The RJafroc Book

Dev P. Chakraborty, PhD

2020-06-14

Contents

Preface **5**

A note on the online distribution mechanism of the book **7**

Contributing to this book **9**

1 Example running external scripts **11**

2 Obuchowski Rockette Hillis2 (ORH) Analysis **13**

2.1 Introduction 13

2.2 Single-reader multiple-treatment model 14

2.3 Multiple-reader multiple-treatment ORH model 27

2.4 Discussion/Summary 39

2.5 References 40

Preface

- This book, an extended documentation of the **RJafroc** package, is currently (as of April 2020) in preperation. It is intended to bypass the file size limits of **CRAN**, which severely limits the extent of the documentation that can be included with the CRAN package.

A note on the online distribution mechanism of the book

- In the hard-copy version of my book (Chakraborty, 2017) the online distribution mechanisms was `BitBucket`.
- `BitBucket` allows code sharing within a *closed* group of a few users (e.g., myself and a student).
- Since the purpose of open-source code is to encourage collaborations, this was, in hindsight, an unfortunate choice. Moreover, as my experience with R-packages grew, it became apparent that the vast majority of R-packages are shared on `GitHub`, not `BitBucket`.
- For these reasons I have switched to `GitHub`. Any previous instructions pertaining to `BitBucket` are obsolete.
- In order to access `GitHub` material one needs to create a (free) account.
- Go to this link and click on `Sign Up`.

Contributing to this book

- I appreciate any feedback on this document, e.g., corrections, comments, etc.

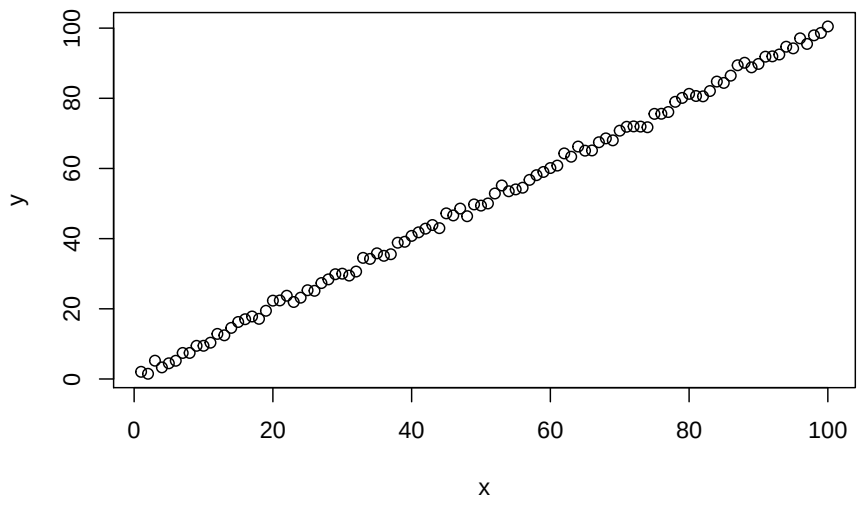
- To do this raise an `Issue` on the `GitHub` interface.
- Click on the `Issues` tab under `dpc10ster/RJafrocBook`, then click on `New issue`.
- Contributions from users automatically become part of the `GitHub` documentation/history of the book.

Chapter 1

Example running external scripts

source method

```
source(here("R/example.R"))
```



```
# source(here("R/example2.R"))
print(head(data.frame(x,y)))
#>   x        y
#> 1 1 2.039574
#> 2 2 1.485255
#> 3 3 5.214021
#> 4 4 3.305117
#> 5 5 4.471868
#> 6 6 5.191855
```

Chapter 2

Obuchowski Rockette Hillis2 (ORH) Analysis

2.1 Introduction

The previous chapter described the DBM significance testing procedure (Dorfman et al., 1992) for analyzing MRMC ROC data, along with improvements suggested by Hillis. Because the method assumes that jackknife pseudovalues can be regarded as independent and identically distributed case-level figures of merit, it has been criticized by Hillis who states that the method "works" but lacks firm statistical foundations (Hillis et al., 2005; Hillis, 2007; Hillis et al., 2008). If a method works there must be good reasons why it works and the last section of the previous chapter, §9.13, gave a justification for why the method works. Specifically, the empirical AUC pseudovalues qualify as case-level FOMs - this property was also 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 desirable.

This chapter presents Hillis' preferred alternative to the DBMH approach. He has argued that the DBMH method can be regarded as a "working model that gives the right results", but a method based on an earlier publication (Obuchowski and Rockette, 1995) by Obuchowski and Rockette, which does not depend on pseudovalues, and predicts more or less the same results, is preferable from a conceptual viewpoint. Since, besides showing the correspondence, Hillis has made significant improvements to the original methodology, this chapter is named "ORH Analysis", where ORH stands for Obuchowski, Rockette and Hillis. The ORH method has advantages in being able to handle more complex study designs (Hillis, 2014) that are outside the scope of this book (the author acknowledges a private communication from Dr. Obuchowski, ca. 2006, that demonstrated the flexibility afforded by the OR approach) and it is likely that

applications to other FOMs (e.g., the FROC paradigm uses a rather different FOM from empirical ROC-AUC) are better performed with the ORH method.

This chapter starts with a gentle introduction to the Obuchowski and Rockette method. The reason is that the method was rather opaque to me, an I suspect, most users. Part of the problem, in my opinion, is the notation, namely lack of usage of the *case-set* index $\{c\}$. A key difference of the Obuchowski and Rockette method from DBMH is in how the error term is modeled by a non-diagonal covariance matrix. The structure of the covariance matrix is examined in some detail as it is key to understanding the ORH method.

In the first step of the introduction a single reader interpreting a case-set in multiple treatments is modeled and the results compared to those obtained using DBMH fixed-reader analysis described in the previous chapter. In the second step multiple readers interpreting a case-set in multiple treatments is modeled. The two analyses, DBMH and ORH, 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.

