Chapter 10: Obuchowski-Rockette-Hillis (ORH) analysis

Table of Contents

Online Supplementary Material

A. Online Appendix 10.A: The DeLong method for estimating the covariance matrix
B. Online Appendix 10.B: Estimation of covariance matrix: single-reader multiple-treatment
C. Online Appendix 10.C: Comparing DBMH and ORH methods for single-reader multiple-treatment
D. Online Appendix 10.D: Minimal implementation of ORH method
E. Online Appendix 10.E: Proof of Eqn. (10.64).
F. Online Appendix 10.F: Single-treatment multiple-reader analysis

10.1: Introduction

The previous chapter described the DBM significance testing procedure[1] for analyzing MRMC ROC data, along with improvements suggested by Hillis. Because the method depends on the assumption 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[2-4]. The physicist in the author believes that 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 FOM-like quantities; this property was also noted in 1997 by Hanley and Hajian-Tilaki[5]. However, this justification only applies to the empirical AUC, so an alternate approach is desirable.

This chapter presents Hillis' preferred alternative to the DBMH approach. He has shown that the DBMH method can be regarded as a "working model that gives the right results", but a method based on an earlier

publication[6] 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[7] that are outside the scope of this book (the author acknowledges a private communication from Dr. Obuchowski, ca. 2006, that demonstrates the flexibility afforded by the OR approach) and it is possible that applications to other paradigms (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 for the "gentle" introduction is that, in the author's opinion, the method is rather opaque to the user community (as distinct from statisticians). Part of the problem is the notation, namely lack of usage of the case-set index $\{c\}$, which while implicit to statisticians, it's absence can be confusing to those from other disciplines, e.g., physics, as in the author's case. The notational issue is highlighted in a key difference of the Obuchowski and Rockette method from DBMH, namely in how the error term is modeled by a covariance matrix. In this chapter the structure of the covariance matrix is examined in detail, as it is key to understanding the ORH method.

In the first step of the gentle 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. As before, for organizational reasons illustrative **R** code is relegated to Appendices, but is essential reading.

10.2: Single-reader multiple-treatment model

Consider a single-reader providing ROC interpretations of a common case $\{c\}$ set in multiple-treatments i (i = *1, 2, ..., I*). Before proceeding, we note that this is *not* homologous (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:

$$\theta_{i\{c\}} = \mu + \tau_i + \varepsilon_{i\{c\}} \quad . \qquad \textbf{(10.1)}$$

In the Obuchowski and Rockette method [6] one models the figure-of-merit, *not the pseudovalues*, indeed this is the key difference from the DBMH method.

Recall that $\{c\}$ denotes a set of cases. Eqn. (10.1) 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:

$$\sum_{i=1}^{I} \tau_i = 0 \quad . \qquad \textbf{(10.2)}$$

The *c*-index was introduced in **Chapter 07**. The left hand side of Eqn. (10.1) is the figure-of-merit $\theta_{i\{c\}}$ for treatment *i* and case-set index $\{c\}$, where $c = 1, 2, \ldots, C$ denote 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 variability.

Obuchowski and Rockette use a *k* index, defined as the "*k^{th} 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 direct way to this end 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 new comers to this field have difficulty understanding what is going on, there is an additive random error term $\varepsilon_{i\{c\}}$ whose sampling behavior is described by a *multivariate normal* distribution with an *I*-dimensional zero mean vector and an *I x I* dimensional covariance matrix Σ:

$$\boldsymbol{\varepsilon}_{i\{c\}} \sim N_I\left(\vec{0}, \Sigma\right) \quad . \qquad \textbf{(10.3)}$$

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 model, but here one restricts to the simpler one):

$$\Sigma \equiv Cov\left(\boldsymbol{\varepsilon}_{i\{c\}}, \boldsymbol{\varepsilon}_{i'\{c\}}\right) = \begin{cases} Var & i = i' \\ Cov_1 & i \neq i' \end{cases} \quad . \qquad \textbf{(10.4)}$$

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

$$\Sigma = \begin{pmatrix} Var & Cov_1 & \ldots & Cov_1 & Cov_1 \\ Cov_1 & Var & \ldots & Cov_1 & Cov_1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ Cov_1 & Cov_1 & \ldots & Var & Cov_1 \\ Cov_1 & Cov_1 & \ldots & Cov_1 & Var \end{pmatrix} \quad . \qquad \textbf{(10.5)}$$

If $I = 2$ then Σ is a symmetric 2 x 2 matrix, whose diagonal terms are the common variances in the two treatments (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 covariances to depend on the specific treatment pairings. For

obvious reasons ("Occam's Razor[a]" or the law of parsimony[b]) one wishes to start with the simplest model that, one hopes, captures essential characteristics of the data.

Some elementary statistical results are presented next.

10.2.1: Definitions of covariance and correlation

The covariance of two scalar random variables X and Y is defined by[8]:

$$Cov(X,Y) = E\left[\left(X - E(X)\right)\left(Y - E(Y)\right)\right] = E(XY) - E(X)E(Y) \quad . \qquad \textbf{(10.6)}$$

$E(X)$ is the expectation value of X, i.e., the integral of x multiplied by its *pdf*:

$$E(X) = \int x \,\mathrm{pdf}(x)\, dx \quad . \qquad \textbf{(10.7)}$$

The integral is over the range of x. The covariance can be thought of as variance of two random variables that is *common* to both of them. The variance, a special case of covariance, of X is defined by:

$$Var(X) \equiv Cov(X,X) = E\left(X^2\right) - \left(E(X)\right)^2 = \sigma_X^2 \quad . \qquad \textbf{(10.8)}$$

It can be shown using the Cauchy–Schwarz inequality[9]:

$$\left|Cov(X,Y)\right|^2 \le Var(X)Var(Y) \quad . \qquad \textbf{(10.9)}$$

A related quantity, the correlation ρ is defined by (the $\sigma's$ are standard deviations):

$$\rho_{xy} \equiv Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad . \qquad \textbf{(10.10)}$$

According to Eqn. (10.9) it has the property:

---

[a] A scientific and philosophic rule that entities should not be multiplied unnecessarily which is interpreted as requiring that the simplest of competing theories be preferred to the more complex or that explanations of unknown phenomena be sought first in terms of known quantities (Merriam-Webster dictionary).

[b] The scientific principle that things are usually connected or behave in the simplest or most economical way.

$$\left|\rho_{xy}\right| \le 1 \quad . \qquad \textbf{(10.11)}$$

For perfect correlation, ρ_{xy} equals one and for perfect anti-correlation ρ_{xy} equals minus one. For uncorrelated variables, ρ_{xy} equals zero. Statistical independence implies zero correlation but the converse is not true.

