0.02771
#> trt1-trt5  0.036775 0.0152 12  2.4191 0.0324  0.00365  0.06990
#> trt2-trt3  0.040425 0.0152 12  2.6592 0.0208  0.00730  0.07355
#> trt2-trt4 -0.000312 0.0152 12 -0.0206 0.9839 -0.03344  0.03281
#> trt2-trt5  0.041875 0.0152 12  2.7546 0.0175  0.00875  0.07500
#> trt3-trt4 -0.040737 0.0152 12 -2.6797 0.0200 -0.07386 -0.00761
#> trt3-trt5  0.001450 0.0152 12  0.0954 0.9256 -0.03167  0.03457
#> trt4-trt5  0.042187 0.0152 12  2.7751 0.0168  0.00906  0.07531
```

```
#>
#> $RRFC$ciAvgRdrEachTrt
#>      Estimate StdErr DF CILower CIUpper
#> Trt1    0.845 0.0246  3   0.767   0.923
#> Trt2    0.850 0.0114  3   0.814   0.887
#> Trt3    0.810 0.0215  3   0.741   0.878
#> Trt4    0.851 0.0254  3   0.770   0.931
#> Trt5    0.808 0.0210  3   0.742   0.875
```

8.5.1 Random-Reader Random-Case (RRRC) analysis

- `st2$RRRC$FTests` contains the results of the F-test.
- In this example `ndf` = 4 because there are I = 5 treatments. Since the p-value is less than 0.05, at least one treatment-pairing FOM difference is significantly different from zero.
- `st2$RRRC$ciDiffTrt` contains the confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- With I = 5 treatments there are 10 distinct treatment-pairings.
- Looking at the `PrGTt` (for probability greater than `t`) column, one finds six pairings that are significant: `trt1-trt3`, `trt1-trt5`, etc. The smallest p-value is for the `trt4-trt5` pairing.
- `st2$RRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}}$.
- Looking at the `Estimate` column one confirms that `trt5` has the smallest FOM while `trt4` has the highest.

8.5.2 Fixed-Reader Random-Case (FRRC) analysis

- `st2$FRRC$FTests` contains results of the F-tests, which in this situation is actually a chi-square test of the NH.
- Again, `ndf` = 4 because there are I = 5 treatments. Since the p-value is less than 0.05, at least one treatment-pairing FOM difference is significantly different from zero.
- `st2$FRRC$ciDiffTrt` contains confidence intervals for the inter-treatment paired difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- With I = 5 treatments there are 10 distinct treatment-pairings.

- Looking at the `PrGTt` column, one finds six pairings that are significant: `trt1-trt3`, `trt1-trt5`, etc. The smallest p-value is for the `trt4-trt5` pairing.
- `st2$FRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}}$.
- The `Estimate` column confirms that `trt5` has the smallest FOM while `trt4` has the highest.

8.5.3 Random-Reader Fixed-Case (RRFC) analysis

- `st2$RRFC$FTests` contains the results of the F-test of the NH.
- Again, `ndf` = 4 because there are I = 5 treatments. Since the p-value is less than 0.05, at least one treatment-pairing FOM difference is significantly different from zero.
- `st2$RRFC$ciDiffTrt` contains confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- With I = 5 treatments there are 10 distinct treatment-pairings.
- The `PrGTt` column shows that six pairings are significant: `trt1-trt3`, `trt1-trt5`, etc. The smallest p-value is for the `trt4-trt5` pairing.
- `st2$RRFC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}}$.
- The `Estimate` column confirms that `trt5` has the smallest FOM while `trt4` has the highest (the `Estimates` column is identical for RRRC, FRRC and RRFC analyses).

8.6 RJafroc: dataset04, FROC

- The fourth example uses `dataset04`, but this time we use the FROC data, specifically, we do not convert it to inferred-ROC.
- Since this is an FROC dataset, one needs to use an FROC figure of merit.
- In this example the weighted AFROC figure of merit `FOM = "wAFROC"` is specified. This is the recommended figure of merit when both normal and abnormal cases are present in the dataset.
- If the dataset does not contain normal cases, then the weighted AFROC1 figure of merit `FOM = "wAFROC1"` should be specified.
- The results are contained in `st3`.
- As noted earlier, this time the object is listed in its entirety.

```
ds <- dataset04 # do NOT convert to ROC
FOM <- "wAFROC"
st3 <- StSignificanceTesting(ds, FOM = FOM, method = "OR")
print(st3, digits = 3)
#> $FOMs
#> $FOMs$foms
#>       rdr1  rdr3  rdr4  rdr5
#> trt1 0.779 0.725 0.704 0.805
#> trt2 0.787 0.727 0.723 0.804
#> trt3 0.730 0.716 0.672 0.773
#> trt4 0.810 0.743 0.694 0.829
#> trt5 0.749 0.682 0.655 0.771
#>
#> $FOMs$trtMeans
#>      Estimate
#> trt1    0.753
#> trt2    0.760
#> trt3    0.723
#> trt4    0.769
#> trt5    0.714
#>
#> $FOMs$trtMeanDiffs
#>           Estimate
#> trt1-trt2 -0.00686
#> trt1-trt3  0.03061
#> trt1-trt4 -0.01604
#> trt1-trt5  0.03884
#> trt2-trt3  0.03747
#> trt2-trt4 -0.00918
#> trt2-trt5  0.04570
#> trt3-trt4 -0.04665
#> trt3-trt5  0.00823
#> trt4-trt5  0.05488
#>
#>
#> $ANOVA
#> $ANOVA$TRanova
#>         SS DF      MS
#> T  0.00927  4 0.00232
#> R  0.03540  3 0.01180
#> TR 0.00204 12 0.00017
#>
#> $ANOVA$VarCom
#>       Estimates  Rhos
#> VarR   0.002209    NA
```

```
#> VarTR -0.000305    NA
#> Cov1   0.000422 0.455
#> Cov2   0.000336 0.362
#> Cov3   0.000304 0.328
#> Var    0.000928    NA
#>
#> $ANOVA$IndividualTrt
#>      DF msREachTrt varEachTrt cov2EachTrt
#> trt1  3    0.00221   0.000877    0.000333
#> trt2  3    0.00171   0.000939    0.000380
#> trt3  3    0.00171   0.000970    0.000297
#> trt4  3    0.00386   0.000859    0.000311
#> trt5  3    0.00298   0.000995    0.000359
#>
#> $ANOVA$IndividualRdr
#>      DF msTEachRdr varEachRdr cov1EachRdr
#> rdr1  4   0.001014   0.000883    0.000412
#> rdr3  4   0.000509   0.000897    0.000436
#> rdr4  4   0.000698   0.001171    0.000495
#> rdr5  4   0.000604   0.000762    0.000345
#>
#>
#> $RRRC
#> $RRRC$FTests
#>             DF       MS FStat        p
#> Treatment  4.0 0.002317   7.8 0.000117
#> Error     36.8 0.000297    NA       NA
#>
#> $RRRC$ciDiffTrt
#>           Estimate StdErr   DF      t    PrGTt  CILower  CIUpper
#> trt1-trt2 -0.00686 0.0122 36.8 -0.563 5.77e-01 -0.03155  0.01784
#> trt1-trt3  0.03061 0.0122 36.8  2.512 1.65e-02  0.00592  0.05531
#> trt1-trt4 -0.01604 0.0122 36.8 -1.316 1.96e-01 -0.04073  0.00866
#> trt1-trt5  0.03884 0.0122 36.8  3.188 2.92e-03  0.01415  0.06354
#> trt2-trt3  0.03747 0.0122 36.8  3.075 3.96e-03  0.01278  0.06217
#> trt2-trt4 -0.00918 0.0122 36.8 -0.753 4.56e-01 -0.03387  0.01552
#> trt2-trt5  0.04570 0.0122 36.8  3.750 6.07e-04  0.02100  0.07040
#> trt3-trt4 -0.04665 0.0122 36.8 -3.828 4.85e-04 -0.07135 -0.02195
#> trt3-trt5  0.00823 0.0122 36.8  0.675 5.04e-01 -0.01647  0.03292
#> trt4-trt5  0.05488 0.0122 36.8  4.504 6.52e-05  0.03018  0.07957
#>
#> $RRRC$ciAvgRdrEachTrt
#>      Estimate StdErr    DF CILower CIUpper     Cov2
#> trt1    0.753 0.0298  7.71   0.684   0.822 0.000333
#> trt2    0.760 0.0284 10.69   0.697   0.823 0.000380
```

```
#> trt3    0.723 0.0269  8.62   0.661   0.784 0.000297
#> trt4    0.769 0.0357  5.24   0.679   0.860 0.000311
#> trt5    0.714 0.0333  6.59   0.635   0.794 0.000359
#>
#>
#> $FRRC
#> $FRRC$FTests
#>                  MS Chisq DF       p
#> Treatment 0.002317  15.4  4 0.00393
#> Error     0.000602    NA NA      NA
#>
#> $FRRC$ciDiffTrt
#>           Estimate StdErr      z   PrGTz  CILower CIUpper
#> trt1-trt2 -0.00686 0.0173 -0.395 0.69260 -0.04085  0.0271
#> trt1-trt3  0.03061 0.0173  1.765 0.07753 -0.00338  0.0646
#> trt1-trt4 -0.01604 0.0173 -0.925 0.35518 -0.05003  0.0180
#> trt1-trt5  0.03884 0.0173  2.240 0.02511  0.00485  0.0728
#> trt2-trt3  0.03747 0.0173  2.161 0.03073  0.00348  0.0715
#> trt2-trt4 -0.00918 0.0173 -0.529 0.59662 -0.04317  0.0248
#> trt2-trt5  0.04570 0.0173  2.635 0.00841  0.01171  0.0797
#> trt3-trt4 -0.04665 0.0173 -2.690 0.00715 -0.08064 -0.0127
#> trt3-trt5  0.00823 0.0173  0.474 0.63515 -0.02576  0.0422
#> trt4-trt5  0.05488 0.0173  3.164 0.00155  0.02089  0.0889
#>
#> $FRRC$ciAvgRdrEachTrt
#>      Estimate StdErr  DF CILower CIUpper
#> trt1    0.753 0.0217 199   0.711   0.796
#> trt2    0.760 0.0228 199   0.715   0.805
#> trt3    0.723 0.0216 199   0.680   0.765
#> trt4    0.769 0.0212 199   0.728   0.811
#> trt5    0.714 0.0228 199   0.670   0.759
#>
#> $FRRC$ciDiffTrtEachRdr
#>                  Estimate StdErr       z   PrGTz  CILower   CIUpper
#> rdr1::trt1-trt2 -0.00773 0.0307 -0.2520 0.80105 -0.06788  0.052416
#> rdr1::trt1-trt3  0.04957 0.0307  1.6154 0.10622 -0.01057  0.109724
#> rdr1::trt1-trt4 -0.03087 0.0307 -1.0058 0.31451 -0.09102  0.029282
#> rdr1::trt1-trt5  0.03047 0.0307  0.9928 0.32083 -0.02968  0.090616
#> rdr1::trt2-trt3  0.05731 0.0307  1.8674 0.06185 -0.00284  0.117457
#> rdr1::trt2-trt4 -0.02313 0.0307 -0.7538 0.45097 -0.08328  0.037016
#> rdr1::trt2-trt5  0.03820 0.0307  1.2448 0.21322 -0.02195  0.098349
#> rdr1::trt3-trt4 -0.08044 0.0307 -2.6212 0.00876 -0.14059 -0.020293
#> rdr1::trt3-trt5 -0.01911 0.0307 -0.6226 0.53352 -0.07926  0.041041
#> rdr1::trt4-trt5  0.06133 0.0307  1.9986 0.04566  0.00118  0.121482
#> rdr3::trt1-trt2 -0.00201 0.0304 -0.0661 0.94726 -0.06152  0.057504
```