2.2 Single-reader multiple-treatment model

Consider a single-reader providing ROC interpretations of a common case-set $\{c\}$ in multiple-treatments i ($i = 1, 2, ..., I$). Before proceeding, we note that this is not homologous (i.e., formally equivalent) to multiple-readers providing ROC interpretations in a single treatment, §10.7; this is because reader is a random factor while treatment is not. The figure of merit θ is modeled as:

$$\theta_{i\{c\}} = \mu + \tau_i + \epsilon_{i\{c\}} \tag{2.1}$$

In the (Obuchowski and Rockette, 1995) one models the figure-of-merit, not the pseudovalues, indeed this is one of the key differences from the DBMH method.

Recall that $\{c\}$ denotes a *set of cases*. (2.1) models the observed figure-of-merit $\theta_{i\{c\}}$ as a constant term μ plus a treatment dependent term τ_i (the treatment-effect) with the constraint:

$$\sum_{i=1}^{I} \tau_i = 0 \tag{2.2}$$

The *c-index* was introduced in (book) Chapter 07. The left hand side of (2.1) 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, *not individual cases.*

This is one place 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 any index allowing for sampling related variability.

Obuchowski and Rockette use a *k-index*, defined as the "kth repetition of the study involving the same diagnostic test, reader and patient (sic)". In the author's opinion, what is meant is a case-set index instead of a repetition index. Repeating a study with the same treatment, reader and cases yields *within-reader* variability, which is different from sampling the population of cases with new case-sets, which yields *case-sampling plus within-reader* variability. As noted earlier, within-reader variability cannot be "turned off" and affects the interpretations of all case-sets.

Interest is in extrapolating to the population of cases and the only way to do this is to sample different case-sets. It is shown below that usage of the case-set index interpretation yields the same results using the DBMH or the ORH methods.

Finally, and this is where I had some difficulty understanding what is going on, there is 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:

$$\epsilon_{i\{c\}} \sim N_I\left(\vec{0}, \Sigma\right) \tag{2.3}$$

Here N_I is the I-variate normal distribution (i.e., each sample yields I random numbers). Obuchowski and Rockette assumed the following structure for the covariance matrix (they describe a more general multi-reader model, but here one restricts to the simpler single-reader case):

$$\Sigma_{ii'} = Cov\left(\epsilon_{i\{c\}}, \epsilon_{i'\{c\}}\right) = \begin{cases} Var & (i = i') \\ Cov_1 & (i \neq i') \end{cases} \tag{2.4}$$

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

$$\Sigma = \begin{pmatrix} Var & Cov_1 & ... & Cov_1 & Cov_1 \\ Cov_1 & Var & ... & Cov_1 & Cov_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Cov_1 & Cov_1 & ... & Var & Cov_1 \\ Cov_1 & Cov_1 & ... & Cov_1 & Var \end{pmatrix} \tag{2.5}$$

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.

A more complex model, with more parameters and therefore more difficult to work with, would allow the variances to be treatment dependent, and the co-variances 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.

Some elementary statistical results are presented next.

2.2.1 Definitions of covariance and correlation

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

$$Cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i - x_\bullet)(y_i - y_\bullet)}{N-1} = E(XY) - E(X) - E(Y) \qquad (2.6)$$

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:

$$E(X) = \int pdf(x)xdx$$

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:

$$Var(X,Y) = Cov(X,X) = E(X^2) - (E(X))^2 = \sigma_x^2$$

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

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

A related quantity, 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:

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

2.2.2 Special case when variables have equal variances

Assuming X and Y have the same variance:

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

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

$$Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y) = 2(Var - Cov) \qquad (2.7)$$

The left part of the above equation is general, the right part specializes to the OR single-reader multiple-treatment model where the variances are equal and likewise all covariances in (2.5) 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 Cov_1, which will be explained later):

$$\rho_1 = \frac{Cov_1}{Var}$$

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 elements at and above the diagonal are shown):

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & & \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & & & \sigma^2 & \rho_1\sigma^2 \\ & & & & \sigma^2 \end{bmatrix} \qquad (2.8)$$

2.2.3 Estimation of the covariance matrix

An unbiased estimate of the covariance (2.4) follows from:

$$\Sigma_{ii'} = \frac{1}{C-1} \sum_{c=1}^{C} \left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right) \left(\theta_{i'\{c\}} - \theta_{i'\{\bullet\}}\right) \qquad (2.9)$$

Sampling different case-sets, as required (2.9), is unrealistic and in reality one is stuck with $C = 1$, 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 can use 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:

$$\Sigma_{ii'}\,|_{bs} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\theta_{i\{b\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i'\{b\}} - \theta_{i'\{\bullet\}}\right) \tag{2.10}$$

(2.10), the bootstrap method of estimating the covariance matrix, is a direct translation of (2.9). Alternatively, one could have used the jackknife FOM values $\theta_{i(k)}$, i.e., the figure of merit with a particular case removed, for all cases, to estimate the covariance matrix:

$$\Sigma_{ii'}\,|_{jk} = \frac{(K-1)^2}{K}\left[\frac{1}{K-1}\sum_{k=1}^{K}\left(\theta_{i(k)} - \theta_{i(\bullet)}\right)\left(\theta_{i'(k)} - \theta_{i'(\bullet)}\right)\right] \tag{2.11}$$

Note the subtle difference in notation between (2.9) and EstimateSigmaJackknife and. In the former, the subscript $\{c\}$ denotes a set of K cases while in the latter, (k) denotes the original case set with a particular case k removed, leaving $K-1$ cases.

For simplicity, in this section we depart from the usual two-subscript convention to index each case. So k ranges from 1 to K, where the first K_1 values represent non-diseased and the following K_2 values represent diseased cases. Jackknife figure of merit values are not to be confused with jackknife pseudovalues. The jackknife FOM value corresponding to a particular case is the FOM with the particular case removed. Unlike pseudovalues, jackknife FOM values cannot be regarded as independent and identically distributed. Notice the use of the subscript enclosed in parenthesis (k) to denote the FOM with case k removed, i.e., a single case, while in the bootstrap equation one uses the curly brackets $\{b\}$ to denote the bth 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. Furthermore, the expression for the jackknife covariance contains a *variance inflation factor*:

$$\frac{(K-1)^2}{K} \tag{2.12}$$

This factor multiplies the traditional expression for the covariance, shown in square brackets in (2.11). A third method of estimating the covariance, namely the DeLong et al. method (DeLong et al., 1988), applicable only to the empirical AUC, is described later.