10.2.2: Special case applicable to Eqn. (10.4)

Assuming X and Y have the same variance:

$$Var(X) = Var(Y) = Var \equiv \sigma^2 \quad . \qquad \textbf{(10.12)}$$

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

$$\left.\begin{aligned} &Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y) \\ &= 2\left(Var - Cov_1\right) \end{aligned}\right\} \quad . \qquad \textbf{(10.13)}$$

The first line of the above equation is general, the second line specializes to the OR single-reader multiple-treatment model where the variances are equal and likewise all covariances in Eqn. (10.5) 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):

$$\rho_1 = \frac{Cov_1}{Var} \quad . \qquad \textbf{(10.14)}$$

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 one need only show elements at and above the diagonal):

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & & \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & & & \sigma^2 & \rho_1\sigma^2 \\ & & & & \sigma^2 \end{pmatrix} \quad . \qquad \textbf{(10.15)}$$

An unbiased estimate of the covariance Eqn. (10.4) follows from:

$$\Sigma_{ii'} = \frac{1}{C-1}\sum_{c=1}^{C}\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i'\{c\}} - \theta_{i'\{\bullet\}}\right) \quad . \qquad \textbf{(10.16)}$$

Sampling different case-sets, as required by Eqn. (10.16), 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 **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 number of bootstraps, then the estimation formula is:

$$\Sigma_{ii'}\big|_{bs} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\theta_{i\{b\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i'\{b\}} - \theta_{i'\{\bullet\}}\right) \quad . \qquad \textbf{(10.17)}$$

The bootstrap method of estimating the covariance matrix, Eqn. (10.17), is a direct translation of Eqn. (10.16). Alternatively one could have used the jackknife FOM values $\theta_{i(k)}$, i.e., the figure of merit with a particular case removed, to estimate the covariance matrix:

$$\Sigma_{ii'}\big|_{jk} = \frac{(K-1)^2}{K}\left[\frac{1}{K-1}\sum_{k=1}^{K}\left(\theta_{i(k)} - \theta_{i(\bullet)}\right)\left(\theta_{i'(k)} - \theta_{i'(\bullet)}\right)\right] \quad . \qquad \textbf{(10.18)}$$

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 simply 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 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. Furthermore, the expression for the jackknife covariance contains a *variance inflation factor*:

$$\frac{(K-1)^2}{K} \quad . \qquad \textbf{(10.19)}$$

This factor multiplies the traditional expression for the covariance[10], shown in square brackets in Eqn. (10.18). A third method of estimating the covariance, namely the DeLong et al. method[11], applicable to the empirical AUC, is described later.

10.2.4: Meaning of the covariance matrix in Eqn. (10.5)

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:

$$\sigma_i^2 \equiv Var_i = \frac{1}{C-1}\sum_{c=1}^{C}\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right) \quad . \qquad \textbf{(10.20)}$$

The process is repeated for all treatments and the *I*-variances are averaged. This is the final estimate of *Var* appearing in Eqn. (10.5).

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 $\theta_{i'\{c\}}$:

$$\rho_{1;ii'} = \frac{1}{C-1}\frac{\sum_{c=1}^{C}\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i'\{c\}} - \theta_{i'\{\bullet\}}\right)}{\sigma_i \sigma_{i'}} \quad . \qquad \textbf{(10.21)}$$

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 Eqn. (10.15). According to Eqn. (10.14) one expects the covariance to be smaller than the variance determined as in the previous paragraph.

In most situations one expects ρ_1 to be positive. There is, perhaps unlikely, a scenario that could lead to anti-correlation and negative ρ_1. 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. The author is unaware of a practical demonstration of this expectation.

10.2.5: Code illustrating the covariance matrix

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[11]. In **Chapter 07**, these methods were described in the context of estimating the variance of AUC. Eqn. (10.17) and Eqn. (10.18) 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 and implemented in file **`VarCovMtrxDLStr.R`**. It has been confirmed by the author that the implementation of the DeLong method[11] in file **`VarCovMtrxDLStr.R`** gives *identical* results to those yielded by the SAS macro attributed to DeLong. The file name stands for "variance covariance matrix according to the DeLong structural components method" described in five unnumbered equation following Eqn. 4 in the cited reference.

The jackknife, bootstrap and the DeLong methods are used in file **`mainVarCov1.R`**, a listing and explanation of which appears in Online Appendix 10.B. **`Source`** the file yielding the following code output:

10.2.5.1: Code Output

```
> source('~/book2/03 B Statistics of ROC analysis/B3 ORH Analysis/software/mainVarCov1.R')
data file =  CXRinvisible3-20mm.xlsx
number of treatments =  4 , number of non-diseased cases =  52 , number of diseased cases =  106
reader =  1
OR variance components using jackknife resampling
Variance =  0.001614554 , Cov1 =  0.0004970402 , rho =  0.3078498
OR variance components using bootstrap resampling
Variance =  0.001575106 , Cov1 =  0.0005271459 , rho =  0.3346733
OR variance components using DeLong method
Variance =  0.001600124 , Cov1 =  0.0004926574 , rho =  0.3078871
```

The dataset is from a recent study comparing 2D digital chest x-rays, 2D dual energy and 3D digital chest tomosynthesis[12], that was used to demonstrate DBMH analysis in §9.11. The output shows that while all three estimates of Var and Cov_1 are comparable to within 10%, the DeLong and jackknife estimates are very close (to within 1%; the correlations are even closer). There is seed dependence associated with the bootstrap, but not with the jackknife (estimating sampling variability of a jackknife estimate requires other techniques[10]). For example, running the code with a different **`seed`** will lead to a different bootstrap estimate, but the jackknife estimate, which does not involve random sampling, rather a systematic leave-one-out procedure, is unaffected.

10.2.6: Significance testing

Why does one go through the trouble of estimating the covariance matrix? The simple reason is that it is needed for significance testing. Define the mean square corresponding to the treatment effect, denoted MST, by:

$$MST = \frac{1}{I-1}\sum_{i=1}^{I}\left(\theta_i - \theta_\bullet\right)^2 \quad . \qquad \textbf{(10.22)}$$

Unlike the previous chapter, all mean square quantities defined in this chapter are based on FOMs; specifically, they are not based on pseudovalues. Converting between them is described in Ref. [2-4] and is implemented in the **RJafroc** package.