```
#> rdr3::trt1-trt3  0.00913 0.0304  0.3008 0.76357 -0.05038  0.068646
#> rdr3::trt1-trt4 -0.01822 0.0304 -0.6002 0.54836 -0.07774  0.041287
#> rdr3::trt1-trt5  0.04262 0.0304  1.4035 0.16046 -0.01690  0.102129
#> rdr3::trt2-trt3  0.01114 0.0304  0.3669 0.71367 -0.04837  0.070654
#> rdr3::trt2-trt4 -0.01622 0.0304 -0.5341 0.59329 -0.07573  0.043296
#> rdr3::trt2-trt5  0.04462 0.0304  1.4697 0.14165 -0.01489  0.104137
#> rdr3::trt3-trt4 -0.02736 0.0304 -0.9010 0.36758 -0.08687  0.032154
#> rdr3::trt3-trt5  0.03348 0.0304  1.1027 0.27014 -0.02603  0.092996
#> rdr3::trt4-trt5  0.06084 0.0304  2.0037 0.04510  0.00133  0.120354
#> rdr4::trt1-trt2 -0.01899 0.0368 -0.5166 0.60543 -0.09104  0.053061
#> rdr4::trt1-trt3  0.03132 0.0368  0.8519 0.39429 -0.04074  0.103370
#> rdr4::trt1-trt4  0.00927 0.0368  0.2521 0.80099 -0.06279  0.081320
#> rdr4::trt1-trt5  0.04845 0.0368  1.3179 0.18753 -0.02360  0.120503
#> rdr4::trt2-trt3  0.05031 0.0368  1.3685 0.17116 -0.02174  0.122361
#> rdr4::trt2-trt4  0.02826 0.0368  0.7687 0.44209 -0.04379  0.100311
#> rdr4::trt2-trt5  0.06744 0.0368  1.8345 0.06658 -0.00461  0.139495
#> rdr4::trt3-trt4 -0.02205 0.0368 -0.5998 0.54864 -0.09410  0.050003
#> rdr4::trt3-trt5  0.01713 0.0368  0.4661 0.64118 -0.05492  0.089186
#> rdr4::trt4-trt5  0.03918 0.0368  1.0659 0.28649 -0.03287  0.111236
#> rdr5::trt1-trt2  0.00131 0.0289  0.0453 0.96385 -0.05526  0.057881
#> rdr5::trt1-trt3  0.03243 0.0289  1.1237 0.26116 -0.02414  0.089006
#> rdr5::trt1-trt4 -0.02432 0.0289 -0.8425 0.39953 -0.08089  0.032256
#> rdr5::trt1-trt5  0.03384 0.0289  1.1724 0.24102 -0.02273  0.090414
#> rdr5::trt2-trt3  0.03112 0.0289  1.0783 0.28089 -0.02545  0.087698
#> rdr5::trt2-trt4 -0.02563 0.0289 -0.8878 0.37466 -0.08220  0.030948
#> rdr5::trt2-trt5  0.03253 0.0289  1.1271 0.25969 -0.02404  0.089106
#> rdr5::trt3-trt4 -0.05675 0.0289 -1.9661 0.04929 -0.11332 -0.000177
#> rdr5::trt3-trt5  0.00141 0.0289  0.0488 0.96109 -0.05516  0.057981
#> rdr5::trt4-trt5  0.05816 0.0289  2.0149 0.04391  0.00159  0.114731
#>
#> $FRRC$IndividualRdrVarCov1
#>      varEachRdr cov1EachRdr
#> rdr1   0.000883    0.000412
#> rdr3   0.000897    0.000436
#> rdr4   0.001171    0.000495
#> rdr5   0.000762    0.000345
#>
#>
#> $RRFC
#> $RRFC$FTests
#>    DF      MS    F        p
#> T   4 0.00232 13.7 0.000202
#> TR 12 0.00017   NA       NA
#>
#> $RRFC$ciDiffTrt
```

```
#>           Estimate  StdErr DF      t    PrGTt CILower  CIUpper
#> trt1-trt2 -0.00686 0.00921 12 -0.745 4.71e-01 -0.0269  0.01321
#> trt1-trt3  0.03061 0.00921 12  3.324 6.06e-03  0.0106  0.05068
#> trt1-trt4 -0.01604 0.00921 12 -1.741 1.07e-01 -0.0361  0.00403
#> trt1-trt5  0.03884 0.00921 12  4.218 1.19e-03  0.0188  0.05891
#> trt2-trt3  0.03747 0.00921 12  4.069 1.56e-03  0.0174  0.05754
#> trt2-trt4 -0.00918 0.00921 12 -0.997 3.39e-01 -0.0292  0.01089
#> trt2-trt5  0.04570 0.00921 12  4.963 3.29e-04  0.0256  0.06576
#> trt3-trt4 -0.04665 0.00921 12 -5.066 2.77e-04 -0.0667 -0.02659
#> trt3-trt5  0.00823 0.00921 12  0.894 3.89e-01 -0.0118  0.02829
#> trt4-trt5  0.05488 0.00921 12  5.959 6.62e-05  0.0348  0.07494
#>
#> $RRFC$ciAvgRdrEachTrt
#>      Estimate StdErr DF CILower CIUpper
#> Trt1    0.753 0.0235  3   0.678   0.828
#> Trt2    0.760 0.0207  3   0.694   0.826
#> Trt3    0.723 0.0207  3   0.657   0.788
#> Trt4    0.769 0.0311  3   0.670   0.868
#> Trt5    0.714 0.0273  3   0.627   0.801
```

8.6.1 Random-Reader Random-Case (RRRC) analysis

- `st3$RRRC$FTests` contains the results of the F-tests.
- The p-value is much smaller than that obtained after converting to an ROC dataset. Specifically, for FROC analysis, the p-value is $1.17105004 \times 10^{-4}$ while that for ROC analysis is 0.03054456. The F-statistic and the `ddf` are both larger for FROC analysis, both of of which result in increased probability of rejecting the NH, i.e., FROC analysis has greater power than ROC analysis.
- The increased power of FROC analysis has been confirmed in simulation studies (Chakraborty, 2002).
- `st3$RRRC$ciDiffTrt` contains the confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- With I = 5 treatments there are 10 distinct treatment-pairings.
- Looking at the `PrGTt` (for probability greater than `t`) column, one finds six pairings that are significant: `trt1-trt3`, `trt1-trt5`, etc. The smallest p-value is for the `trt4-trt5` pairing. The findings are consistent with the prior ROC analysis, the difference being the smaller p-values.
- `st3$RRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}}$.

- Looking at the `Estimate` column one confirms that `trt5` has the smallest FOM while `trt4` has the highest (the `Estimates` column is identical for RRRC, FRRC and RRFC analyses).
- `st3$RRRC$st1$RRRC$ciDiffTrtEachRdr` contains confidence intervals for inter-treatment difference FOMs, for each reader, i.e., $CI_{1-\alpha,RRRC,\theta_{ij}-\theta_{i'j}}$.

8.6.2 Fixed-Reader Random-Case (FRRC) analysis

- `st3$FRRC$FTests` contains results of the F-test of the NH.
- Again, `ndf` = 4 because there are I = 5 treatments. Since the p-value is less than 0.05, at least one treatment-pairing FOM difference is significantly different from zero.
- `st3$FRRC$ciDiffTrt` contains the confidence intervals for the inter-treatment paired difference FOMs averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- With I = 5 treatments there are 10 distinct treatment-pairings.
- Looking at the `PrGTt` (for probability greater than `t`) column, one finds six pairings that are significant: `trt1-trt3`, `trt1-trt5`, etc. The smallest p-value is for the `trt4-trt5` pairing. The findings are consistent with the prior ROC analysis, the difference being the smaller p-values.
- `st3$FRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}}$.
- Looking at the `Estimate` column one confirms that `trt5` has the smallest FOM while `trt4` has the highest.
- `st3$FRRC$st1$FRRC$ciDiffTrtEachRdr` contains confidence intervals for inter-treatment difference FOMs, for each reader, i.e., $CI_{1-\alpha,FRRC,\theta_{ij}-\theta_{i'j}}$.

8.6.3 Random-Reader Fixed-Case (RRFC) analysis

- `st3$RRFC$FTests` contains results of the F-test of the NH.
- Again, `ndf` = 4 because there are I = 5 treatments. Since the p-value is less than 0.05, at least one treatment-pairing FOM difference is significantly different from zero.
- `st3$RRFC$ciDiffTrt` contains confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- `st3$RRFC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}}$.

- The `Estimate` column confirms that `trt5` has the smallest FOM while `trt4` has the highest (the `Estimates` column is identical for RRRC, FRRC and RRFC analyses).

8.7 RJafroc: dataset04, FROC/DBM

- The fourth example again uses `dataset04`, i.e., FROC data, *but this time using DBM analysis*.
- The key difference below is in the call to `StSignificanceTesting()` function, where we set `method = "DBM"`.
- Since DBM analysis is pseudovalue based, and the figure of merit is not the empirical AUC under the ROC, one expects to see differences from the previously presented OR analysis, contained in `st3`.

```
st4 <- StSignificanceTesting(ds, FOM = FOM, method = "DBM")
# Note: using DBM analysis
print(st4, digits = 3)
#> $FOMs
#> $FOMs$foms
#>       rdr1  rdr3  rdr4  rdr5
#> trt1 0.779 0.725 0.704 0.805
#> trt2 0.787 0.727 0.723 0.804
#> trt3 0.730 0.716 0.672 0.773
#> trt4 0.810 0.743 0.694 0.829
#> trt5 0.749 0.682 0.655 0.771
#>
#> $FOMs$trtMeans
#>      Estimate
#> trt1    0.753
#> trt2    0.760
#> trt3    0.723
#> trt4    0.769
#> trt5    0.714
#>
#> $FOMs$trtMeanDiffs
#>            Estimate
#> trt1-trt2 -0.00686
#> trt1-trt3  0.03061
#> trt1-trt4 -0.01604
#> trt1-trt5  0.03884
#> trt2-trt3  0.03747
#> trt2-trt4 -0.00918
#> trt2-trt5  0.04570
#> trt3-trt4 -0.04665
```

```
#> trt3-trt5  0.00823
#> trt4-trt5  0.05488
#>
#>
#> $ANOVA
#> $ANOVA$TRCanova
#>             SS   DF     MS
#> T       1.853    4 0.4633
#> R       7.081    3 2.3603
#> C     289.602  199 1.4553
#> TR      0.407   12 0.0339
#> TC     95.772  796 0.1203
#> RC    126.902  597 0.2126
#> TRC   226.479 2388 0.0948
#> Total 748.096 3999     NA
#>
#> $ANOVA$VarCom
#>        Estimates
#> VarR    0.002209
#> VarC    0.060862
#> VarTR  -0.000305
#> VarTC   0.006369
#> VarRC   0.023545
#> VarErr  0.094841
#>
#> $ANOVA$IndividualTrt
#>       DF  Trt1  Trt2  Trt3  Trt4  Trt5
#> msR    3 0.442 0.343 0.342 0.772 0.597
#> msC  199 0.375 0.416 0.372 0.358 0.415
#> msRC 597 0.109 0.112 0.134 0.110 0.127
#>
#> $ANOVA$IndividualRdr
#>       DF   rdr1   rdr3  rdr4   rdr5
#> msT    4 0.2027 0.1019 0.140 0.1208
#> msC  199 0.5064 0.5278 0.630 0.4285
#> msTC 796 0.0942 0.0922 0.135 0.0833
#>
#>
#> $RRRC
#> $RRRC$FTests
#>             DF     MS FStat        p
#> Treatment  4.0 0.4633   7.8 0.000117
#> Error     36.8 0.0594    NA       NA
#>
#> $RRRC$ciDiffTrt
```