2.2.4 Meaning of the covariance matrix in (2.5)

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,...,I and c = 1,2,...,C. Considering treatment i, the variance of the FOM-values for the different case-sets c = 1,2,...,C, is an estimate of Var_i for this treatment:

$$\sigma_i^2 \equiv Var_i = \frac{1}{C-1}\sum_{c=1}^{C}\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right) \tag{2.13}$$

The process is repeated for all treatments and the I-variance values are averaged. This is the final estimate of Var appearing in (2.3).

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.* Since they are derived from the same case-set, one expects the values to be correlated. For a particularly easy case-set one expects all I-estimates to be collectively higher than usual. The process is repeated for different case-sets and one calculates the correlation $\rho_{1;ii'}$ between the two C-length arrays $\theta_{i\{c\}}$ and $\theta_{i'\{c\}}$:

$$\rho_{1;ii'} = \frac{1}{C-1}\sum_{c=1}^{C}\frac{\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i'\{c\}} - \theta_{i'\{\bullet\}}\right)}{\sigma_i\sigma_{i'}} \tag{2.14}$$

The entire process is repeated for different treatment pairings and the resulting $I(I-1)/2$ distinct values are averaged yielding the final estimate of ρ_1 in (2.8). According to @ref(eq: EstimateRho) one expects the covariance to be smaller than the variance determined as in the previous paragraph.

In most situations one expects ρ_1 (for ROC studies) to be positive. There is, perhaps unlikely, a scenario that could lead to anti-correlation and negative. This could occur, with "complementary" treatments, e.g., CT vs. MRI, where one treatment is good for bone imaging and the other for soft-tissue imaging. In this situation what constitutes an easy case-set in one treatment could be a difficult case-set in the other treatment.

2.2.5 Code illustrating the covariance matrix (TBA)

As indicated above, the covariance matrix can be estimated using the jackknife or the bootstrap. If the figure of merit is the Wilcoxon statistic, then one can also use the DeLong et al method (DeLong et al., 1988). In (book) Chapter 07, these methods were described in the context of estimating the variance of AUC. (2.10) and (2.11) 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). The extension of the DeLong method to covariances is described in Online Appendix 10.A (TBA) and implemented

in file `VarCovMtrxDLStr.R`. The file name stands for "variance covariance matrix according to the DeLong structural components method" *described in five unnumbered equations following Eqn. 4 in the cited reference.*

- The functions (for `Var` and `Cov1` using bootstrap, jackknife, and the DeLong methods) are not displayed, but they are compiled. To display them download the repository and look at the `Rmd` file corresponding to this output and the sourced files listed below:

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

The following code chunk extracts (using the `DfExtractDataset` function) a single-reader multiple-treatment ROC dataset corresponding to the first reader from `dataset02`, i.e., the Van Dyke or `VD` 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 = boostrap method, with B = number of bootstraps and `seed` = value.
- dl = DeLong method
- rjjk = RJafroc, covEstMethod = "jackknife"
- rjbs = RJafroc, covEstMethod = "bootstrap"

For exammple, `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 `UtilVarComponentsOR`:

```
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 <- UtilVarComponentsOR(rocData1R, FOM = "Wilcoxon")$varComp # default `covEstMethod` is jack
data.frame ("Cov_rjjk" = ret4$cov1, "Var_rjjk" = ret4$var)
#> data frame with 0 columns and 0 rows
```

Note that the estimates are identical and that the Cov_1 estimate is smaller than the Var estimate (their ratio is the correlation $\rho_1 = Cov_1/Var = 0.5343623$).

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 `UtilVarComponentsOR`:

```
ret5 <- UtilVarComponentsOR(rocData1R, FOM = "Wilcoxon", covEstMethod = "bootstrap", nBoots = 200
data.frame ("Cov_rjbs" = ret5$cov1, "Var_rjbs" = ret5$var)
#> data frame with 0 columns and 0 rows
```

Note that the two estimates are identical *provided the seeds are identical*.

Following are results of the DeLong covariance estimation method, the first output using this repository code and the second using the `RJafroc` function `UtilVarComponentsOR` 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 <- UtilVarComponentsOR(rocData1R, FOM = "Wilcoxon", covEstMethod = "DeLong")$varC
data.frame ("Cov_rjdl" = ret5$cov1, "Var_rjdl" = ret5$var)
#> data frame with 0 columns and 0 rows
```

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

2.2.6 Significance testing

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

$$MS(T) = \frac{1}{I-1}\sum_{i=1}^{I}(\theta_i - \theta_\bullet)^2 \tag{2.15}$$

Unlike the previous chapter, all mean square quantities defined 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.,

$$\chi_{1R} \equiv \frac{(I-1)MS(T)}{Var - Cov_1} \sim \chi^2_{I-1} \tag{2.16}$$

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

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

$$F_{1R} \equiv \frac{MS(T)}{Var - Cov_1} \sim F_{I-1,\infty} \tag{2.17}$$

2.2.6.1 An aside on the relation between the chisquare and the F-distribution with infinite ddf

Define $D_{1-\alpha}$, the $(1-\alpha)$ quantile of distribution D, as that "cutoff" value such that the probability of observing a random sample d less than or equal to $D_{1-\alpha}$ is $(1-\alpha)$. In other words,

$$\Pr(d \leq D_{1-alpha} \mid d \sim D) = 1 - \alpha \tag{2.18}$$

With definition (2.18), 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):

$$\frac{\chi^2_{1-\alpha,I-1}}{I-1} = F_{1-\alpha,I-1,\infty} \tag{2.19}$$

(2.19) 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
```

2.2.7 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:

$$p \equiv \Pr(f > F_{1R} \mid f \sim F_{I-1,\infty}) \tag{2.20}$$

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

$$CI_{1-\alpha,1RMT} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,\infty}\sqrt{2(Var - Cov_1)} \tag{2.21}$$

Comparing (2.21) to (2.7) shows that the term $\sqrt{2(Var - Cov_1)}$ 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.]