It can be shown[2] that under the null hypothesis that all treatments have identical performances the test statistic F_{1R} defined below (the *1R* subscript is meant to denote single-reader analysis) is distributed approximately as a central F-distribution with $I-1$ numerator degrees of freedom (*ndf)* and infinite denominator degrees of freedom (*ddf*), i.e.,

$$\left.\begin{aligned} \frac{(I-1)MST}{Var-Cov_1} &\sim \chi^2_{I-1} \\ F_{1R} \equiv \frac{MST}{Var-Cov_1} &\sim F_{I-1,\infty} \end{aligned}\right\} \quad . \qquad \textbf{(10.23)}$$

[The first form is from §5.4 ibid. with two other covariance terms "zeroed out" because they are multiplied by $J-1=0$. Dividing a χ^2 distributed random variable with $I-1$ degrees of freedom by $I-1$ yields an F-distributed random variable with *ndf* = $I-1$ and *ddf* = ∞, as in the second form in Eqn. (10.23). Here is an **R** example: **pf(3.1,4,Inf)** = 0.9853881; **pchisq(3.1*4,4)** = 0.9853881. The first form shows that the CDF of the F-distribution with 4 and infinite degrees of freedom at 3.1 equals the CDF of the χ^2 distribution with 4 degrees of freedom at 3.1 times 4. A little "mulling over it" should convince the reader about the truth of these statements.]

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

$$p \equiv P\left(f \geq F_{1R} \,|\, f \sim F_{I-1,\infty}\right) \quad . \qquad \textbf{(10.24)}$$

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

$$CI_{1-\alpha} = \left(\theta_{i\bullet} - \theta_{i'\bullet}\right) \pm t_{\alpha/2;\infty}\sqrt{2\left(Var - Cov_1\right)} \quad . \qquad \textbf{(10.25)}$$

Comparing Eqn. (10.25) to Eqn. (10.13) shows that the term $\sqrt{2\left(Var - Cov_1\right)}$ is the standard error of the inter-treatment FOM difference; this should make intuitive sense; the covariance (i.e., scaled correlation) tends to reduce the variance of the difference (recall the analogy from the Introduction of **Chapter 09** due to the late Dr. Wagner). The multiplier $t_{\alpha/2;\infty}$ equals 1.96. One has probably encountered the rule that a confidence interval is plus or minus two standard deviations from the central value; the "2" comes from rounding up 1.96, good enough, as they say, for "government work".

10.2.7: 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 analysis described in §9.8 of the previous chapter (with just one reader the conclusions apply to the specific reader, so reader must be a fixed factor). **`Source`** the file **`MainOrDbmh1R.R`**, a listing of which appears in Online Appendix 10.C. For convenience, a few relevant lines are shown here:

10.2.7.1: Code listing (partial)

```
rm(list = ls()) #mainOrDbmh1R.R
library(RJafroc)
...
ret1 <- SignificanceTesting(rocData,fom = "Wilcoxon", method = "DBMH", option = "FRRC")
cat("DBMH: F-stat = ", ret1$fFRRC, ", ddf = ", ret1$ddfFRRC, ", P-val = ", ret1$pFRRC,"\n")

ret2 <- SignificanceTesting(rocData,fom = "Wilcoxon", method = "ORH", option = "FRRC")
cat("ORH (Jackknife):  F-stat = ", ret2$fFRRC, ", ddf = ", ret2$ddfFRRC, ", P-val = ",
ret2$pFRRC,"\n")

ret3 <- SignificanceTesting(rocData,fom = "Wilcoxon", method = "ORH", option = "FRRC",
                            covEstMethod = "DeLong")
cat("ORH (DeLong):  F-stat = ", ret3$fFRRC, ", ddf = ", ret3$ddfFRRC, ", P-val = ",
ret3$pFRRC,"\n")

ret4 <- SignificanceTesting(rocData,fom = "Wilcoxon", method = "ORH", option = "FRRC",
                            covEstMethod = "Bootstrap")
cat("ORH (Bootstrap):  F-stat = ", ret4$fFRRC, ", ddf = ", ret4$ddfFRRC, ", P-val = ",
ret4$pFRRC,"\n")
```

The code illustrates different ways of performing the significance testing; Fig. 10.1 shows the help-screen for this function. Limners 15 – 16 selects reader 1 data for the four treatments in this dataset (the reader should experiment with different choices of selected reader). The first form illustrates usage of the function

SignificanceTesting() with method specified as "**DBMH**" and option specified as "**FRRC**", for DBMH fixed-reader random-case analysis (with one reader, as always, one must regard reader as a fixed factor). The second form uses the same function with method specified as "**ORH**", which uses the default jackknife method for estimating the covariance matrix. The third form overrides the default with **covEstMethod = "DeLong"**, which uses the DeLong method for estimating the covariance matrix. The last form uses the **covEstMethod = "Bootstrap"**, which uses the bootstrap method with 200 bootstraps, the default, which can be overridden, as shown in Fig. 10.1, using option **nBoots**.