```
#>           Estimate StdErr   DF      t    PrGTt  CILower  CIUpper
#> trt1-trt2 -0.00686 0.0122 36.8 -0.563 5.77e-01 -0.03155  0.01784
#> trt1-trt3  0.03061 0.0122 36.8  2.512 1.65e-02  0.00592  0.05531
#> trt1-trt4 -0.01604 0.0122 36.8 -1.316 1.96e-01 -0.04073  0.00866
#> trt1-trt5  0.03884 0.0122 36.8  3.188 2.92e-03  0.01415  0.06354
#> trt2-trt3  0.03747 0.0122 36.8  3.075 3.96e-03  0.01278  0.06217
#> trt2-trt4 -0.00918 0.0122 36.8 -0.753 4.56e-01 -0.03387  0.01552
#> trt2-trt5  0.04570 0.0122 36.8  3.750 6.07e-04  0.02100  0.07040
#> trt3-trt4 -0.04665 0.0122 36.8 -3.828 4.85e-04 -0.07135 -0.02195
#> trt3-trt5  0.00823 0.0122 36.8  0.675 5.04e-01 -0.01647  0.03292
#> trt4-trt5  0.05488 0.0122 36.8  4.504 6.52e-05  0.03018  0.07957
#>
#> $RRRC$ciAvgRdrEachTrt
#>      Estimate StdErr    DF CILower CIUpper
#> trt1    0.753 0.0298  7.71   0.684   0.822
#> trt2    0.760 0.0284 10.69   0.697   0.823
#> trt3    0.723 0.0269  8.62   0.661   0.784
#> trt4    0.769 0.0357  5.24   0.679   0.860
#> trt5    0.714 0.0333  6.59   0.635   0.794
#>
#>
#> $FRRC
#> $FRRC$FTests
#>            DF    MS FStat       p
#> Treatment   4 0.463  3.85 0.00416
#> Error     796 0.120    NA      NA
#>
#> $FRRC$ciDiffTrt
#>           Estimate StdErr  DF      t   PrGTt  CILower CIUpper
#> trt1-trt2 -0.00686 0.0173 796 -0.395 0.69271 -0.04090  0.0272
#> trt1-trt3  0.03061 0.0173 796  1.765 0.07791 -0.00343  0.0647
#> trt1-trt4 -0.01604 0.0173 796 -0.925 0.35546 -0.05008  0.0180
#> trt1-trt5  0.03884 0.0173 796  2.240 0.02539  0.00480  0.0729
#> trt2-trt3  0.03747 0.0173 796  2.161 0.03103  0.00343  0.0715
#> trt2-trt4 -0.00918 0.0173 796 -0.529 0.59677 -0.04322  0.0249
#> trt2-trt5  0.04570 0.0173 796  2.635 0.00858  0.01166  0.0797
#> trt3-trt4 -0.04665 0.0173 796 -2.690 0.00730 -0.08069 -0.0126
#> trt3-trt5  0.00823 0.0173 796  0.474 0.63528 -0.02581  0.0423
#> trt4-trt5  0.05488 0.0173 796  3.164 0.00161  0.02084  0.0889
#>
#> $FRRC$ciAvgRdrEachTrt
#>      Estimate StdErr  DF CILower CIUpper
#> trt1    0.753 0.0217 199   0.711   0.796
#> trt2    0.760 0.0228 199   0.715   0.805
#> trt3    0.723 0.0216 199   0.680   0.765
```

```
#> trt4    0.769 0.0212 199   0.728   0.811
#> trt5    0.714 0.0228 199   0.669   0.759
#>
#> $FRRC$ciDiffTrtEachRdr
#>                 Estimate StdErr  DF       t   PrGTt   CILower   CIUpper
#> rdr1::trt1-trt2 -0.00773 0.0307 199 -0.2520 0.80131 -0.068250  0.052784
#> rdr1::trt1-trt3  0.04957 0.0307 199  1.6154 0.10781 -0.010942  0.110092
#> rdr1::trt1-trt4 -0.03087 0.0307 199 -1.0058 0.31573 -0.091384  0.029650
#> rdr1::trt1-trt5  0.03047 0.0307 199  0.9928 0.32203 -0.030050  0.090984
#> rdr1::trt2-trt3  0.05731 0.0307 199  1.8674 0.06332 -0.003209  0.117825
#> rdr1::trt2-trt4 -0.02313 0.0307 199 -0.7538 0.45186 -0.083650  0.037384
#> rdr1::trt2-trt5  0.03820 0.0307 199  1.2448 0.21469 -0.022317  0.098717
#> rdr1::trt3-trt4 -0.08044 0.0307 199 -2.6212 0.00944 -0.140959 -0.019925
#> rdr1::trt3-trt5 -0.01911 0.0307 199 -0.6226 0.53423 -0.079625  0.041409
#> rdr1::trt4-trt5  0.06133 0.0307 199  1.9986 0.04702  0.000816  0.121850
#> rdr3::trt1-trt2 -0.00201 0.0304 199 -0.0661 0.94733 -0.061885  0.057868
#> rdr3::trt1-trt3  0.00913 0.0304 199  0.3008 0.76389 -0.050743  0.069010
#> rdr3::trt1-trt4 -0.01822 0.0304 199 -0.6002 0.54904 -0.078102  0.041652
#> rdr3::trt1-trt5  0.04262 0.0304 199  1.4035 0.16202 -0.017260  0.102493
#> rdr3::trt2-trt3  0.01114 0.0304 199  0.3669 0.71406 -0.048735  0.071018
#> rdr3::trt2-trt4 -0.01622 0.0304 199 -0.5341 0.59389 -0.076093  0.043660
#> rdr3::trt2-trt5  0.04462 0.0304 199  1.4697 0.14323 -0.015252  0.104502
#> rdr3::trt3-trt4 -0.02736 0.0304 199 -0.9010 0.36867 -0.087235  0.032518
#> rdr3::trt3-trt5  0.03348 0.0304 199  1.1027 0.27148 -0.026393  0.093360
#> rdr3::trt4-trt5  0.06084 0.0304 199  2.0037 0.04645  0.000965  0.120718
#> rdr4::trt1-trt2 -0.01899 0.0368 199 -0.5166 0.60600 -0.091485  0.053502
#> rdr4::trt1-trt3  0.03132 0.0368 199  0.8519 0.39531 -0.041177  0.103810
#> rdr4::trt1-trt4  0.00927 0.0368 199  0.2521 0.80125 -0.063227  0.081760
#> rdr4::trt1-trt5  0.04845 0.0368 199  1.3179 0.18904 -0.024044  0.120944
#> rdr4::trt2-trt3  0.05031 0.0368 199  1.3685 0.17271 -0.022185  0.122802
#> rdr4::trt2-trt4  0.02826 0.0368 199  0.7687 0.44300 -0.044235  0.100752
#> rdr4::trt2-trt5  0.06744 0.0368 199  1.8345 0.06807 -0.005052  0.139935
#> rdr4::trt3-trt4 -0.02205 0.0368 199 -0.5998 0.54932 -0.094544  0.050444
#> rdr4::trt3-trt5  0.01713 0.0368 199  0.4661 0.64168 -0.055360  0.089627
#> rdr4::trt4-trt5  0.03918 0.0368 199  1.0659 0.28778 -0.033310  0.111677
#> rdr5::trt1-trt2  0.00131 0.0289 199  0.0453 0.96389 -0.055610  0.058227
#> rdr5::trt1-trt3  0.03243 0.0289 199  1.1237 0.26251 -0.024485  0.089352
#> rdr5::trt1-trt4 -0.02432 0.0289 199 -0.8425 0.40055 -0.081235  0.032602
#> rdr5::trt1-trt5  0.03384 0.0289 199  1.1724 0.24242 -0.023077  0.090760
#> rdr5::trt2-trt3  0.03112 0.0289 199  1.0783 0.28219 -0.025794  0.088044
#> rdr5::trt2-trt4 -0.02563 0.0289 199 -0.8878 0.37573 -0.082544  0.031294
#> rdr5::trt2-trt5  0.03253 0.0289 199  1.1271 0.26105 -0.024385  0.089452
#> rdr5::trt3-trt4 -0.05675 0.0289 199 -1.9661 0.05068 -0.113669  0.000169
#> rdr5::trt3-trt5  0.00141 0.0289 199  0.0488 0.96113 -0.055510  0.058327
#> rdr5::trt4-trt5  0.05816 0.0289 199  2.0149 0.04526  0.001240  0.115077
```

```
#>
#>
#> $RRFC
#> $RRFC$FTests
#>           DF     MS FStat        p
#> Treatment  4 0.4633  13.7 0.000202
#> Error     12 0.0339    NA       NA
#>
#> $RRFC$ciDiffTrt
#>           Estimate  StdErr DF      t    PrGTt CILower  CIUpper
#> trt1-trt2 -0.00686 0.00921 12 -0.745 4.71e-01 -0.0269  0.01321
#> trt1-trt3  0.03061 0.00921 12  3.324 6.06e-03  0.0106  0.05068
#> trt1-trt4 -0.01604 0.00921 12 -1.741 1.07e-01 -0.0361  0.00403
#> trt1-trt5  0.03884 0.00921 12  4.218 1.19e-03  0.0188  0.05891
#> trt2-trt3  0.03747 0.00921 12  4.069 1.56e-03  0.0174  0.05754
#> trt2-trt4 -0.00918 0.00921 12 -0.997 3.39e-01 -0.0292  0.01089
#> trt2-trt5  0.04570 0.00921 12  4.963 3.29e-04  0.0256  0.06576
#> trt3-trt4 -0.04665 0.00921 12 -5.066 2.77e-04 -0.0667 -0.02659
#> trt3-trt5  0.00823 0.00921 12  0.894 3.89e-01 -0.0118  0.02829
#> trt4-trt5  0.05488 0.00921 12  5.959 6.62e-05  0.0348  0.07494
#>
#> $RRFC$ciAvgRdrEachTrt
#>      Estimate StdErr DF CILower CIUpper
#> trt1    0.753 0.0235  3   0.678   0.828
#> trt2    0.760 0.0207  3   0.694   0.826
#> trt3    0.723 0.0207  3   0.657   0.788
#> trt4    0.769 0.0311  3   0.670   0.868
#> trt5    0.714 0.0273  3   0.627   0.801
```

8.7.1 Random-Reader Random-Case (RRRC) analysis

- `st4$RRRC$FTests` contains the results of the F-test of the NH.
- `st4$RRRC$ciDiffTrt` contains the confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- `st4$RRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}}$.

8.7.2 Fixed-Reader Random-Case (FRRC) analysis

- `st4$FRRC$FTests` contains results of the F-test of the NH, which is actually a chi-square statistic.

- `st4$FRRC$ciDiffTrt` contains confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- With I = 5 treatments there are 10 distinct treatment-pairings.
- Looking at the `PrGTt` (for probability greater than `t`) column, one finds six pairings that are significant: `trt1-trt3`, `trt1-trt5`, etc. The smallest p-value is for the `trt4-trt5` pairing. The findings are consistent with the prior ROC analysis, the difference being the smaller p-values.
- `st4$FRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}}$.
- `st4$FRRC$ciDiffTrtEachRdr` contains confidence intervals for inter-treatment difference FOMs, for each reader, i.e., $CI_{1-\alpha,FRRC,\theta_{ij}-\theta_{i'j}}$.

8.7.3 Random-Reader Fixed-Case (RRFC) analysis

- `st4$RRFC$FTests` contains the results of the F-test of the NH.
- `st4$RRFC$ciDiffTrt` contains confidence intervals for the inter-treatment paired difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- `st4$RRFC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}}$.
- The `Estimate` column confirms that `trt5` has the smallest FOM while `trt4` has the highest (the `Estimates` column is identical for RRRC, FRRC and RRFC analyses).

8.8 Summary

8.9 Discussion

8.10 Tentative

```
ds1 <- dataset04 # do NOT convert to ROC
# comment/uncomment following code to disable/enable unequal weights
# K2 <- length(ds1$ratings$LL[1,1,,1])
# weights <- array(dim = c(K2, max(ds1$lesions$perCase)))
# perCase <- ds1$lesions$perCase
# for (k2 in 1:K2) {
#   sum <- 0
```