2.2.8 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 DBMH 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 to get the covariance matrix, or some other methods such as the DeLong method, if one restricts to the Wilcoxon statistic for the figure of merit (empirical ROC-AUC). Since one is dealing with a single reader in multiple treatments, for DBMH one needs the fixed-reader random-case analysis described in §9.8 of the previous chapter (with one reader the conclusions obviousl apply to the specific reader, so reader must be regarded as a fixed factor).

Shown below are results obtained using RJafroc function `StSignificanceTesting` with `analysisOption = "FRRC"` for DBMH (which uses the jackknife), and for ORH using 3 different ways of estimating the covarince matrix (i.e., Cov_1 and Var).

```
ret1 <- StSignificanceTesting(rocData1R,FOM = "Wilcoxon", method = "DBMH", analysisOpti
data.frame("DBMH:F" = ret1$FTestStatsFRRC$fFRRC,
           "DBMH:ddf" = ret1$FTestStatsFRRC$ddfFRRC,
           "DBMH:P-val" = ret1$FTestStatsFRRC$pFRRC)
#> data frame with 0 columns and 0 rows

ret2 <- StSignificanceTesting(rocData1R,FOM = "Wilcoxon", method = "ORH", analysisOptio
data.frame("ORHJack:F" = ret2$FTestStatsFRRC$fFRRC,
           "ORHJack:ddf" = ret2$FTestStatsFRRC$ddfFRRC,
           "ORHJack:P-val" = ret2$FTestStatsFRRC$pFRRC)
#> data frame with 0 columns and 0 rows

ret3 <- StSignificanceTesting(rocData1R,FOM = "Wilcoxon", method = "ORH", analysisOptio
                              covEstMethod = "DeLong")
data.frame("ORHDeLong:F" = ret3$FTestStatsFRRC$fFRRC,
           "ORHDeLong:ddf" = ret3$FTestStatsFRRC$ddfFRRC,
           "ORHDeLong:P-val" = ret3$FTestStatsFRRC$pFRRC)
#> data frame with 0 columns and 0 rows

ret4 <- StSignificanceTesting(rocData1R,FOM = "Wilcoxon", method = "ORH", analysisOptio
                              covEstMethod = "bootstrap")
data.frame("ORHBoot:F" = ret4$FTestStatsFRRC$fFRRC,
           "ORHBoot:ddf" = ret4$FTestStatsFRRC$ddfFRRC,
           "ORHBoot:P-val" = ret4$FTestStatsFRRC$pFRRC)
#> data frame with 0 columns and 0 rows
```

The DBMH and ORH-jackknife methods yield identical F-statistics, but the

denominator degrees of freedom are different, $(I-1)(K-1) = 113$ for DBMH and ∞ for ORH. The F-statistics for ORH-bootstrap and ORH-DeLong are different.

Shown below is a first-principles implementation of signficance 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
```

Next, how does it compare to `RJafroc` `FRRC` analysis using the `StSignificanceTesting` function?