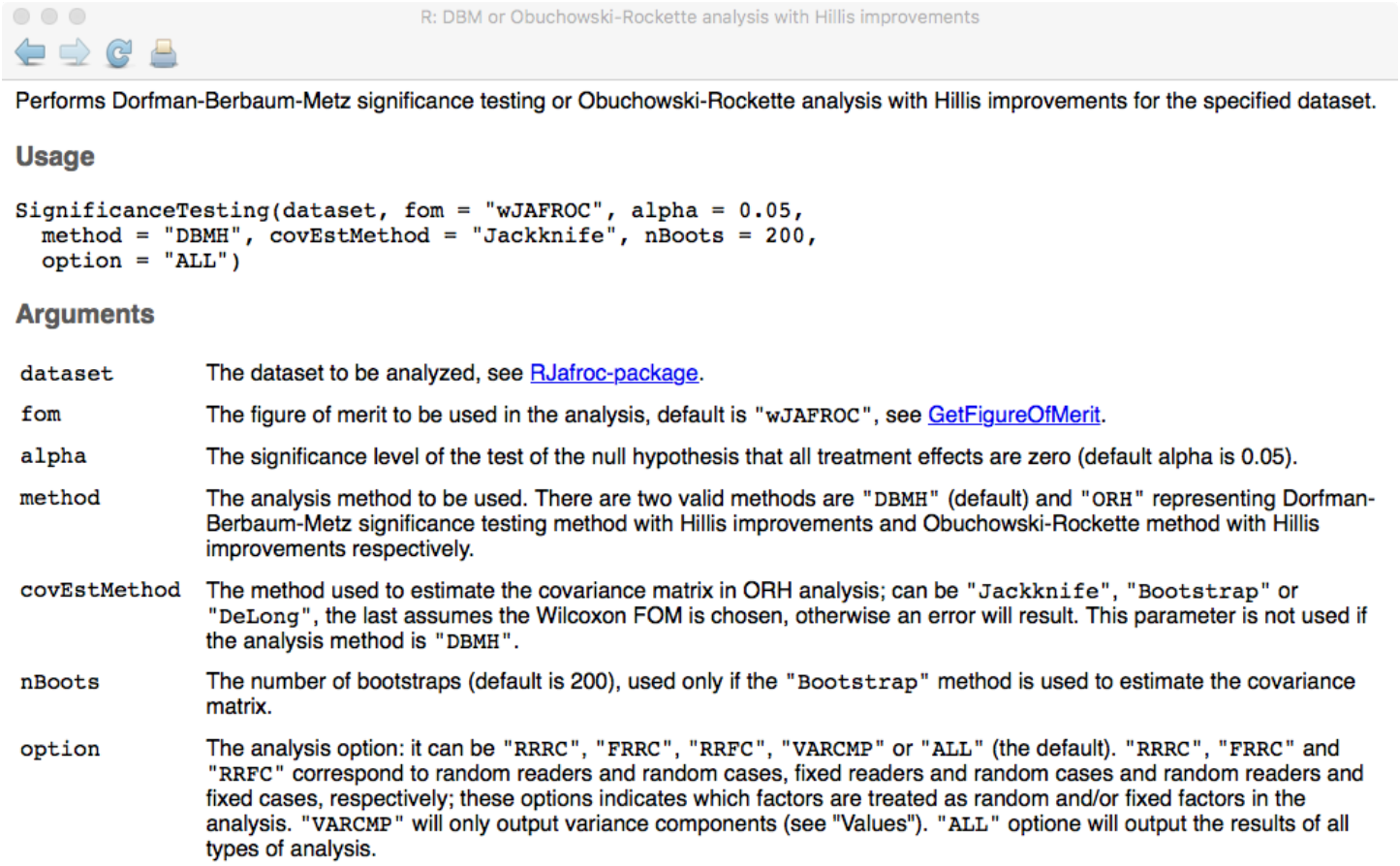

Fig. 10.1: This figure shows a screen-shot of the help page on the **SignificanceTesting()** function. It illustrates different usages of the function illustrated in §10.2.7.1 and §10.2.7.2.

Sourcing the code yields the following output:

10.2.7.2: Code Output

```
> source('~/Desktop/book3/03 B Statistics of ROC analysis/B10 ORH
Analysis/software/mainOrDbmh1R.R')
data file =  CXRinvisible3-20mm.xlsx
selected reader =  1
DBMH: F-stat =  2.200775 , ddf =  471 , P-val =  0.0871945
ORH (Jackknife):  F-stat =  2.200775 , ddf =  Inf , P-val =  0.08571326
ORH (DeLong):  F-stat =  2.220742 , ddf =  Inf , P-val =  0.08347962
ORH (Bootstrap):  F-stat =  2.09761 , ddf =  Inf , P-val =  0.09820081
```

The output lists the results for four methods. Listed first are the results using the DBMH fixed-reader method. The next three lines list the results of the ORH method using different methods of estimating the covariance matrix: the jackknife, the DeLong and the bootstrap method. The F-statistics used by the DBMH and the ORH/jackknife methods are identical; this is because the jackknife method was used to estimate the covariance matrix needed for the OR method and the DBMH approach always uses jackknife-derived pseudovalues. However, the degrees of freedom are different. The 471 in DBMH comes from $(I-1)(K-1)$ = 3 x 157 whereas in the OR method it is infinite; see Eqn. 22 and 23 in cited paper[2]. Because *K* is generally a large number the effect of the difference in degrees of freedom on p-values is minimal. As expected, the DeLong method gives results very similar to the jackknife method, while the bootstrap yields slightly different results, due to the different method used to estimate the covariance matrix.

The demonstration should convince one that the "replication" index in the original Obuchowski and Rockette publication[6] is being interpreted correctly, namely as in Eqn. (10.1), a case-set index is needed, not a "replication" index.

10.3: 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 Eqn. (10.1), 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 appropriately to account for the additional reader factor.

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

$$\theta_{ij\{c\}} = \mu + \tau_i + R_j + (\tau R)_{ij} + \varepsilon_{ij\{c\}} \quad . \qquad \textbf{(10.26)}$$

The fixed treatment effect τ_i is subject to the usual constraint, Eqn. (10.2). The first two terms on the right hand side of Eqn. (10.26) 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 is a 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:

$$\left.\begin{array}{l} R_j \sim N\left(0,\sigma_R^2\right) \\ (\tau R)_{ij} \sim N\left(0,\sigma_{\tau R}^2\right) \end{array}\right\} \quad . \qquad \textbf{(10.27)}$$

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):

$$\mu = \theta_{\bullet\bullet\{1\}} \quad . \qquad \textbf{(10.28)}$$

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 τ_i:

$$\tau_i = \theta_{i\bullet\{1\}} - \theta_{\bullet\bullet\{1\}} \quad . \qquad \textbf{(10.29)}$$

The τ_i estimates obey the sum rule Eqn. **Error! Reference source not found.**. For example, with two treatments, the values of τ_i must be the negatives of each other.

The error term on the right hand side of Eqn. (10.26) is more complex than the corresponding DBM model error term. Obuchowski and Rockette model this term with a multivariate normal distribution with a length $(I \times J)$ zero-mean vector and a $(I \times J) \times (I \times J)$ covariance matrix Σ. In other words,

$$\varepsilon_{ij\{c\}} \sim N_{I \times J}\left(\vec{0}, \Sigma\right) \quad . \qquad \textbf{(10.30)}$$

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

$$Cov\left(\boldsymbol{\varepsilon}_{ij\{c\}},\boldsymbol{\varepsilon}_{i'j'\{c\}}\right)=\begin{cases} Var & i=i',j=j' \\ Cov_1 & i\neq i',j=j' \\ Cov_2 & i=i',j\neq j' \\ Cov_3 & i\neq i',j\neq j' \end{cases} \quad . \qquad \textbf{(10.31)}$$

Apart from fixed effects, the model in Eqn. (10.31) contains 6 parameters:

$$\sigma_R^2,\sigma_{\tau R}^2,Var,Cov_1,Cov_2,Cov_3 \quad . \qquad \textbf{(10.32)}$$

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 Eqn. (10.26) "looks" simpler because four covariance terms are "hidden" in the $\boldsymbol{\varepsilon}$ 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,Cov_3 estimates need to be averaged over all applicable treatment-reader combinations.