```
#   for (el in 1:perCase[k2]) {
#     weights[k2,el] <- 1/el
#     sum <- sum + 1/el
#   }
#   weights[k2,1:perCase[k2]] <- weights[k2,1:perCase[k2]] / sum
# }
# ds1$lesions$weights <- weights
ds <- ds1
FOM <- "wAFROC" # also try wAFROC1, MaxLLF and MaxNLF
st5 <- StSignificanceTesting(ds, FOM = FOM, method = "OR")
print(st5, digits = 4)
```

A comparison was run between results of OR and DBM for the FROC dataset. Except for `FRRC`, where differences are expected (because `ddf` in the former is ∞, while that in the later is $(I-1)\times(J-1)$), the results for the p-values were identical. This was true for the following FOMs: `wAFROC`, with equal and unequal weights, and `MaxLLF`. The confidence intervals (again, excluding `FRRC`) were identical for `FOM` = `wAFROC`. Slight differences were observed for `FOM` = `MaxLLF`.

8.11 Chapter References

Chapter 9

Sample size estimation: DBM method

9.1 TBA How much finished

80%

9.2 Introduction

The question addressed here is "how many readers and cases", usually abbreviated to "sample-size", should one employ to conduct a "well-planned" ROC study. The reasons for the quotes around "well-planned" will shortly become clear. If cost were no concern, the reply would be: "as many readers and cases as one can get". There are other causes affecting sample-size, e.g., the data collection paradigm and analysis, however, this chapter is restricted to the MRMC ROC data collection paradigm, with data analyzed by the DBM method described in a previous chapter. The next chapter will deal with data analyzed by the OR method.

It turns out that provided one can specify conceptually valid effect-sizes between different paradigms (i.e., in the same "units"), the methods described in this chapter are extensible to other paradigms; see TBA Chapter 19 for sample size estimation for FROC studies. *For this reason it is important to understand the concepts of sample-size estimation in the simpler ROC context.*

For simplicity and practicality, this chapter, and the next, is restricted to analysis of two-treatment data ($I = 2$). The purpose of most imaging system assessment studies is to determine, for a given diagnostic task, whether radiologists perform better using a new treatment over the conventional treatment, and

whether the difference is statistically significant. Therefore, the two-treatment case is the most common one encountered. While it is possible to extend the methods to more than two treatments, the extensions are not, in my opinion, clinically interesting.

Assume the figure of merit (FOM) θ is chosen to be the area AUC under the ROC curve (empirical or fitted is immaterial as far as the formulae are concerned; however, the choice will affect statistical power). The statistical analysis determines the significance level of the study, i.e., the probability or p-value for incorrectly rejecting the null hypothesis (NH) that the two θs are equal: $NH : \theta_1 = \theta_2$, where the subscripts refer to the two treatments and the bullet represents the average over the reader index. If the p-value is smaller than a pre-specified α, typically set at 5%, one rejects the NH and declares the treatments different at the α significance level. Statistical power is the probability of correctly rejecting the null hypothesis when the alternative hypothesis $AH : \theta_1 \neq \theta_2$ is true, (TBA Chapter 08).

The value of the *true* difference between the treatments, known as the *true effect-size* is, of course, unknown. If it were known, there would be no need to conduct the ROC study. One would simply adopt the treatment with the higher θ. Sample-size estimation involves making an educated guess regarding the true effect-size , called the *anticipated effect size*, and denoted by d. To quote Harold Kundel (ICRU, 1996): "any calculation of power amounts to specification of the anticipated effect-size". Increasing the anticipated effect size will increase statistical power but may represent an unrealistic expectation of the true difference between the treatments, in the sense that it overestimates the ability of technology to achieve this much improvement. Conversely, an unduly small d might be clinically insignificant, besides requiring a very large sample-size to achieve sufficient statistical power.

Statistical power depends on the magnitude of d divided by the standard deviation $\sigma(d)$ of d, i.e. $D = \frac{|d|}{\sigma(d)}$. The sign is relevant as it determines whether the project is worth pursuing at all (see TBA §11.8.4). The ratio is termed (Cohen, 1988) Cohen's D. When this signal-to-noise-ratio-like quantity is large, statistical power approaches 100%. Reader and case variability and data correlations determine $\sigma(d)$. No matter how small the anticipated d, as long as it is finite, then, using sufficiently large numbers of readers and cases $\sigma(d)$ can be made sufficiently small to achieve near 100% statistical power. Of course, a very small effect-size may not be clinically significant. There is a key difference between *statistical significance* and *clinical significance*. An effect-size in AUC units could be so small, e.g., 0.001, as to be clinically insignificant, but by employing a sufficiently large sample size one could design a study to detect this small - and clinically meaningless - difference with near unit probability, i.e., high statistical power.

What determines clinical significance? A small effect-size, e.g., 0.01 AUC units, could be clinically significant if it applies to a large population, where the small benefit in detection rate is amplified by the number of patients benefiting from

the new treatment. In contrast, for an "orphan" disease, i.e., one with very low prevalence, an effect-size of 0.05 might not be enough to justify the additional cost of the new treatment. The improvement might have to be 0.1 before it is worth it for a new treatment to be brought to market. One hates to monetize life and death issues, but there is no getting away from it, as cost/benefit issues determine clinical significance. The arbiters of clinical significance are engineers, imaging scientists, clinicians, epidemiologists, insurance companies and those who set government health care policies. The engineers and imaging scientists determine whether the effect-size the clinicians would like is feasible from technical and scientific viewpoints. The clinician determines, based on incidence of disease and other considerations, e.g., altruistic, malpractice, cost of the new device and insurance reimbursement, what effect-size is justifiable. Cohen has suggested that d values of 0.2, 0.5, and 0.8 be considered small, medium, and large, respectively, but he has also argued against their indiscriminate usage. However, after a study is completed, clinicians often find that an effect-size that biostatisticians label as small may, in certain circumstances, be clinically significant and an effect-size that they label as large may in other circumstances be clinically insignificant. Clearly, this is a complex issue. Some suggestions on choosing a clinically significant effect size are made in (TBA §11.12).

Having developed a new imaging modality the R&D team wishes to compare it to the existing standard with the short-term goal of making a submission to the FDA to allow them to perform pre-market testing of the device. The long-term goal is to commercialize the device. Assume the R&D team has optimized the device based on physical measurements, (TBA Chapter 01), perhaps supplemented with anecdotal feedback from clinicians based on a few images. Needed at this point is a pilot study. A pilot study, conducted with a relatively small and practical sample size, is intended to provide estimates of different sources of variability and correlations. It also provides an initial estimate of the effect-size, termed the *observed effect-size*, d. Based on results from the pilot the sample-size tools described in this chapter permit estimation of the numbers of readers and cases that will reduce $\sigma(d)$ sufficiently to achieve the desired power for the larger "pivotal" study. [A distinction could be made in the notation between observed and anticipated effect sizes, but it will be clear from the context. Later, it will be shown how one can make an educated guess about the anticipated effect size from an observed effect size.]

This chapter is concerned with multiple-reader MRMC studies that follow the fully crossed factorial design meaning that each reader interprets a common case-set in all treatments. Since the resulting pairings (i.e., correlations) tend to decrease $\sigma(d)$ (since the variations occur in tandem, they tend to cancel out in the difference, see (TBA Chapter 09, Introduction), for Dr. Robert Wagner's sailboat analogy) it yields more statistical power compared to an unpaired design, and consequently this design is frequently used. Two sample-size estimation procedures for MRMC are the Hillis-Berbaum method (Hillis and Berbaum, 2004) and the Obuchowski-Rockette (Obuchowski, 1998) method. With recent work by Hillis, the two methods have been shown to be substantially equivalent.