```
ret_rj <- StSignificanceTesting(rocData1R, FOM = "Wilcoxon", method = "ORH", analysisO
print(data.frame("theta_1" = ret_rj$fomArray[1],
                 "theta_2" = ret_rj$fomArray[2],
                 "Var" = ret_rj$varComp$var,
                 "Cov1" = ret_rj$varComp$cov,
                 "MS_T" = ret_rj$msT,
                 "F_1R" = ret_rj$FTestStatsFRRC$fFRRC,
                 "pValue" = ret_rj$FTestStatsFRRC$pFRRC,
                 "Lower" = ret_rj$ciDiffTrtFRRC$CILower,
                 "Mid" = ret_rj$ciDiffTrtFRRC$Estimate,
                 "Upper" = ret_rj$ciDiffTrtFRRC$CIUpper))
#> data frame with 0 columns and 0 rows
```

The first-principles and the `RJafroc` values agree exactly with each other. This above code also shows how to extract the different estimates (Var, Cov_1, etc.) from the object `ret_rj` returned by `RJafroc`.

- Var: ret_rj$varComp$var
- Cov1: ret_rj$varComp$cov
- F-statistic: ret_rj$FTestStatsFRRC$fFRRC
- ddf: ret_rj$FTestStatsFRRC$ddfFRRC
- p-value: ret_rj$FTestStatsFRRC$pFRRC
- CI Lower: ret_rj$ciDiffTrtFRRC$CILower
- Mid Value: ret_rj$ciDiffTrtFRRC$Estimate
- CI Upper: ret_rj$ciDiffTrtFRRC$CIUpper

2.2.8.1 Jumping ahead

If RRRC analysis were conducted, the values would be:

- msR: ret_rj$meanSquares$msR
- msT: ret_rj$meanSquares$msT
- msTR: ret_rj$meanSquares$msTR
- Var: ret_rj$varComp$var
- Cov1: ret_rj$varComp$cov
- Cov2: ret_rj$varComp$cov2
- Cov3: ret_rj$varComp$cov3
- varR: ret_rj$varComp$varR

- varTR: ret_rj$varComp$varTR
- F-statistic: ret_rj$FTestStatsRRRC$fRRRC
- ddf: ret_rj$FTestStatsRRRC$ddfRRRC
- p-value: ret_rj$FTestStatsRRRC$pRRRC
- CI Lower: ret_rj$ciDiffTrtRRRC$CILower
- Mid Value: ret_rj$ciDiffTrtRRRC$Estimate
- CI Upper: ret_rj$ciDiffTrtRRRC$CIUpper

And similarly, for RRFC analysis, one replaces RRRC with RRFC.]

2.3 Multiple-reader multiple-treatment ORH model

The previous sections served as a gentle introduction to the single-reader multiple-treatment Obuchowski and Rockette method. 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 (2.1), one needs an additional j index to index readers, and additional random 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 general Obuchowski and Rockette model for fully paired multiple-reader multiple-treatment interpretations is:

$$\theta_{ij\{c\}} = \mu + \tau_i + R_j + (\tau R)_{ij} + \epsilon_{ij\{c\}} \tag{2.22}$$

The fixed treatment effect τ_i is subject to the usual constraint, (2.2). The first two terms on the right hand side of (2.22) 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 following two terms are, by assumption, mutually independent random samples specified as follows: R_j denotes the random treatment-independent contribution to the figure-of-merit of reader j (j = 1,2, …, J), modeled as a sample from 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 $\sigma_{\tau R}^2$. There could be a perceived notational clash with similar variance component terms defined for the DBMH model – except in that case they applied to pseudovalues. The meaning should be clear from the context. Summarizing:

$$\begin{cases} R_j \sim N(0, \sigma_R^2) \\ \tau R \sim N(0, \sigma_{\tau R}^2) \end{cases} \tag{2.23}$$

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):

$$\mu = \theta_{\bullet\bullet\{1\}} \tag{2.24}$$

As before the dot subscript denotes an average over the replaced index. Averaging over the j index and performing a subtraction yields an estimate of :

$$\tau_i = \theta_{i\bullet\{1\}} - \theta_{\bullet\bullet\{1\}} \tag{2.25}$$

The τ_i estimates obey the constraint (2.2). 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 (2.22) 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,

$$\epsilon_{ij\{c\}} \sim N_{IJ}(\vec{0}, \Sigma) \tag{2.26}$$

Here N_{IJ} is the N_{IJ} variate normal distribution. The covariance matrix Σ is defined by 4 parameters, Var, Cov_1, Cov_2, Cov_3, defined as follows:

$$Cov(\epsilon_{ij\{c\}}, \epsilon_{i'j'\{c\}}) = \left\{ \begin{matrix} Var\ (i = i', j = j') \\ Cov1\ (i \neq i', j = j') \\ Cov2\ (i = i', j \neq j') \\ Cov3\ (i \neq i', j \neq j') \end{matrix} \right\} \tag{2.27}$$

Apart from fixed effects, the model implied by (2.22) and (2.27) contains 6 parameters:

$$\sigma_R^2, \sigma_{\tau R}^2, Var, Cov_1, Cov_2, Cov_3$$

This is the same number of variance component parameters as in the DBMH model, which should not be a surprise since one is modeling the data with equivalent models. The Obuchowski and Rockette model (2.22) "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, as allowing treatment and reader dependencies would greatly increase the number of parameters that would need to be estimated.

It is implicit in the Obuchowski-Rockette model that the Var, Cov_1, Cov_2$, and Cov_3, estimates need to be averaged over all applicable treatment-reader combinations.

2.3.1 Structure of the covariance matrix

To understand the structure of this matrix, recall that the diagonal elements of a (square) covariance matrix are variances and the off-diagonal elements are covariances. With two indices ij one can still imagine a square matrix where 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 6x6 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 (2.27) the entries in these cells would be 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 (2.27) the entries in cells with *different treatment index pairs but identical reader index pairs* would be Cov_1 (as an example, the cell (21,11) has the same reader index, namely reader 1, but different treatment indices, namely 2 and 1, so it is replaced by Cov_1):

$$\Sigma = \begin{bmatrix} Var & (12,11) & (13,11) & Cov_1 & (22,11) & (23,11) \\ & Var & (13,12) & (21,12) & Cov_1 & (23,12) \\ & & Var & (21,13) & (22,13) & Cov_1 \\ & & & Var & (22,21) & (23,21) \\ & & & & Var & (23,22) \\ & & & & & Var \end{bmatrix}$$

Similarly, the entries in cells with identical treatment index pairs but different reader index pairs would be Cov_2:

$$\Sigma = \begin{bmatrix} Var & Cov_2 & Cov_2 & Cov_1 & (22,11) & (23,11) \\ & Var & Cov_2 & (21,12) & Cov_1 & (23,12) \\ & & Var & (21,13) & (22,13) & Cov_1 \\ & & & Var & Cov_2 & Cov_2 \\ & & & & Var & Cov_2 \\ & & & & & Var \end{bmatrix}$$

Finally, the entries in cells with different treatment index pairs and different reader index pairs would be Cov_3:

$$\Sigma = \begin{bmatrix} Var & Cov_2 & Cov_2 & Cov_1 & Cov_3 & Cov_3 \\ & Var & Cov_2 & Cov_3 & Cov_1 & Cov_3 \\ & & Var & Cov_3 & Cov_3 & Cov_1 \\ & & & Var & Cov_2 & Cov_2 \\ & & & & Var & Cov_2 \\ & & & & & Var \end{bmatrix}$$

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 term Var can be estimated as follows:

$$Var = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})^2 (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})^2 \right\rangle_{ij} \quad \epsilon_{ij\{c\}} \sim N_{IJ}(\vec{0}, \Sigma) \tag{2.28}$$

Of course, in practice one would use the bootstrap or the jackknife as a stand-in for the c-index, but for pedagogic purpose, one maintains the fiction that one has a large number of case-sets at one's disposal (not to mention the time spent by the readers interpreting them). Notice that the left-hand-side of (2.28) 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:

$$Cov = \begin{cases} Cov_1 = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})^2 (\theta_{i'j\{c\}} - \theta_{i'j\{\bullet\}})^2 \right\rangle_{ii',jj} \\ Cov_2 = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})^2 (\theta_{ij'\{c\}} - \theta_{ij'\{\bullet\}})^2 \right\rangle_{ii,jj'} \\ Cov_3 = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})^2 (\theta_{i'j'\{c\}} - \theta_{i'j'\{\bullet\}})^2 \right\rangle_{ii',jj'} \end{cases} \quad (2.29)$$

In (2.29) *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 Cov_1 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 $\langle\rangle_{ii',jj'}$.

2.3.2 Physical meanings of the covariance terms

The meanings of the different terms follow a similar description to that given in 2.3.1. The diagonal term Var of the covariance matrix Σ is the variance of the figure-of-merit values obtained when reader j interprets different case-sets 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 and within-reader variability.

$\rho_{1;ii'jj}$ is the correlation of the figure-of-merit values obtained when the same reader j interprets a case-set in different treatment i, i'. Each case-set, starting with $c = 1$, yields two numbers $\theta_{ij\{1\}}$ and $\theta_{i'j\{1\}}$; the process is repeated for C case-sets. The correlation of the two pairs of C-length arrays, averaged over all pairings of different treatments and same readers, is ρ_1. Because of the common contribution due to the shared reader, ρ_1 will be non-zero. For large common variation, the two arrays become almost perfectly correlated, and ρ_1 approaches unity. For zero common variation, the two arrays become independent, and ρ_1 equals zero. Translating to covariances, one has $Cov_1 < Var$.

$\rho_{2;iijj'}$ is the correlation of the figure-of-merit values obtained when different readers j, j' interpret the same case-set in the same treatment i. As before this yields two numbers and upon repeating over C case-sets one has 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 on different treatments than they do with other readers on the same treatment. Translating to covariances, one has $Cov_2 < Cov_1 < Var$.

$\rho_{3;ii'jj'}$ is the correlation of the figure-of-merit values obtained 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:

$$Cov_3 < Cov_2 < Cov_1 < Var \tag{2.30}$$

2.3.3 ORH random-reader random-case analysis

A model such as (2.22) cannot be analyzed by standard analysis of variance (ANOVA) techniques. Because of the correlated structure of the error term a customized ANOVA is needed (in standard ANOVA models, such as used in DBMH, the covariance matrix of the error term is diagonal with all diagonal elements equal to a common variance, represented by the epsilon term in the DBM model).

One starts with the null hypothesis (NH) that the true figures-of-merit of all treatments are identical, i.e.,

$$NH : \tau_i = 0 \;\; (i = 1, 2, ..., I) \tag{2.31}$$

The analysis described next considers both readers and cases as random effects. Because of the special nature of the covariance matrix, a modified F-statistic is needed 1-4,7, denoted F_{ORH}, defined by:

$$F_{ORH} = \frac{MS(T)}{MS(TR) + J \max(Cov_2 - Cov_3, 0)} \tag{2.32}$$

(2.32) incorporates Hillis' modification, which ensures that the constraint (2.30) is always obeyed and avoids a possibly negative (hence illegal) F-statistic. The mean square (MS) terms are defined by (these are calculated directly using FOM values, not pseudovalues):

$$\left.\begin{aligned} MS(T) &= \tfrac{J}{I-1} \textstyle\sum_{i=1}^{I} (\theta_{i\bullet} - \theta_{\bullet\bullet})^2 \\ MS(TR) &= \tfrac{1}{(I-1)(J-1)} \textstyle\sum_{i=1}^{I} \sum_{j=1}^{J} (\theta_{ij} - \theta_{i\bullet} - \theta_{\bullet j} + \theta_{\bullet\bullet}) \end{aligned}\right\} \tag{2.33}$$

In their original paper (Obuchowski and Rockette, 1995) Obuchowski and Rockette state that their proposed test statistic F (basically (2.32) without the constraint implied by the max function) is distributed as an F-statistic with numerator degree of freedom $ndf = I - 1$ and denominator degree of freedom $ddf = (I-1)(J-1)$. It turns out that then the test is unduly conservative, meaning it is unusually reluctant to reject the null hypothesis.

In this connection the author has two historical anecdotes. The late Dr. Robert F. Wagner once stated to the author (ca. 2001) that the sample-size tables published by Obuchowski (Obuchowski, 1998, 2000), using the unmodified version of (2.32), 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 the author once conducted NH simulations using the Roe-Metz simulator described in the preceding chapter and the significance testing as 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. The author recalls (ca. 2004) telling Dr. Steve Hillis about this issue, and he suggested a different value for the denominator degrees of freedom (ddf), substitution of which magically solved the problem, i.e., the simulations rejected the null hypothesis about 5% of the time; the new ddf is defined below (ndf is unchanged), with the subscript H denoting the Hillis modification:

$$ndf = I - 1 \tag{2.34}$$

$$ddf_H = \frac{\left[MS(TR) + J\max(Cov_2 - Cov_3)\right]^2}{\frac{\left[MS(TR)\right]^2}{(I-1)(J-1)}} \tag{2.35}$$

If $Cov_2 < Cov_3$ this reduces to the expression originally suggested by Obuchowski and Rockette. With these changes, under the null hypothesis, the observed statistic F_{ORH}, defined in (2.32), is distributed as an F-statistic with $I-1$ and ddf_H degrees of freedom (Hillis et al., 2005; Hillis, 2007; Hillis et al., 2008):