10.3.1: 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:

$$\Sigma = \begin{pmatrix} (11,11) & (12,11) & (13,11) & (21,11) & (22,11) & (23,11) \\ (11,12) & (12,12) & (13,12) & (21,12) & (22,12) & (23,12) \\ (11,13) & (12,13) & (13,13) & (21,13) & (22,13) & (23,13) \\ (11,21) & (12,21) & (13,21) & (21,21) & (22,21) & (23,21) \\ (11,22) & (12,22) & (13,22) & (21,22) & (22,22) & (23,22) \\ (11,23) & (12,23) & (13,23) & (21,23) & (22,23) & (23,23) \end{pmatrix} \quad . \qquad \textbf{(10.33)}$$

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. (10.31) the entries in these cells would be Var.

$$\Sigma = \begin{pmatrix} Var & (12,11) & (13,11) & (21,11) & (22,11) & (23,11) \\ & Var & (13,12) & (21,12) & (22,12) & (23,12) \\ & & Var & (21,13) & (22,13) & (23,13) \\ & & & Var & (22,21) & (23,21) \\ & & & & Var & (23,22) \\ & & & & & Var \end{pmatrix} \quad . \qquad \textbf{(10.34)}$$

According to Eqn. (10.31) 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, 2 and 1, so it is replaced by Cov_1):

$$\Sigma = \begin{pmatrix} Var & (12,11) & (13,11) & Cov_1 & (22,11) & (23,11) \\ & Var & (13,12) & (21,12) & Cov_1 & (23,12) \\ & & Var & (21,13) & (22,13) & Cov_1 \\ & & & Var & (22,21) & (23,21) \\ & & & & Var & (23,22) \\ & & & & & Var \end{pmatrix} \quad . \qquad \textbf{(10.35)}$$

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

$$\Sigma = \begin{pmatrix} Var & Cov_2 & Cov_2 & Cov_1 & (22,11) & (23,11) \\ & Var & Cov_2 & (21,12) & Cov_1 & (23,12) \\ & & Var & (21,13) & (22,13) & Cov_1 \\ & & & Var & Cov_2 & Cov_2 \\ & & & & Var & Cov_2 \\ & & & & & Var \end{pmatrix} \quad . \qquad \textbf{(10.36)}$$

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

$$\Sigma = \begin{pmatrix} Var & Cov_2 & Cov_2 & Cov_1 & Cov_3 & Cov_3 \\ & Var & Cov_2 & Cov_3 & Cov_1 & Cov_3 \\ & & Var & Cov_3 & Cov_3 & Cov_1 \\ & & & Var & Cov_2 & Cov_2 \\ & & & & Var & Cov_2 \\ & & & & & Var \end{pmatrix} \quad . \qquad \textbf{(10.37)}$$

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:

$$Var = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} \left(\theta_{ij\{c\}} - \theta_{ij\{\bullet\}} \right) \left(\theta_{ij\{c\}} - \theta_{ij\{\bullet\}} \right) \right\rangle_{ij} \quad . \qquad \textbf{(10.38)}$$

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 Eqn. (10.38) does not have 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:

$$\left.\begin{aligned} Cov_1 &= \left\langle \frac{1}{C-1}\sum_{c=1}^{C}\left(\theta_{ij\{c\}}-\theta_{ij\{\bullet\}}\right)\left(\theta_{i'j\{c\}}-\theta_{i'j\{\bullet\}}\right)\right\rangle_{ii'jj} \\ Cov_2 &= \left\langle \frac{1}{C-1}\sum_{c=1}^{C}\left(\theta_{ij\{c\}}-\theta_{ij\{\bullet\}}\right)\left(\theta_{ij'\{c\}}-\theta_{ij'\{\bullet\}}\right)\right\rangle_{iijj'} \\ Cov_3 &= \left\langle \frac{1}{C-1}\sum_{c=1}^{C}\left(\theta_{ij\{c\}}-\theta_{ij\{\bullet\}}\right)\left(\theta_{i'j'\{c\}}-\theta_{i'j'\{\bullet\}}\right)\right\rangle_{ii'jj'} \end{aligned}\right\} \quad . \qquad \textbf{(10.39)}$$

In Eqn. (10.39) 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:

(i) For Cov_1 one averages over all combinations of *different* treatments and *same* readers, as denoted by $\langle\ \rangle_{ii'jj}$.

(ii) For Cov_2 one averages over all combinations of *same* treatment and *different* readers, as denoted by $\langle\ \rangle_{iijj'}$.

(iii) For Cov_3 one averages over all combinations of *different* treatments and *different* readers, as denoted by $\langle\ \rangle_{ii'jj'}$.

10.3.2: Physical meanings of the covariance terms

The meanings of the different terms follow a similar description to that given in §10.2.4. 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 x 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:

$$Var \geq Cov_1 \geq Cov_2 \geq Cov_3 \quad . \qquad \textbf{(10.40)}$$

10.3.3: ORH random-reader random-case analysis

A model such as Eqn. (10.26) 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.,

$$NH: \tau_i = 0 \qquad (i = 1,2,...,I) \qquad . \qquad \textbf{(10.41)}$$

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[2-4,6,7], denoted F_{ORH}, defined by:

$$F_{ORH} = \frac{MST}{MSTR + J\max\left(Cov_2 - Cov_3, 0\right)} \qquad . \qquad \textbf{(10.42)}$$

Eqn. (10.42) incorporates Hillis' modification, which ensures that the constraint $Cov_2 \geq Cov_3$ is always obeyed and avoids negative denominators. The mean square (MS) terms are defined by (*these are calculated directly using FOM values, not pseudovalues*):

$$\left.\begin{aligned} MST &= \frac{J}{I-1}\sum_{i=1}^{I}\left(\theta_{i\bullet} - \theta_{\bullet\bullet}\right)^2 \\ MSTR &= \frac{1}{\left(I-1\right)\left(J-1\right)}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\theta_{ij} - \theta_{i\bullet} - \theta_{\bullet j} + \theta_{\bullet\bullet}\right)^2 \end{aligned}\right\} \qquad . \qquad \textbf{(10.43)}$$