This chapter will focus on the DBM approach. Since it is based on a standard ANOVA model, it is easier to extend the NH testing procedure described in Chapter 09 to the alternative hypothesis, which is relevant for sample size estimation. [TBA Online Appendix 11.A shows how to translate the DBM formulae to the OR method (Hillis et al., 2011).]

Given an effect-size, and choosing this wisely is the most difficult part of the process, the method described in this chapter uses pseudovalue variance components estimated by the DBM method to predict sample-sizes (i.e., different combinations of numbers of readers and cases) necessary to achieve a desired power.

9.3 Statistical Power

The concept of statistical power was introduced in [TBA Chapter 08] but is worth repeating. There are two possible decisions following a test of a null hypothesis (NH): reject or fail to reject the NH. Each decision is associated with a probability on an erroneous conclusion. If the NH is true and one rejects it, the probability of the ensuing Type-I error is denoted α. If the NH is false and one fails to reject it, the probability of the ensuing Type II- error is denoted β. Statistical power is the complement of β, i.e.,

$$Power = 1 - \beta \tag{9.1}$$

Typically, one aims for $\beta = 0.2$ or less, i.e., a statistical power of 80% or more. Like $\alpha = 0.05$, this is a *convention* and more nuanced cost-benefit considerations may cause the researcher to adopt a different value.

9.3.1 Observed vs. anticipated effect-size

Assuming no other similar studies have already been conducted with the treatments in question, the observed effect-size, although "merely an estimate", is the best information available at the end of the pilot study regarding the value of the true effect-size. From the two previous chapters one knows that the significance testing software will report not only the observed effect-size, but also a 95% confidence interval associate with it. It will be shown later how one can use this information to make an educated guess regarding the value of the anticipated effect-size.

9.3.2 Dependence of statistical power on estimates of model parameters

Examination of the expression for , Eqn. (11.5), shows that statistical power increases if:

- The numerator is large. This occurs if: (a) the anticipated effect-size d is large. Since effect-size enters as the *square*, TBA Eqn. (11.8), it is has a particularly strong effect; (b) If $J \times K$ is large. Both of these results should be obvious, as a large effect size and a large sample size should result in increased probability of rejecting the NH.
- The denominator is small. The first term in the denominator is $(\sigma_\epsilon^2 + \sigma_{\tau RC}^2)$. These two terms cannot be separated. This is the residual variability of the jackknife pseudovalues. It should make sense that the smaller the variability, the larger is the non-centrality parameter and the statistical power.
- The next term in the denominator is $K\sigma_{\tau R}^2$, the treatment-reader variance component multiplied by the total number of cases. The reader variance σ_R^2 has no effect on statistical power, because it has an equal effect on both treatments and cancels out in the difference. Instead, it is the treatment-reader variance σ_R^2 that contributes "noise" tending to confound the estimate of the effect-size.
- The variance components estimated by the ANOVA procedure are realizations of random variables and as such subject to noise (there actually exists a beast such as variance of a variance). The presence of the K term, usually large, can amplify the effect of noise in the estimate of σ_R^2, making the sample size estimation procedure less accurate.
- The final term in the denominator is $J\sigma_{\tau C}^2$. The variance σ_C^2 has no impact on statistical power, as it cancels out in the difference. The treatment-case variance component introduces "noise" into the estimate of the effect size, thereby decreasing power. Since it is multiplied by J, the number of readers, and typically $J << K$, the error amplification effect on accuracy of the sample size estimate is not as bad as with the treatment-reader variance component.
- Accuracy of sample size estimation, essentially estimating confidence intervals for statistical power, is addressed in (Chakraborty, 2010).

9.3.3 Formulae for random-reader random-case (RRRC) sample size estimation

9.3.4 Significance testing

9.3.5 p-value and confidence interval

9.3.6 Comparing DBM to Obuchowski and Rockette for single-reader multiple-treatments

Having performed a pilot study and planning to perform a pivotal study, sample size estimation follows the following procedure, which assumes that both reader and case are treated as random factors. Different formulae, described later, apply when either reader or case is treated as a fixed factor.

- Perform DBM analysis on the pilot data. This yields the observed effect size as well as estimates of all relevant variance components and mean squares appearing in TBA Eqn. (11.5) and Eqn. (11.7).
- This is the difficult but critical part: make an educated guess regarding the effect-size, d, that one is interested in "detecting" (i.e., hoping to reject the NH with probability $1 - \beta$). The author prefers the term "anticipated" effect-size to "true" effect-size (the latter implies knowledge of the true difference between the modalities which, as noted earlier, would obviate the need for a pivotal study).
- Two scenarios are considered below. In the first scenario, the effect-size is assumed equal to that observed in the pilot study, i.e., $d = d_{obs}$.
- In the second, so-called "best-case" scenario, one assumes that the anticipate value of d is the observed value plus two-sigma of the confidence interval, in the correct direction, of course, i.e., $d = |d_{obs}| + 2\sigma$. Here σ is one-fourth the width of the 95% confidence interval for d_{obs}. Anticipating more than 2σ greater than the observed effect-size would be overly optimistic. The width of the CI implies that chances are less than 2.5% that the anticipated value is at or beyond the overly optimistic value. These points will become clearer when example datasets are analyzed below.
- Calculate statistical power using the distribution implied by Eqn. (11.4), to calculate the probability that a random value of the relevant F-statistic will exceed the critical value, as in §11.3.2.
- If power is below the desired or "target" power, one tries successively larger value of J and / or K until the target power is reached.

9.4 Formulae for fixed-reader random-case (FRRC) sample size estimation

It was shown in TBA §9.8.2 that for fixed-reader analysis the non-centrality parameter is defined by:

$$\Delta = \frac{JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2} \tag{9.2}$$

The sampling distribution of the F-statistic under the AH is:

$$F_{AH|R} \equiv \frac{MST}{MSTC} \sim F_{I-1,(I-1)(K-1),\Delta} \tag{9.3}$$

9.4.1 Formulae for random-reader fixed-case (RRFC) sample size estimation

It is shown in TBA §9.9 that for fixed-case analysis the non-centrality parameter is defined by:

$$\Delta = \frac{JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2} \tag{9.4}$$

Under the AH, the test statistic is distributed as a non-central F-distribution as follows:

$$F_{AH|C} \equiv \frac{MST}{MSTR} \sim F_{I-1,(I-1)(J-1),\Delta} \tag{9.5}$$

9.4.2 Fixed-reader random-case (FRRC) analysis TBA

It is a realization of a random variable, so one has some leeway in the choice of anticipated effect size - more on this later. Here J^* and K^* refer to the number of readers and cases in the *pilot* study.

9.4.3 Random-reader fixed-case (RRFC) analysis

9.4.4 Single-treatment multiple-reader analysis

9.5 Discussion/Summary/2

9.6 Chapter References

RJafroc Vignettes

Chapter 10

F-distribution

10.1 TBA How much finished

10%

10.2 Introduction

Since it plays an important role in sample size estimation, it is helpful to examine the behavior of the F-distribution. In the following `ndf` = numerator degrees of freedom, `ddf` = denominator degrees of freedom and `ncp` = non-centrality parameter (i.e., the Δ appearing in Eqn. (11.6) of (Chakraborty, 2017)).

The use of three `R` functions is demonstrated.

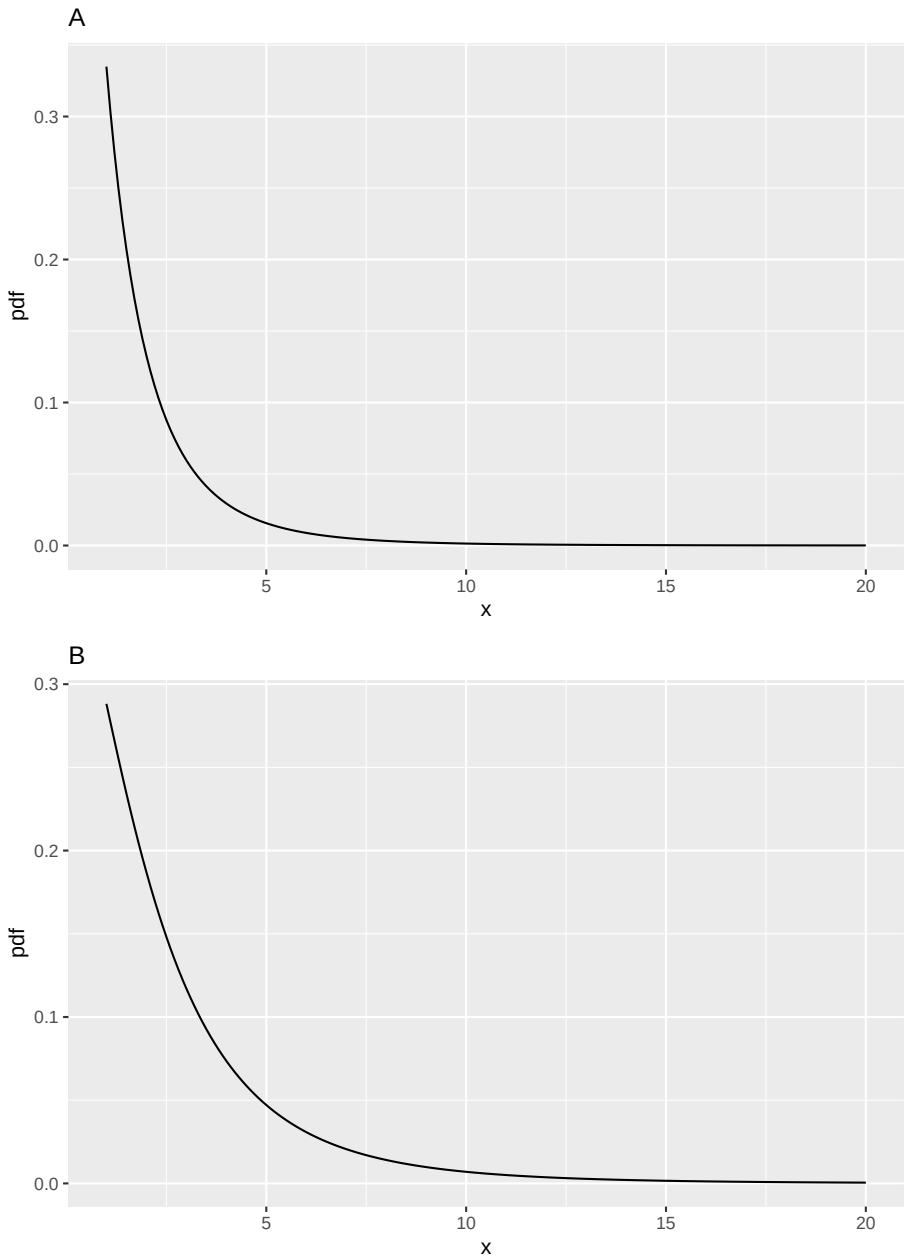
- `qf(p,ndf,ddf)` is the *quantile* function of the F-distribution for specified values of `p`, `ndf` and `ddf`, i.e., the value `x` such that fraction `p` of the area under the F-distribution lies to the right of `x`. Since `ncp` is not included as a parameter, the default value, i.e., zero, is used. This is called the *central* F-distribution.
- `df(x,ndf,ddf,ncp)` is the probability density function (*pdf*) of the F-distribution, as a function of `x`, for specified values of `ndf`, `ddf` and `ncp`.
- `pf(x,ndf,ddf,ncp)` is the probability (or cumulative) distribution function of the F-distribution for specified values of `ndf`, `ddf` and `ncp`.

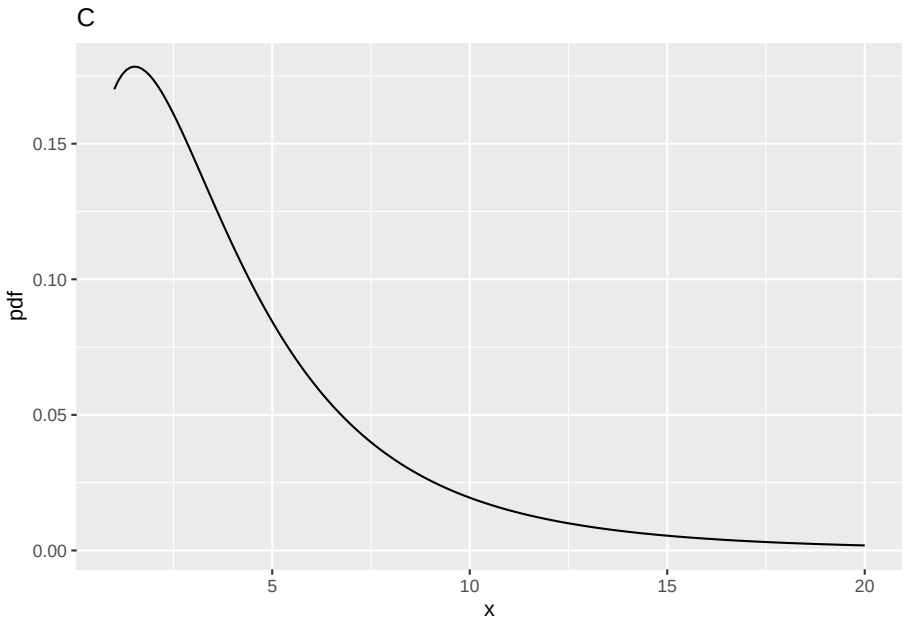
10.3 Effect of `ncp` for `ndf` = 2 and `ddf` = 10

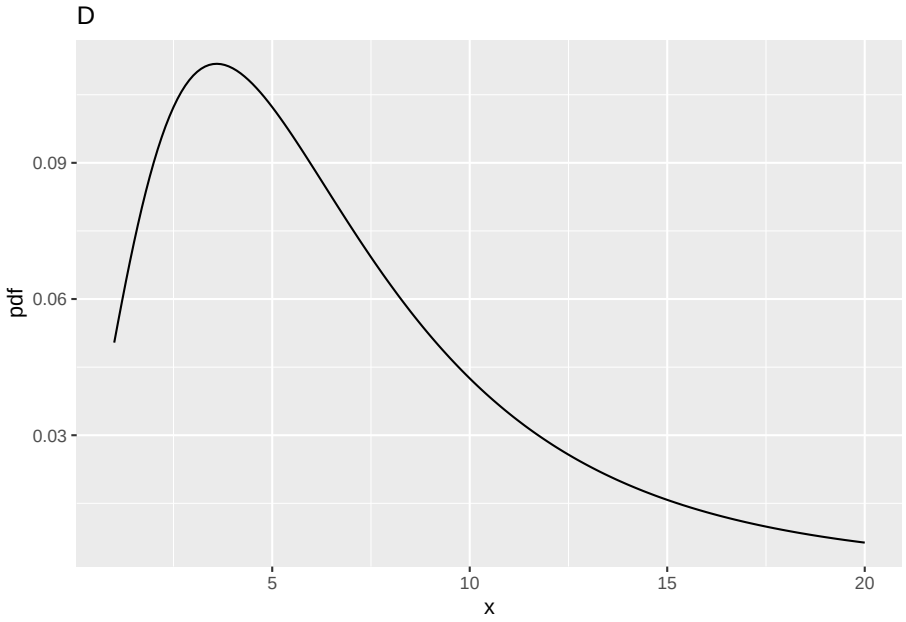
- Four values of `ncp` are considered (0, 2, 5, 10) for `ddf` = 10.

- `fCrit` is the critical value of the F distribution, i.e., that value such that fraction α of the area is to the right of the critical value, i.e., `fCrit` is identical in statistical notation to $F_{1-\alpha,ndf,ddf}$.