$$F_{ORH} \sim F_{ndf,ddf_H} \tag{2.36}$$

2.3.3.1 Decision rule, p-value and confidence interval

The critical value of the F-statistic for rejection of the null hypothesis is $F_{1-\alpha,ndf,ddf_H}$, i.e., that value such that fraction $(1-\alpha)$ of the area under the distribution lies to the left of the critical value. From definition (2.32), rejection of the NH is more likely if $MS(T)$ increases, meaning the treatment effect is larger; $MS(TR)$ decreases meaning there is less contamination of the treatment

effect by treatment-reader variability; the greater of $Cov2$ or $Cov3$ decreases, meaning there is less contamination of the treatment effect by between-reader and treatment-reader variability, α increases, meaning one is allowing a greater probability of Type I errors, ndf increases, meaning the more the number of treatment pairings, the greater the chance that at least one pair will reject the NH or 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_{ndf,ddf_H} that lies above the observed value F_{ORH}:

$$p = \Pr(F > F_{ORH} \mid F \sim F_{ndf,ddf_H}) \tag{2.37}$$

The $(1-\alpha)$ confidence interval for $\theta_{i\bullet} - \theta_{i'\bullet}$ is given by (the average is over the reader index; the case-set index $\{1\}$ is suppressed):

$$CI_{1-\alpha,RRRC} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,(ddf_H}\sqrt{\frac{2}{J}(MS(TR) + J\max(Cov_2 - Cov_3, 0))} \tag{2.38}$$

2.3.4 Fixed-reader random-case (FRRC) analysis

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

$$F_{ORH|R} = \frac{MS(T)}{Var - Cov_1 + (J-1)\max(Cov_2 - Cov_3, 0)} \tag{2.39}$$

$F_{ORH|R}$, a realization (i.e., observation) of a random variable, is distributed as an F-statistic with:

$$\left.\begin{array}{c} ndf = I - 1 \\ ddf = \infty \\ F_{ORH|R} \sim F_{ndf,ddf} \end{array}\right\} \tag{2.40}$$

Alternatively, as with (2.16),

$$(I-1)F_{ORH|R} \sim t_{I-1}$$

For $J = 1$, (2.39) reduces to (2.17).

The critical value of the statistic is $F_{1-\alpha,I-1,\infty}$ which is that value such that fraction $(1-\alpha)$ of the area under 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, i.e.,:

$$F_{ORH|R} > F_{1-\alpha,I-1,\infty}$$

The p-value of the test is the probability that a random sample from the distribution $F_{I-1,\infty}$ exceeds the observed value of the F statistic defined in (2.39):

$$p = \Pr(F > F_{ORH|R} \mid F \sim F_{I-1,\infty}) \tag{2.41}$$

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

$$CI_{1-\alpha,FRRC} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,\infty} \sqrt{\frac{2}{J}(Var - Cov_1 + (J-1)\max(Cov_2 - Cov_3, 0))} \tag{2.42}$$

One can think of the numerator terms on the right hand side of (2.42) as the variance of the inter-treatment FOM difference per reader, and the division by J is needed as the readers, as a group, have smaller variance in inverse proportion to their numbers.

The NH is rejected if any of the following equivalent conditions is met:

- The observed value of the F-statistic exceeds the critical value $F_{1-\alpha,I-1,\infty}$.
- The p-value defined by (2.41) is less than α.
- The $(1-\alpha)$ confidence interval does not include zero.

Notice that for J = 1, (2.42) reduces to (2.21).

2.3.5 Random-reader fixed-case (RRFC) analysis

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:

$$F_{ORH|C} = \frac{MS(T)}{MS(TR)} \tag{2.43}$$

$F_{ORH|C}$ is distributed as an F-statistic with:

$$\left.\begin{array}{c} ndf = I-1 \\ ddf = (I-1)(J-1) \\ F_{ORH|C} \sim F_{ndf,ddf} \end{array}\right\} \tag{2.44}$$

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 is given by:

$$CI_{1-\alpha,RRFC} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,(I-1)(J-1)} \sqrt{\frac{2}{J} MS(TR)} \tag{2.45}$$

It is time to reinforce the formulae with examples.

2.3.6 Single-treatment multiple-reader analysis

Suppose one has data in a single treatment i and multiple readers are involved. One wishes to determine if the performance of the readers as a group equals some specified value. *Since only a single treatment is involved, an implicit i dependence in subsequent formulae is ignored.*

In 2.2 single-reader multiple-treatment analysis was described. It is not identical to single-treatment multiple-reader analysis. Treatment is a fixed factor while reader is a random factor. Therefore, one cannot simply use the previous analysis with reader and treatment interchanged (a graduate student tried to do just that, and he is quite smart, hence the reason for this warning; one can use the previous analysis if reader is regarded as a fixed factor, and a function in RJafroc called TBA `StSignificanceTestingSingleFixedFactor()` does just that).

In the analysis described in this section reader is regarded as a random effect. The average performance of the readers is estimated and compared to a specified value. Hillis has described the appropriate modifications. [TBA Two approaches are described, one using the DBM pseudovalue based model and

the other based on the OR model with appropriate modification. The second approach is summarized below. TBA]

For single-treatment multiple-reader ORH analysis, the figure of merit model is (contrast the following equation to (2.1) noting the absence of an i index. If multiple modalities are present the current analysis is applicable to data in each treatment analyzed one at a time):

$$\theta_{j\{c\}} = \mu + R_j + \epsilon_{j\{c\}} \tag{2.46}$$

One wishes to test the NH: $\mu = \mu_0$ where μ_0 is some pre-specified value. (since $C = 1$, in the interest of brevity one can suppress the c index):

$$\mu = \theta_\bullet \tag{2.47}$$

The variance of the reader-averaged FOM can be shown (Obuchowski and Rockette, 1995) to be given by (the reference is to the original OR publication, specifically Eqn. 2.3):

$$\sigma^2_{\theta_\bullet} = \frac{1}{J}(\sigma^2_R + Var + (J-1)Cov_2) \tag{2.48}$$

2.3.7 Connection to existing literature

Rather than attempt to derive the preceding equation, it is shown how it follows from the existing literature (Obuchowski and Rockette, 1995). For convenience Eqn. 2.3 in cited reference is reproduced below.

$$Var(\theta_{i\bullet\bullet}) = \frac{1}{J}(\sigma^2_b + \sigma^2_{ab} + (\sigma^2_w/K) + \sigma^2_c(1 + J(J-1)r_2)) \tag{2.49}$$