In their original paper[6] Obuchowski and Rockette state that their proposed test statistic *F* (basically Eqn. (10.42) without the constraint implied by the *max* function) is distributed as an F-statistic with numerator degree of freedom $ndf = \left(I-1\right)$ and denominator degree of freedom $ddf = \left(I-1\right)\left(J-1\right)$. 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[13,14], using the unmodified version of Eqn. (10.42), 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* value is defined below (*ndf* is unchanged), with the subscript H denoting the Hillis modification:

$$ndf = I - 1 \quad . \qquad \textbf{(10.44)}$$

$$ddf_H = \frac{\left[MSTR + \max\left(J\left(Cov_2 - Cov_3 \right), 0 \right) \right]^2}{\frac{\left[MSTR \right]^2}{\left(I-1 \right)\left(J-1 \right)}} \quad . \qquad \textbf{(10.45)}$$

If $Cov_2 \le Cov_3$ 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 Eqn. (10.42), is distributed as an F-statistic with $ndf = \left(I-1 \right)$ and $ddf = ddf_H$ degrees of freedom[2-4]:

$$F_{ORH} \sim F_{ndf, ddf_H} \quad . \qquad \textbf{(10.46)}$$

10.3.4: 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 $\left(1-\alpha\right)$ of the area under the distribution lies to the left of the critical value. From the definition of F_{ORH}, rejection of the NH is more likely if $MS\left(T\right)$ increases, meaning the treatment effect is larger, $MS\left(TR\right)$ decreases (there is less contamination of the treatment effect by treatment-reader variability), the greater of Cov_2 or Cov_3 decreases (there is less contamination of the treatment effect by between-reader and treatment-reader variability), α increases (allowing a greater probability of Type I errors), *ndf* increases (the more the number of treatment pairings, the greater the chance that at least one pair will reject the NH) or ddf_H increases (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_{DBMH} 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_{DBMH}:

$$p = P\left(F \ge F_{ORH} \mid F \sim F_{ndf, ddf_H} \right) \quad . \qquad \textbf{(10.47)}$$

The $(1-\alpha)100$ percent 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):

$$CI_{1-\alpha}=(\theta_{i\bullet}-\theta_{i'\bullet})\pm t_{\alpha/2;ddf_H}\sqrt{\frac{2}{J}\left(MSTR+J\max\left(Cov_2-Cov_3,0\right)\right)} \quad . \quad \textbf{(10.48)}$$

The next section describes special cases of ORH analysis.

10.4: Special cases

The following extends the analysis to fixed-reader, fixed-case and single treatment analyses. The relevant results from Hillis papers[2-4] are quoted below.

10.4.1: Fixed-reader random-case (FRRC) analysis

Using the vertical bar notation $|R$ to denote that reader is regarded as a fixed effect[15], the appropriate F -statistic for testing the null hypothesis $NH:\tau_i=0\ (i=1,2...,I)$ is[3]:

$$F_{ORH|R}=\frac{MST}{\left[Var-Cov_1+(J-1)\max\left(Cov_2-Cov_3,0\right)\right]} \quad . \quad \textbf{(10.49)}$$

$F_{ORH|R}$, a realization of a random variable, is distributed as an F-statistic with:

$$\left.\begin{aligned} ndf&=I-1\\ ddf&=\infty \end{aligned}\right\} \quad . \quad \textbf{(10.50)}$$

$$F_{ORH|R}\sim F_{I-1,\infty} \quad . \quad \textbf{(10.51)}$$

Alternatively, as with Eqn. (10.23),

$$(I-1)F_{ORH|R}\sim\chi^2_{I-1} \quad . \quad \textbf{(10.52)}$$

For $J = 1$, Eqn. (10.49) reduces to Eqn. (10.23).

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:

$$F_{ORH|R} > F_{1-\alpha;I-1,\infty} \quad . \qquad \textbf{(10.53)}$$

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 Eqn. (10.49):

$$p = \mathrm{P}\left(F \geq F_{ORH|R} \mid F \sim F_{I-1,\infty}\right) \quad . \qquad \textbf{(10.54)}$$

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

$$CI_{1-\boldsymbol{\alpha}} = \left(\boldsymbol{\theta}_{i\bullet} - \boldsymbol{\theta}_{i'\bullet}\right) \pm t_{\boldsymbol{\alpha}/2;\infty}\sqrt{\frac{2}{J}\left[Var - Cov_1 + \left(J-1\right)\max\left(Cov_2 - Cov_3, 0\right)\right]} \quad . \qquad \textbf{(10.55)}$$

One can think of the numerator terms on the right hand side of Eqn. (10.55) 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 Eqn. (10.54) is less than α.

Notice that for *J* = 1, Eqn. (10.55) reduces to Eqn. (10.25).

10.4.2: Random-reader fixed-case (RRFC) analysis

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

$$F_{ORH|C} = \frac{MST}{MSTR} \quad . \qquad \textbf{(10.56)}$$

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

$$\left.\begin{aligned} ndf &= I-1 \\ ddf &= (I-1)(J-1) \end{aligned}\right\} \quad . \qquad \textbf{(10.57)}$$

On other words,

$$F_{ORH|C} \sim F_{I-1,(I-1)(J-1)} \quad . \qquad \textbf{(10.58)}$$

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:

$$F_{ORH|C} > F_{1-\alpha;I-1,(I-1)(J-1)} \quad . \qquad \textbf{(10.59)}$$

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

$$p = \mathrm{P}\left(F > F_{ORH|C} \mid F \sim F_{(I-1),(I-1)(J-1)}\right) \quad . \qquad \textbf{(10.60)}$$

The $(1-\alpha)$ confidence interval is given by:

$$CI_{1-\boldsymbol{\alpha}} = \left(\boldsymbol{\theta}_{i\bullet} - \boldsymbol{\theta}_{i'\bullet}\right) \pm t_{\boldsymbol{\alpha}/2;(I-1)(J-1)} \sqrt{\frac{2}{J} MSTR} \quad . \qquad \textbf{(10.61)}$$

It is time to reinforce the formulae with examples.