```
ndf <- 2;ddf <- 10;ncp <- c(0,2,5,10)
alpha <- 0.05
fCrit <- qf(1-alpha, ndf,ddf)
x <- seq(1, 20, 0.1)
myLabel <- c("A", "B", "C", "D")
myLabelIndx <- 1
pFgtFCrit <- NULL
for (i in 1:length(ncp))
{
  y <- df(x,ndf,ddf,ncp=ncp[i])
  pFgtFCrit <- c(pFgtFCrit, 1-pf(fCrit, ndf, ddf, ncp = ncp[i]))
}
for (i in 1:length(ncp))
{
  y <- df(x,ndf,ddf,ncp=ncp[i])
  curveData <- data.frame(x = x, pdf = y)
  curvePlot <- ggplot(data = curveData, mapping = aes(x = x, y = pdf)) +
    geom_line() +
    ggtitle(myLabel[myLabelIndx]);myLabelIndx <- myLabelIndx + 1
  print(curvePlot)
}
fCrit_2_10 <- fCrit # convention fCrit_ndf_ddf
```


A
0.3
0.2
0.1
0.0
pdf
5
10
15
20
x
B
0.3
0.2
0.1
0.0
pdf
5
10
15
20
x





	ndf	ddf	fCrit	ncp	pFgtFCrit
A	2	10	4.102821	0	0.0500000
B	2	10	4.102821	2	0.1775840
C	2	10	4.102821	5	0.3876841
D	2	10	4.102821	10	0.6769776

10.4 Comments

10.4.1 Fig. A

- This corresponds to `ncp = 0`, i.e., the *central* F-distribution.
- The integral under this distribution is unity (this is also true for all plots in this vignette).
- The critical value, `fCrit` in the above code block, is the value of `x` such that the probability of exceeding `x` is α. The corresponding parameter `alpha` is defined above as 0.05.
- In the current example `fCrit` = 4.102821. Notice the use of the quantile function `qf()` to determine this value, and the default value of `ncp`, namely zero, is used; specifically, one does not pass a 4th argument to `qf()`.
- **The decision rule for rejecting the NH uses the NH distribution of the F-statistic**, i.e., reject the NH if F >= `fCrit`. As expected, `prob > fCrit` = 0.05 because this is how `fCrit` was defined.

10.4.2 Fig. B

- This corresponds to `ncp = 2`, `ndf` = 2 and `ddf` = 10.
- The distribution is slightly shifted to the right as compared to Fig. A, thereby making it more likely that the observed value of the F-statistic will exceed the critical value determined for the NH distribution.
- In fact, `prob > fCrit` = 0.177584, i.e., the *statistical power* (compare this to Fig. A where `prob > fCrit` was 0.05).

10.4.3 Fig. C

- This corresponds to `ncp = 5`, `ndf` = 2 and `ddf` = 10.
- Now `prob > fCrit` = 0.3876841.
- Power has increased compared to Fig. B.

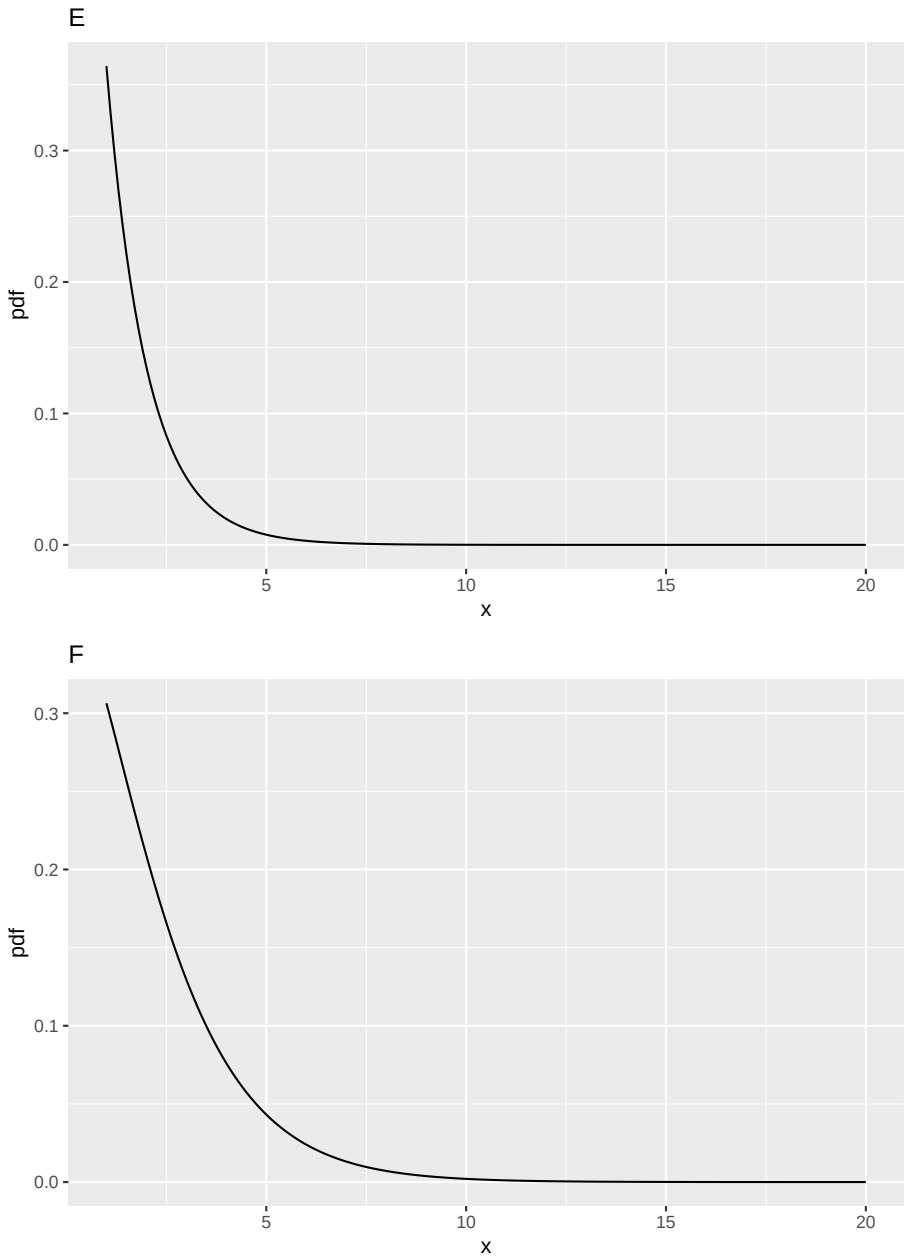
10.4.4 Fig. D

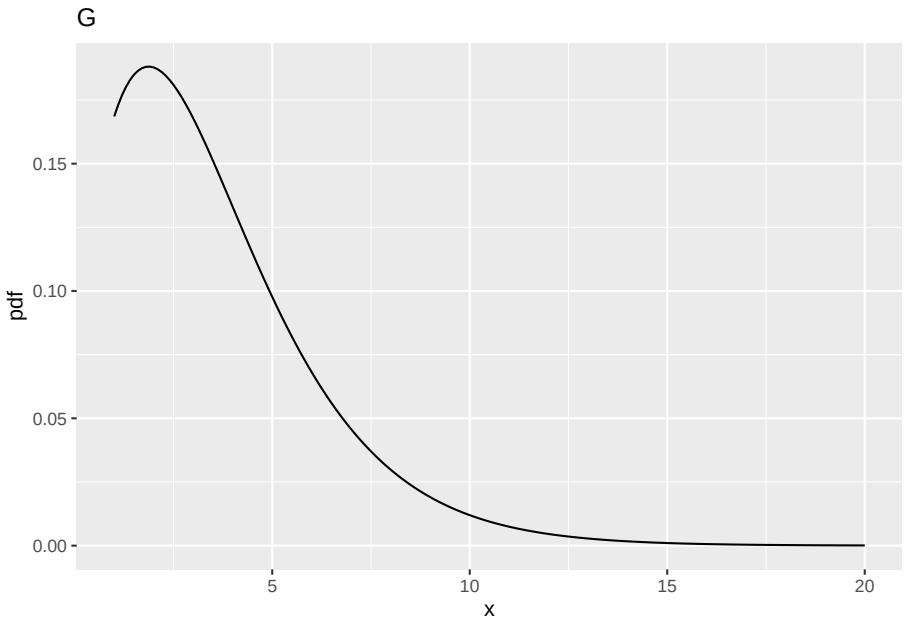
- This corresponds to `ncp = 10`, `ndf` = 2 and `ddf` = 10.
- Now `prob > fCrit` is 0.6769776.
- Power has increased compared to Fig. C.
- The effect of the shift is most obvious in Fig. C and Fig. D.
- Considering a vertical line at `x` = 4.102821, fraction 0.6769776 of the probability distribution in Fig. D lies to the right of this line
- Therefore the NH is likely to be rejected with probability 0.6769776.

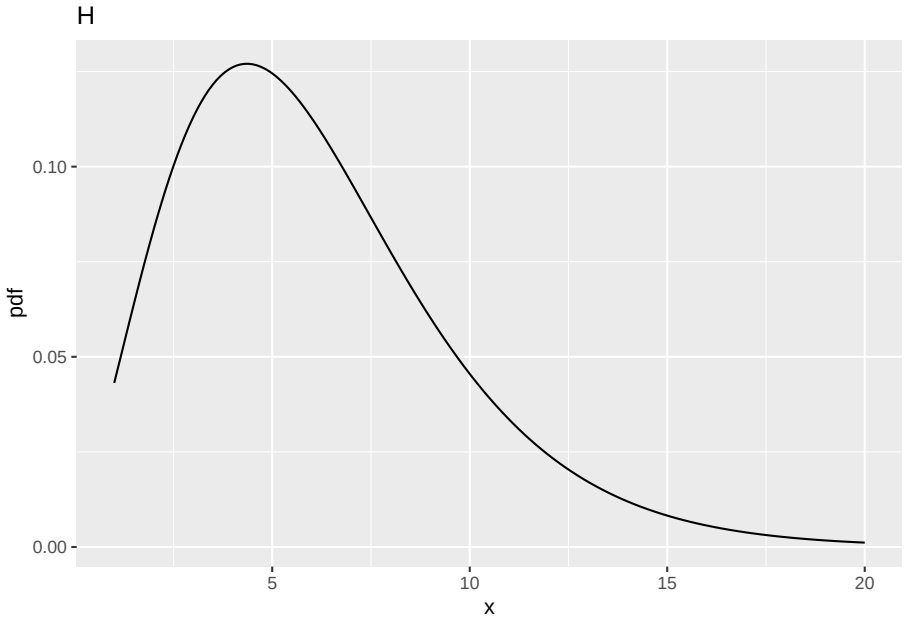
10.4.5 Summary

The larger that non-centrality parameter, the greater the shift to the right of the F-distribution, and the greater the statistical power.

10.5 Effect of ncp for ndf = 2 and ddf = 100




G
0.15
0.10
0.05
0.00
pdf
5
10
15
20
x


H
0.10
0.05
0.00
pdf
5
10
15
20
x

	ndf	ddf	fCrit	ncp	pFgtFCrit
A	2	10	4.102821	0	0.0500000
B	2	10	4.102821	2	0.1775840
C	2	10	4.102821	5	0.3876841
D	2	10	4.102821	10	0.6769776
E	2	100	3.087296	0	0.0500000
F	2	100	3.087296	2	0.2199264
G	2	100	3.087296	5	0.4910802
H	2	100	3.087296	10	0.8029764

10.6 Comments

- All comparisons in this sections are at the same values of `ncp` defined above.
- And between `ddf` = 100 and `ddf` = 10.

10.6.1 Fig. E

- This corresponds to `ncp` = 0, `ndf` = 2 and `ddf` = 100.
- The critical value is `fCrit_2_100` = 3.0872959. Notice the decrease compared to the previous value for `ncp` = 0, i.e., 4.102821, for `ddf` = 10.
- One expects that increasing `ddf` will make it more likely that the NH will be rejected, and this is confirmed below.
- All else equal, statistical power increases with increasing `ddf`.

10.6.2 Fig. F

- This corresponds to `ncp` = 2, `ndf` = 2 and `ddf` = 100.
- The probability of exceeding the critical value is `prob > fCrit_2_100` = 0.2199264, greater than the previous value, i.e., 0.177584 for `ddf` = 10.

10.6.3 Fig. G

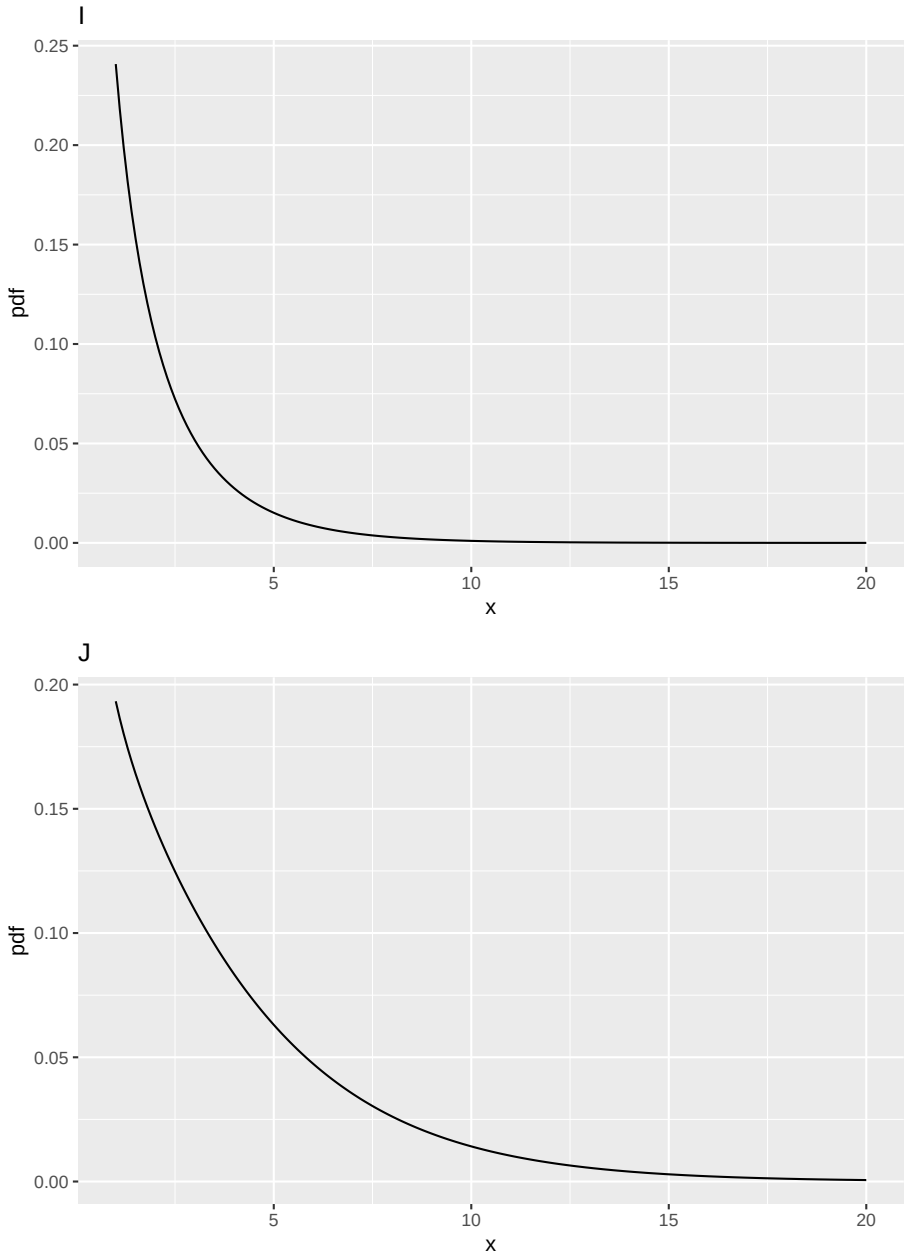
- This corresponds to `ncp = 5`, `ndf` = 2 and `ddf` = 100.
- The probability of exceeding the critical value is `prob > fCrit_2_100` = 0.4910802.
- This is greater than the previous value, i.e., 0.3876841 for `ddf` = 10.

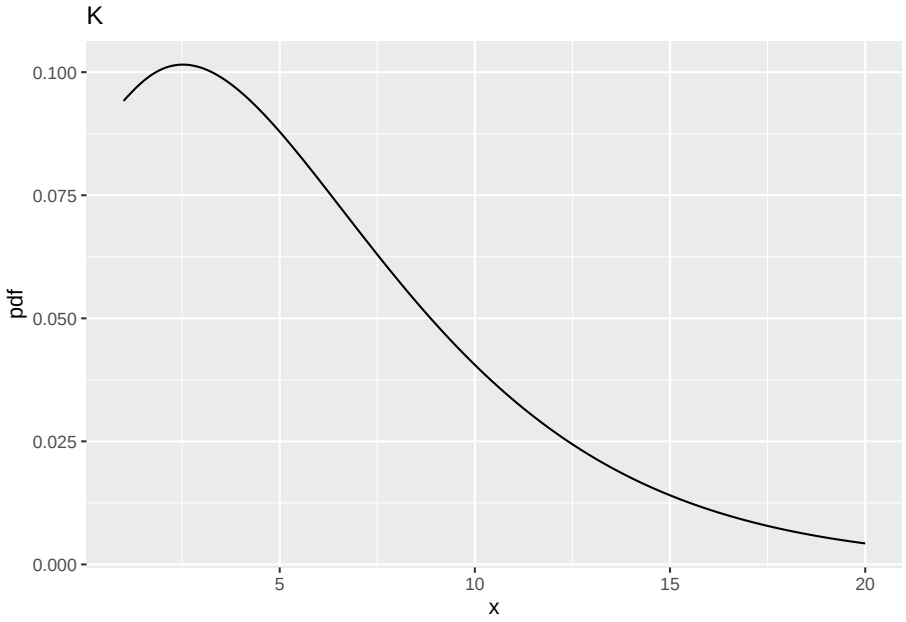
10.6.4 Fig. H

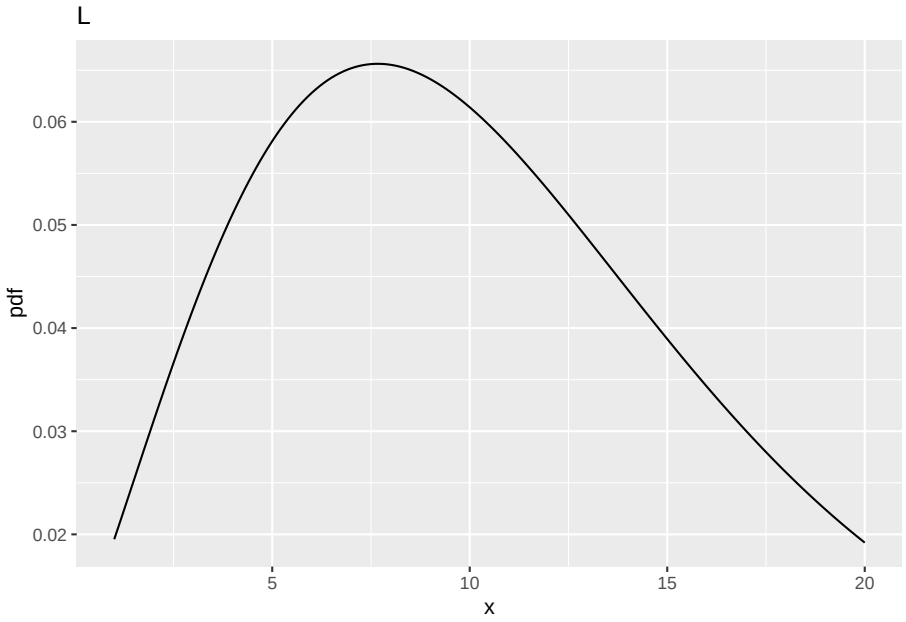
- This corresponds to `ncp = 10`, `ndf` = 2 and `ddf` = 100.

- The probability of exceeding the critical value is `prob > fCrit_2_100` is 0.8029764.
- This is greater than the previous value, i.e., 0.6769776 for `ddf` = 10.

10.7 Effect of ncp for ndf = 1, ddf = 100




K
0.100
0.075
0.050
0.025
0.000
pdf
5
10
15
20
x


L
0.06
0.05
0.04
0.03
0.02
pdf
5
10
15
20
x

	ndf	ddf	fCrit	ncp	pFgtFCrit
A	2	10	4.102821	0	0.0500000
B	2	10	4.102821	2	0.1775840
C	2	10	4.102821	5	0.3876841
D	2	10	4.102821	10	0.6769776
E	2	100	3.087296	0	0.0500000
F	2	100	3.087296	2	0.2199264
G	2	100	3.087296	5	0.4910802
H	2	100	3.087296	10	0.8029764
I	1	100	3.936143	0	0.0500000
J	1	100	3.936143	2	0.2883607
K	1	100	3.936143	5	0.6004962
L	1	100	3.936143	10	0.8793619

10.8 Comments

- All comparisons in this sections are at the same values of `ncp` defined above and at `ddf` = 100.
- And between `ndf` = 1 and `ndf` = 2.

10.8.1 Fig. I

- This corresponds to `ncp` = 0, `ndf` = 1 and `ddf` = 100.
- The critical value is `fCrit_1_100` = 3.936143.
- Notice the increase in the critical value as compared to the corresponding value for `ndf = 2`, i.e., 3.0872959.
- One expects power to decrease: the following code demonstrates that as `ndf` increases, the critical value `fCrit` decreases.
- In significance testing generally `ndf` = I -1.
- It more likely that the NH will be rejected with increasing numbers of treatments.

ndf	ddf	fCrit
1	100	3.936143
2	100	3.087296
5	100	2.305318
10	100	1.926692
12	100	1.850255
15	100	1.767530
20	100	1.676434

10.8.2 Fig. J

- This corresponds to `ncp` = 2, `ndf` = 1 and `ddf` = 100.
- Now `prob > fCrit_1_100` = 0.2883607, 0.1351602, 0.0168844, 8.9992114×10^{-4}, 3.2584757×10^{-4}, 8.1619807×10^{-5}, 1.1084132×10^{-5}, larger than the previous value 0.2199264.
- The power has actually increased.

10.8.3 Fig. K

- This corresponds to `ncp` = 5, `ndf` = 1 and `ddf` = 100‘’,
- Now `prob > fCrit_1_100` = 0.6004962, 0.3632847, 0.0699798, 0.0048836, 0.0018367, 4.6889533×10^{-4}, 6.2058692×10^{-5}, larger than the previous value 0.4910802.
- Again, the power has actually increased.

10.8.4 Fig. L

- This corresponds to `ncp` = 10, `ndf` = 1 and `ddf` = 100
- Now `prob > fCrit_1_100` is 0.8793619, 0.7000168, 0.2459501, 0.0290856, 0.0123033, 0.0035298, 5.1213398×10^{-4}, larger than the previous value 0.8029764.
- The power has actually increased.

10.9 Summary

- Power increases with increasing `ddf` and `ncp`.
- The effect of increasing `ncp` is quite dramatic. This is because power depends on the square of `ncp`.
- As `ndf` increases, `fCrit` decreases, which makes it more likely that the NH will be rejected.
- With increasing numbers of treatments the probability is greater that the F-statistic will be large enough to exceed the critical value.

Chapter 11

Sample size estimation: OR method

11.1 TBA How much finished

30%

11.2 Introduction

11.3 Statistical Power

$$Power = 1 - \beta \tag{11.1}$$

11.3.1 Sample size estimation for random-reader random-cases

For convenience the OR model is repeated below with the case-set index suppressed:

$$Y_{n(ijk)} = \mu + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + (\tau RC)_{ijk} + \epsilon_{n(ijk)} \tag{11.2}$$

As usual, the treatment effects τ_i are subject to the constraint that they sum to zero. The observed effect size (a random variable) is defined by:

$$d = \theta_{1\bullet} - \theta_{2\bullet} \tag{11.3}$$