In the OR notation, the FOM has three indices, θ_{ijk}. One deletes the i index as one is dealing with a single treatment and one can drop the average over the k index, as one is dealing with a single dataset; σ^2_b in the OR notation is what we are calling σ^2_R; for single treatment the treatment-reader interaction term σ^2_{ab} is absent; and for single "replication" the term σ^2_w/K (in OR notation K is the number of replications) is absent, or, more accurately, the within-reader variance σ^2_w is absorbed into the case sampling variance σ^2_c as the two are inseparable); the term σ^2_ϵ is what we are calling Var; and $\sigma^2_c r_2$ in OR paper is what we are calling Cov_2.

An alternative first principles derivation, due to Mr. Xuetong Zhai, is given in TBA Online Appendix 10.E.

One needs to replace σ^2_R in (2.50) with an expected value. Again, rather than attempt to derive the following equation, it is shown how it follows from the

existing literature (Hillis, 2014). We start with Table I ibid: this is a table of expected means squares for the OR model, analogous to TBA Table 9.1 in Chapter 09, for the DBM model. For a single treatment (in the notation of the cited reference, $t = 1$ and the treatment-reader variance component goes away and the term σ_ϵ^2 is what we are calling Var), it follows that:

$$E(MS(R)) = \sigma_R^2 + Var = Cov_2$$

Substituting this equation in (2.50) yields,

$$\sigma_{\theta_\bullet}^2 = \frac{1}{J}(E(MS(R)) + JCov_2) \tag{2.50}$$

An estimate of $MS(R)$ is given by (from here on it is understood that MSR is an estimate defined by:

$$MS(R) = \frac{1}{J-1}\sum_{j=1}^{J}(\theta_j - \theta_\bullet)^2 \tag{2.51}$$

Replacing the expected mean-square value with the estimate and avoiding negative covariance, which could lead to a negative variance estimate, one has:

$$\sigma_{\theta_\bullet}^2 = \frac{1}{J}(MS(R) + J\max(Cov_2, 0)) \tag{2.52}$$

The observed value of the t-statistic for testing the NH is t_{1T} (the supbscript means that this statistic applies to single treatment analysis):

$$t_{1T} = \frac{\mu - \mu_0}{\sigma_{\theta_\bullet}} = (\theta_\bullet - \mu_0)\sqrt{\frac{J}{(MS(R) + J\max(Cov_2, 0)}} \tag{2.53}$$

This is distributed as a t-statistic with $df_H^{I=1}$ degrees of freedom:

$$t_{1T} \sim t_{df_H^{1T}} \tag{2.54}$$

In the above equation, Hillis single-treatment degree of freedom $t_{df_H^{1T}}$ is defined by (Hillis, 2014):

$$df_H^{1T} = (J-1)\left[\frac{MS(R) + J\max(Cov_2, 0)}{MS(R)}\right]^2 \tag{2.55}$$

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

$$p = \Pr(t > |t| \mid t \sim t_{df_H^{1T}}) \tag{2.56}$$

Therefore, a $100 \times (1 - \alpha)$ percent confidence interval for $\theta_\bullet - \mu_0$ is:

$$\theta_\bullet - \mu_0 \pm t_{\alpha/2, df_H^{1T}} \sqrt{\frac{MS(R) + \max(Cov_2, 0)}{J}} \tag{2.57}$$

The single treatment method is implemented in mainSingleTreatment.R. The relevant code is listed in Online Appendix 10.F. Source the code to get the following output.

2.4 Discussion/Summary

This chapter described the Obuchowski-Rockette method as modified by Hillis. As noted earlier, it has the same number of parameters as the DBMH method described in the preceding chapter, but the model (2.22) *appears* simpler as some terms are "hidden" in the structure of the error term. In this chapter the NH condition was considered. Extension to the alternative hypothesis, i.e., estimating statistical power, is deferred to online appendices to Chapter 11. The extension is a little simpler with the DBMH model, as it is a standard ANOVA model. For example the expressions for the DBMH non-centrality parameter was readily defined in Chapter 09, e.g., §9.7.4. Hillis has derived expressions allowing transformation between quantities in the two methods, and this is the approach adopted in this book and implemented in the cited online appendix.

Online Appendix 10.A describes R implementation of the DeLong method for estimating the covariance matrix for empirical AUC. Since the main difficulty understanding the original OR method is conceptualizing the covariance matrix, the author has explained this at an elementary level, using a case-set index which is implicit in the original OR paper4. This was the reason for the gentle introduction analyzing performance of a single reader in multiple treatments. The jackknife, bootstrap and the DeLong methods, all implemented in Online Appendix 10.B, should reinforce understanding of the covariance matrix. The DBM and ORH methods are compared for this special case in Online Appendix 10.C. A minimal implementation of the ORH method for MRMC data is given in Online Appendix 10.D, which is a literal implementation of the relevant formulae. The special case of multiple readers in a single treatment is coded in Online Appendix 10.F. This will be used in Chapter 22 where standalone CAD performance is compared to a group of radiologists interpreting the same cases.

The original publication (Dorfman et al., 1992) and a subsequent one (Obuchowski and Rockette, 1995) were major advances. Hillis' work showing their equivalence unified the two apparently disparate analyses, and this was a major

advance. The Hillis papers, while difficult reads, are ones the author goes to repeatedly.

This concludes two methods used to analyze ROC MRMC datasets. A third method, restricted to the empirical AUC, is also available (Clarkson et al., 2006; Kupinski et al., 2006; Gallas, 2006; Gallas et al., 2009). As noted earlier, the author prefers methods that are applicable to other estimates of AUC, not just the empirical area, and to other data collection paradigms.

The next chapter takes on the subject of sample size estimation using either DBMH or the ORH method.

2.5 References

Bibliography

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.

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*, 44:837–845.

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.

Gallas, B. D. (2006). One-shot estimate of mrmc variance: Auc. *Academic Radiology*, 13(3):353–362.

Gallas, B. D., Bandos, A., Samuelson, F. W., and Wagner, R. F. (2009). A framework for random-effects roc analysis: Biases with the bootstrap and other variance estimators. *Communications in Statistics - Theory and Methods*, 38(15):2586 – 2603.

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.

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.

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. (1997). Variance-component modeling in the analysis of receiver operating characteristic index estimates. *Acad. Radiol.*, 4(8):587–600.