10.5: Example of ORH analysis

A minimal version of ORH analysis is implemented in file **`mainORH.R`** listed and explained in Online Appendix 10.D (the **`RJafroc`** package[16] has the full implementation with more detailed output, Fig. 10.2). **`Source`** this file to get the following output and Fig. 10.2:

10.5.1: Code Output

```
> source('~/Desktop/book3/03 B Statistics of ROC analysis/B10 ORH Analysis/software/mainORH.R')
alpha = 0.05
data file =  CXRinvisible3-20mm.xlsx
number of treatments =  4 , number of readers =  5 , number of non-diseased cases =  52 , number
of diseased cases =  =  106
```

```
Random reader random case analysis
Hillis ddfH =  70.52
F statistic is  13.3 and critical value of F is  2.735
pvalue =  5.645e-07
For pairing 1-2 , mean diff is  0.001959  and 95% CI is  -0.04244 0.04636
For pairing 1-3 , mean diff is  -0.1071  and 95% CI is  -0.1515 -0.06267
For pairing 2-3 , mean diff is  -0.109  and 95% CI is  -0.1534 -0.06463
For pairing 1-4 , mean diff is  -0.08806  and 95% CI is  -0.1325 -0.04366
For pairing 2-4 , mean diff is  -0.09002  and 95% CI is  -0.1344 -0.04562
For pairing 3-4 , mean diff is  0.01901  and 95% CI is  -0.02539 0.06342

Fixed reader random case analysis
ddf =  Inf
F statistic is  10.54 and critical value of F is  2.605
p-value is  6.276e-07
For pairing 1-2 , mean diff is  0.001959  and 95% CI is  -0.04707 0.05099
For pairing 1-3 , mean diff is  -0.1071  and 95% CI is  -0.1561 -0.05805
For pairing 2-3 , mean diff is  -0.109  and 95% CI is  -0.1581 -0.06001
For pairing 1-4 , mean diff is  -0.08806  and 95% CI is  -0.1371 -0.03903
For pairing 2-4 , mean diff is  -0.09002  and 95% CI is  -0.139 -0.04099
For pairing 3-4 , mean diff is  0.01901  and 95% CI is  -0.03001 0.06804

Random reader fixed case analysis
ddf =  12
F statistic is  32.25 and critical value of F is  3.49
p-value is  5.035e-06
For pairing 1-2 , mean diff is  0.001959  and 95% CI is  -0.0292 0.03312
For pairing 1-3 , mean diff is  -0.1071  and 95% CI is  -0.1382 -0.07592
For pairing 2-3 , mean diff is  -0.109  and 95% CI is  -0.1402 -0.07788
For pairing 1-4 , mean diff is  -0.08806  and 95% CI is  -0.1192 -0.0569
For pairing 2-4 , mean diff is  -0.09002  and 95% CI is  -0.1212 -0.05886
For pairing 3-4 , mean diff is  0.01901  and 95% CI is  -0.01215 0.05017
```

After listing relevant details about the data file, the numbers of treatments, readers and cases, the results are presented in three parts: random-reader random-case, fixed-reader random cases and random-reader fixed-case. For each past, listed are *ddf* (*ndf* = 3 is implicit, as there are four treatments), the observed F-statistic, the critical value of the F-statistic, the p-value for testing the NH that all treatments have identical FOMs, the observed inter-treatment FOM differences and corresponding 95% confidence intervals. As in **Chapter 09**, when more than two treatments are involved, one needs to make two sequential tests before declaring a specific inter-treatment difference as significant: the overall F-statistic has to be significant and the $(1-\alpha)$ confidence interval for the specific inter-treatment difference must not include zero. The p-value for the inter-treatment difference is calculated by **RJafroc**, but is not included in the minimal implementation.

As described in §9.11, for this dataset[12] treatment M-1 refers to 2-view digital chest x-rays (CXR) with a flat-panel detector. Treatment M-2 refers to CXR + dual energy images (DE), treatment M-3 refers to chest tomosynthesis images (TOMO) with the GE VolumeRad device, and treatment M-4 to TOMO + DE.

For RRRC the p-value is 5.6 x10^{-7} but that for FRRC is 6.3 x10^{-7}. Usually one expects "freezing" a component of variability to decrease the p-value, but when the p-value is already so small, rounding and other sources of variability can distort this expectation (recall that quantities used in the analysis, e.g., the covariance matrix, are *estimates*, each subject to a sampling error). The reason for such low p-values was described in §9.11. As usual, plotting is always useful to gain insights into the data. Fig. 10.2 shows empirical ROC curves, averaged over

readers, for the four treatments. The differences are visually obvious. With such large differences, between modalities 3 or 4 and 1 or 2, a very small p-value is expected. The plots also show that modalities 3 and 4 are not that different and the same applies to modalities 1 and 2.