It is a realization of a random variable, so one has some leeway in the choice of anticipated effect size. In the significance-testing procedure described in TBA Chapter 09 interest was in the distribution of the F-statistic when the NH is true. For sample size estimation, one needs to know the distribution of the statistic when the NH is false. It was shown that then the observed F-statistic TBA Eqn. (9.35) is distributed as a non-central F-distribution $F_{ndf,ddf,\Delta}$ with non-centrality parameter Δ:

$$F_{DBM|AH} \sim F_{ndf,ddf,\Delta} \tag{11.4}$$

The non-centrality parameter was defined, Eqn. TBA (9.34), by:

$$\Delta = \frac{JK\sigma^2_{Y;\tau}}{\left(\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC}\right) + K\sigma^2_{Y;\tau R} + J\sigma^2_{Y;\tau C}} \tag{11.5}$$

To minimize confusion, this equation has been rewritten here using the subscript Y to explicitly denote pseudo-value derived quantities (in TBA Chapter 09 this subscript was suppressed.

The estimate of $\sigma^2_{Y;\tau C}$ can turn out to bee negative. To avoid a negative denominator, Hillis suggests the following modification:

$$\Delta = \frac{JK\sigma^2_{Y;\tau}}{\left(\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC}\right) + K\sigma^2_{Y;\tau R} + \max\left(J\sigma^2_{Y;\tau C}, 0\right)} \tag{11.6}$$

This expression depends on three variance components, $(\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC})$ - the two terms are inseparable - $\sigma^2_{Y;\tau R}$ and $\sigma^2_{Y;\tau C}$. The ddf term appearing in TBA Eqn. (11.4) was defined by TBA Eqn. (9.24) - this quantity does not change between NH and AH:

$$ddf_H = \frac{\left[MSTR + \max(MSTR - MSTRC, 0)\right]^2}{\frac{[MSTR]^2}{(I-1)(J-1)}} \tag{11.7}$$

The mean squares in this expression can be expressed in terms of the three variance-components appearing in TBA Eqn. (11.6). Hillis and Berbaum (Hillis and Berbaum, 2004) have derived these expression and they will not be repeated here (Eqn. 4 in the cited reference). RJafroc implements a function to calculate the mean squares, `UtilMeanSquares()`, which allows ddf to be calculated using Eqn. TBA (11.7). The sample size functions in this package need only the three variance-components (the formula for ddf_H is implemented internally).

For two treatments, since the individual treatment effects must be the negatives of each other (because they sum to zero), it is easily shown that:

$$\sigma^2_{Y;\tau} = \frac{d^2}{2} \tag{11.8}$$

11.3.2 Dependence of statistical power on estimates of model parameters

Examination of the expression for , Eqn. (11.5), shows that statistical power increases if:

- The numerator is large. This occurs if: (a) the anticipated effect-size d is large. Since effect-size enters as the *square*, TBA Eqn. (11.8), it is has a particularly strong effect; (b) If $J \times K$ is large. Both of these results should be obvious, as a large effect size and a large sample size should result in increased probability of rejecting the NH.
- The denominator is small. The first term in the denominator is $\left(\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC}\right)$. These two terms cannot be separated. This is the residual variability of the jackknife pseudovalues. It should make sense that the smaller the variability, the larger is the non-centrality parameter and the statistical power.
- The next term in the denominator is $K\sigma^2_{Y;\tau R}$, the treatment-reader variance component multiplied by the total number of cases. The reader variance $\sigma^2_{Y;R}$ has no effect on statistical power, because it has an equal effect on both treatments and cancels out in the difference. Instead, it is the treatment-reader variance $\sigma^2_{Y;R}$ that contributes "noise" tending to confound the estimate of the effect-size.
- The variance components estimated by the ANOVA procedure are realizations of random variables and as such subject to noise (there actually exists a beast such as variance of a variance). The presence of the K term, usually large, can amplify the effect of noise in the estimate of $\sigma^2_{Y;R}$, making the sample size estimation procedure less accurate.
- The final term in the denominator is $J\sigma^2_{Y;\tau C}$. The variance $\sigma^2_{Y;C}$ has no impact on statistical power, as it cancels out in the difference. The treatment-case variance component introduces "noise" into the estimate of the effect size, thereby decreasing power. Since it is multiplied by J, the number of readers, and typically $J << K$, the error amplification effect on accuracy of the sample size estimate is not as bad as with the treatment-reader variance component.
- Accuracy of sample size estimation, essentially estimating confidence intervals for statistical power, is addressed in (Chakraborty, 2010).

11.3.3 Formulae for random-reader random-case (RRRC) sample size estimation

11.3.4 Significance testing

11.3.5 p-value and confidence interval

11.3.6 Comparing DBM to Obuchowski and Rockette for single-reader multiple-treatments

Having performed a pilot study and planning to perform a pivotal study, sample size estimation follows the following procedure, which assumes that both reader and case are treated as random factors. Different formulae, described later, apply when either reader or case is treated as a fixed factor.

- Perform OR analysis on the pilot data. This yields the observed effect size as well as estimates of all relevant variance components and mean squares appearing in TBA Eqn. (11.5) and Eqn. (11.7).
- This is the difficult but critical part: make an educated guess regarding the effect-size, d, that one is interested in "detecting" (i.e., hoping to reject the NH with probability $1 - \beta$). The author prefers the term "anticipated" effect-size to "true" effect-size (the latter implies knowledge of the true difference between the modalities which, as noted earlier, would obviate the need for a pivotal study).
- Two scenarios are considered below. In the first scenario, the effect-size is assumed equal to that observed in the pilot study, i.e., $d = d_{obs}$.
- In the second, so-called "best-case" scenario, one assumes that the anticipate value of d is the observed value plus two-sigma of the confidence interval, in the correct direction, of course, i.e., $d = |d_{obs}| + 2\sigma$. Here σ is one-fourth the width of the 95% confidence interval for d_{obs}. Anticipating more than 2σ greater than the observed effect-size would be overly optimistic. The width of the CI implies that chances are less than 2.5% that the anticipated value is at or beyond the overly optimistic value. These points will become clearer when example datasets are analyzed below.
- Calculate statistical power using the distribution implied by Eqn. (11.4), to calculate the probability that a random value of the relevant F-statistic will exceed the critical value, as in §11.3.2.
- If power is below the desired or "target" power, one tries successively larger value of J and / or K until the target power is reached.

11.4 Formulae for fixed-reader random-case (FRRC) sample size estimation

It was shown in TBA §9.8.2 that for fixed-reader analysis the non-centrality parameter is defined by:

$$\Delta = \frac{JK\sigma^2_{Y;\tau}}{\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC} + J\sigma^2_{Y;\tau C}} \tag{11.9}$$

The sampling distribution of the F-statistic under the AH is:

$$F_{AH|R} \equiv \frac{MST}{MSTC} \sim F_{I-1,(I-1)(K-1),\Delta} \tag{11.10}$$

11.4.1 Formulae for random-reader fixed-case (RRFC) sample size estimation

It is shown in TBA §9.9 that for fixed-case analysis the non-centrality parameter is defined by:

$$\Delta = \frac{JK\sigma^2_{Y;\tau}}{\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC} + K\sigma^2_{Y;\tau R}} \tag{11.11}$$

Under the AH, the test statistic is distributed as a non-central F-distribution as follows:

$$F_{AH|C} \equiv \frac{MST}{MSTR} \sim F_{I-1,(I-1)(J-1),\Delta} \tag{11.12}$$

11.4.2 Example 1

In the first example the Van Dyke dataset is regarded as a pilot study. Two implementations are shown, a direct application of the relevant formulae, including usage of the mean squares, which in principle can be calculated from the three variance-components. This is then compared to the `RJafroc` implementation.

Shown first is the "open" implementation.

```
alpha <- 0.05;cat("alpha = ", alpha, "\n")
#> alpha =  0.05
rocData <- dataset02 # select Van Dyke dataset
retDbm <- StSignificanceTesting(dataset = rocData, FOM = "Wilcoxon", method = "DBM")
```

```
varYTR <- retDbm$ANOVA$VarCom["VarTR","Estimates"]
varYTC <- retDbm$ANOVA$VarCom["VarTC","Estimates"]
varYEps <- retDbm$ANOVA$VarCom["VarErr","Estimates"]
effectSize <- retDbm$FOMs$trtMeanDiffs["trt0-trt1","Estimate"]
cat("effect size = ", effectSize, "\n")
#> effect size =  -0.043800322

#RRRC
J <- 10; K <- 163
ncp <- (0.5*J*K*(effectSize)^2)/(K*varYTR+max(J*varYTC,0)+varYEps)
MS <- UtilMeanSquares(rocData, FOM = "Wilcoxon", method = "DBM")
ddf <- (MS$msTR+max(MS$msTC-MS$msTRC,0))^2/(MS$msTR^2)*(J-1)
FCrit <- qf(1 - alpha, 1, ddf)
Power <- 1-pf(FCrit, 1, ddf, ncp = ncp)
data.frame("J"= J,  "K" = K, "FCrit" = FCrit, "ddf" = ddf, "ncp" = ncp, "RRRCPower" = P
#>    J   K     FCrit       ddf       ncp  RRRCPower
#> 1 10 163 4.1270572 34.334268 8.1269825 0.79111255

#FRRC
J <- 10; K <- 133
ncp <- (0.5*J*K*(effectSize)^2)/(max(J*varYTC,0)+varYEps)
ddf <- (K-1)
FCrit <- qf(1 - alpha, 1, ddf)
Power <- 1-pf(FCrit, 1, ddf, ncp = ncp)
data.frame("J"= J,  "K" = K, "FCrit" = FCrit, "ddf" = ddf, "ncp" = ncp, "RRRCPower" = P
#>    J   K    FCrit ddf       ncp  RRRCPower
#> 1 10 133 3.912875 132 7.9873835 0.80111671

#RRFC
J <- 10; K <- 53
ncp <- (0.5*J*K*(effectSize)^2)/(K*varYTR+varYEps)
ddf <- (J-1)
FCrit <- qf(1 - alpha, 1, ddf)
Power <- 1-pf(FCrit, 1, ddf, ncp = ncp)
data.frame("J"= J,  "K" = K, "FCrit" = FCrit, "ddf" = ddf, "ncp" = ncp, "RRRCPower" = P
#>    J  K    FCrit ddf       ncp  RRRCPower
#> 1 10 53 5.117355   9 10.048716 0.80496663
```

For 10 readers, the numbers of cases needed for 80% power is largest (163) for RRRC and least for RRFC (53). For all three analyses, the expectation of 80% power is met - the numbers of cases and readers were chosen to achieve close to 80% statistical power. Intermediate quantities such as the critical value of the F-statistic, `ddf` and `ncp` are shown. The reader should confirm that the code does in fact implement the relevant formulae. Shown next is the `RJafroc` implementation. The relevant file is mainSsDbm.R, a listing of which follows:

11.4.3 Fixed-reader random-case (FRRC) analysis

11.4.4 Random-reader fixed-case (RRFC) analysis

11.4.5 Single-treatment multiple-reader analysis

11.5 Discussion/Summary/3

11.6 Chapter References

Bibliography

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.

Bunch, P. C., Hamilton, J. F., Sanderson, G. K., and Simmons, A. H. (1977). A free response approach to the measurement and characterization of radiographic observer performance. In *Application of Optical Instrumentation in Medicine VI*, volume 127, pages 124–135. International Society for Optics and Photonics.

Chakraborty, D., Breatnach, E., Yester, M., Soto, B., Barnes, G., and Fraser, R. (1986). Digital and conventional chest imaging: A modified ROC study of observer performance using simulated nodules. *Radiology*, 158:35–39.

Chakraborty, D. and Zhai, X. (2022). *RJafroc: Artificial Intelligence Systems and Observer Performance*. R package version 2.1.2.9000.

Chakraborty, D. P. (2002). Statistical power in observer performance studies: A comparison of the ROC and free-response methods in tasks involving localization. *Acad. Radiol.*, 9(2):147–156.

Chakraborty, D. P. (2010). Prediction accuracy of a sample-size estimation method for ROC studies. *Academic radiology*, 17:628–638.

Chakraborty, D. P. (2017). *Observer Performance Methods for Diagnostic Imaging: Foundations, Modeling, and Applications with R-Based Examples*. CRC Press, Boca Raton, FL.

Clarkson, E., Kupinski, M. A., and Barrett, H. H. (2006). A probabilistic model for the MRMC method, part 1: Theoretical development. *Academic Radiology*, 13(11):1410–1421.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, 2 edition.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

Dorfman, D., Berbaum, K., and Metz, C. (1992). ROC characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. *Invest. Radiol.*, 27(9):723–731.

Dorfman, D. D., Berbaum, K. S., and Lenth, R. V. (1995). Multireader, multicase receiver operating characteristic methodology: A bootstrap analysis. *Academic Radiology*, 2(7):626–633.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, Boca Raton.

Gallas, B. D. (2006). One-shot estimate of MRMC variance: AUC. *Academic Radiology*, 13(3):353–362.

Gallas, B. D., Pennello, G. a., and Myers, K. J. (2007). Multireader multicase variance analysis for binary data. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 24(12):70–80.

Hajian-Tilaki, K. O., Hanley, J. A., Joseph, L., and Collet, J. P. (1997). Extension of receiver operating characteristic analysis to data concerning multiple signal detection tasks. *Acad Radiol*, 4:222–229.

Hillis, S., Obuchowski, N., Schartz, K., and Berbaum, K. (2005). A comparison of the dorfman-berbaum-metz and obuchowski-rockette methods for receiver operating characteristic (ROC) data. *Statistics in Medicine*, 24(10):1579–1607.

Hillis, S. L. (2007). A comparison of denominator degrees of freedom methods for multiple observer (ROC) studies. *Statistics in Medicine*, 26:596–619.

Hillis, S. L. (2014). A marginal-mean ANOVA approach for analyzing multireader multicase radiological imaging data. *Statistics in Medicine*, 33(2):330–360.

Hillis, S. L., Berbaum, K., and Metz, C. (2008). Recent developments in the dorfman-berbaum-metz procedure for multireader (ROC) study analysis. *Acad Radiol*, 15(5):647–661.

Hillis, S. L. and Berbaum, K. S. (2004). Power estimation for the dorfman-berbaum-metz method. *Acad. Radiol.*, 11(11):1260–1273.

Hillis, S. L., Obuchowski, N. A., and Berbaum, K. S. (2011). Power estimation for multireader ROC methods: An updated and unified approach. *Academic Radiology*, 18(2):129–142.

ICRU (1996). Medical imaging: the assessment of image quality. *JOURNAL OF THE ICRU*, 54(1):37–40.

Ishwaran, H. and Gatsonis, C. A. (2000). A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *The Canadian Journal of Statistics*, 28(4):731–750.

Kupinski, M. A., Clarkson, E., and Barrett, H. H. (2006). A probabilistic model for the MRMC method, part 2: Validation and applications. *Academic Radiology*, 13(11):1422–1430.

Larsen, R. J. and Marx, M. L. (2005). *An introduction to mathematical statistics*. Prentice Hall Hoboken, NJ.

Niklason, L. T., Hickey, N. M., Chakraborty, D. P., Sabbagh, E. A., Yester, M. V., Fraser, R. G., and Barnes, G. T. (1986). Simulated pulmonary nodules: detection with dual-energy digital versus conventional radiography. *Radiology*, 160:589–593.

Noether, G. E. (1967). Elements of nonparametric statistics. Report, Wiley and Sons.

Obuchowski, N. A. (1998). Sample size calculations in studies of test accuracy. *Statistical Methods in Medical Research*, 7(4):371–392.

Obuchowski, N. A. (2000). Sample size tables for receiver operating characteristic studies. *Am. J. Roentgenol.*, 175(3):603–608.

Obuchowski, N. A. and Rockette, H. (1995). Hypothesis testing of the diagnostic accuracy for multiple diagnostic tests: An ANOVA approach with dependent observations. *Communications in Statistics: Simulation and Computation*, 24:285–308.

Roe, C. and Metz, C. (1997a). Variance-component modeling in the analysis of receiver operating characteristic index estimates. *Acad. Radiol.*, 4(8):587–600.

Roe, C. A. and Metz, C. (1997b). Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: Validation with computer simulation. *Acad Radiol*, 4:298–303.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, 6(5):309–316.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Series in Cognition and Perception. Academic Press, New York, first edition.

Toledano, A. and Gatsonis, C. (1996). Ordinal regression methodology for ROC curves derived from correlated data. *Stat Med*, 15(16):1807–1826.

Toledano, A. Y. (2003). Three methods for analyzing correlated ROC curves: A comparison in real data sets. *Statistics in Medicine*, 22(18):2919–33.

Van Dyke, C., White, R., Obuchowski, N., Geisinger, M., Lorig, R., and Meziane, M. (1993). Cine MRI in the diagnosis of thoracic aortic dissection. *79th RSNA Meetings*.

Zanca, F., Jacobs, J., Van Ongeval, C., Claus, F., Celis, V., Geniets, C., Provost, V., Pauwels, H., Marchal, G., and Bosmans, H. (2009). Evaluation of clinical image processing algorithms used in digital mammography. *Medical Physics*, 36(3):765–775.

Zhou, X.-H., McClish, D. K., and Obuchowski, N. A. (2009). *Statistical methods in diagnostic medicine*, volume 569. John Wiley & Sons.