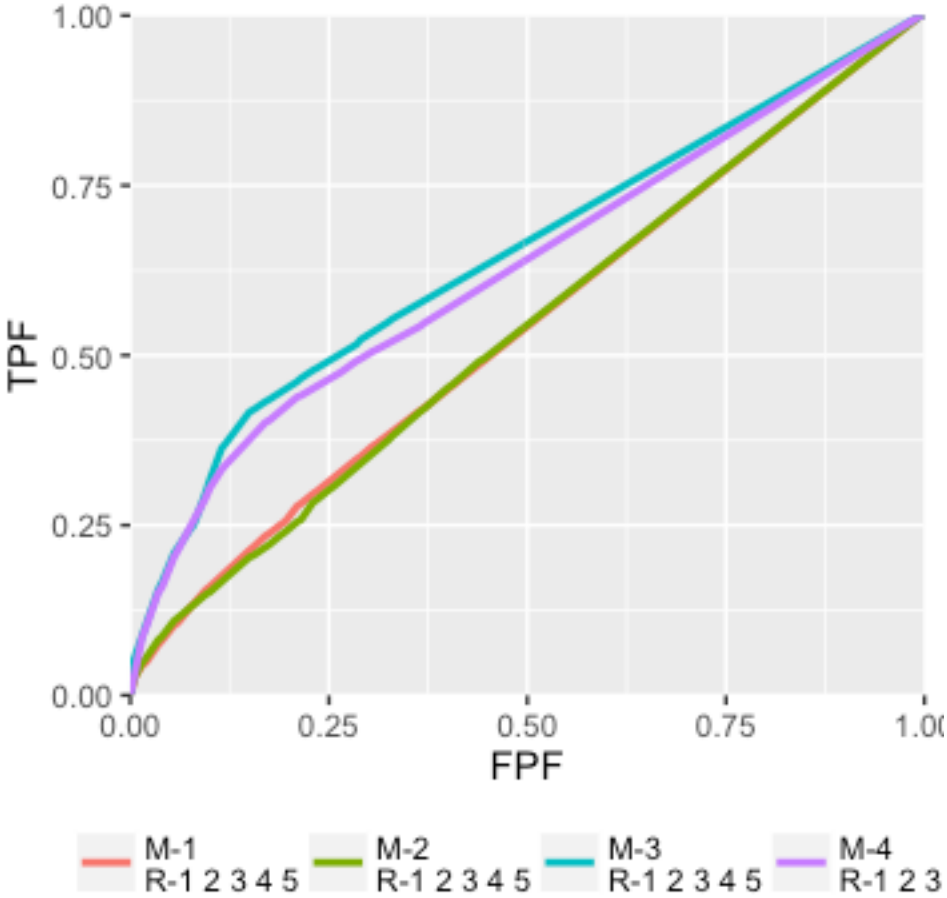

Fig. 10.2: Empirical ROC plots averaged over five radiologists for the four treatments. Since the nodules were invisible on chest x-rays, performance for both M-1 (CXR) and M-2 (CXR+DE) are close to chance level. But tomosynthesis resulted in a highly significant improvement, but tomosynthesis plus dual energy yielded a slightly lower ROC compared to M-3. Differences between modalities M-3 or M-4 and M-1 or M-2 were significant. [M-1 = CXR = 2view chest x-rays, M-2 = DE = dual energy + CXR; M-3 = TOMO = GE VolumeRad tomosynthesis and M-4 = TOMO + DE. The file **`MainOrh.R`** were used to generate this figure.]

The following code snippet lists the reader-averaged empirical AUCs for the four treatments (highlight **`trtMeans`** and click **Run**).

10.5.2: Code snippet

```
> trtMeans
[1] 0.5382801 0.5363208 0.6453556 0.6263425
```

Since this analysis was restricted to lesions not visible on CXR (the full dataset contained lesions of different sizes 3 – 20 mm, all visible on CT images to independent expert radiologist "truthers"[c]), for treatment M1, performance is close to chance-level (empirical AUC = 0.538). Adding dual-energy (M2) had minimal effect (empirical AUC = 0.536, an insignificant decrease). The corresponding confidence intervals for the difference M1 – M2, listed in §10.5.1, include zero for all three analyses. The ROC plot for chest tomosynthesis is clearly above the chance diagonal (empirical AUC = 0.645), and the confidence intervals for the differences M1-M3 and M2-M3 do not include zero (they are both on the negative side of zero, because a larger FOM is being subtracted from a smaller one). Adding dual energy to chest tomosynthesis had minimal effect (empirical AUC = 0.626), so differences M1-M4 and M2-M4 are significant but difference M3-M4 is not (the reader should confirm these statements with the listed confidence intervals). Fig. 10.2 was generated by the last few lines in **mainORH.R**.

10.6: Comparison of ORH and DBMH methods

Source the code in **MainOrhDbmh.R**. This code is brief since practically everything occurs inside **RJafroc**.

10.6.1: Code Listing

```
rm(list = ls()) #mainOrhDbmh.R
library(RJafroc)
ROC <- FALSE
if (ROC) {
  #fileName <- "Franken1.lrc"
  fileName <- "VanDyke.lrc"
  rocData <- ReadDataFile(fileName, format = "MRMC", renumber = "TRUE")
} else {
  fileName <- "CXRinvisible3-20mm.xlsx"
  frocData <- ReadDataFile(fileName, format = "JAFROC", renumber = "TRUE")
  rocData <- FROC2HrROC(frocData)
  rm(frocData)
}

OutputReport(dataset = rocData,fom = "Wilcoxon",method = "DBMH",
             reportFormat = "xlsx",reportFile = "DBMH.xlsx", showWarnings = FALSE)
OutputReport(dataset = rocData,fom = "Wilcoxon",method = "ORH",
             reportFormat = "xlsx",reportFile = "ORH.xlsx", showWarnings = FALSE)
```

After reading the relevant data file, currently set to **CXRinvisible3-20mm.xlsx**, and converting it to a highest rating ROC dataset object named **rocData**, the code uses the **RJafroc** package function **OutputReport()** to analyze the dataset and generate a report using in the first call the DBMH method and in the second call the ORH method. **Source** this file to get the following:

10.6.2: Code Output

```
> source('~/book2/B Statistics of ROC analysis/B3 AnalyzingROCdataORH/software/MainOrhDbmh.R')
...
The report has been saved to DBMH.xlsx.
```

[c] This is the phrase used by all collaborators on the project; some may cringe but this is how new words are ultimately adopted into the English Language.

```
The report has been saved to ORH.xlsx.
```

The output is in two Excel files: **DBMH.xlsx** and **ORH.xlsx**. The reader should compare these files and be convinced that, except for minor differences in fixed-reader p-values, due to the differences in degrees of freedom, they are identical. The reader should also confirm that information in **ORH.xlsx** agrees with that in §10.5.1, where the details of the analysis are not "hidden" inside **RJafroc**.

10.7: Single-treatment multiple-reader Analysis

Often one has data in a single treatment and multiple readers are involved. One wishes to determine if the performance of the readers as a group equal some specified value.

In §10.2 *single-reader multiple-treatment* analysis was described. Attention now turns to *single-treatment multiple-reader* analysis and they are not the same! After all, treatment is a fixed factor while reader is a random factor; so one cannot simply use the previous analysis with reader and treatment interchanged (my 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 **SignificanceTestingSingleFixedFactor()** 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[2,3,7] has described the appropriate modification of the OR model when all readers read cases in a single treatment. 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.

For single-treatment multiple-reader ORH analysis, the figure of merit model is (contrast the following equation to Eqn. (10.1) 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):

$$\theta_{j\{c\}} = \mu + R_j + \varepsilon_{j\{c\}} \quad . \qquad \textbf{(10.62)}$$

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

$$\theta_{\bullet} = \mu \quad . \qquad \textbf{(10.63)}$$

The variance of the reader-averaged FOM can be shown[6] to be given by (the reference is to the original OR publication, specifically Eqn. 2.3):

$$\sigma_{\theta_{\cdot}}^2 = \frac{1}{J}\left(\sigma_R^2 + Var + \left(J-1\right)Cov_2\right) \quad . \qquad \textbf{(10.64)}$$

Connection to existing literature: Rather than attempt to derive the preceding equation, it is shown how it follows from the existing literature[6]. For convenience Eqn. 2.3 *ibid* is reproduced below.

$$Var\left(\hat{\theta}_{i\bullet\bullet}\right) = \left(1/J\right)\left[\sigma_b^2 + \sigma_{ab}^2 + \left(\sigma_w^2/K\right) + \sigma_c^2\left(1+\left[J-1\right]r_2\right)\right] \quad . \qquad \textbf{(10.65)}$$

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_b^2 in the OR notation is what we are calling σ_R^2; for single treatment the treatment-reader interaction term σ_{ab}^2 is absent; and for single "replication" the term σ_w^2/K (in OR notation K is the number of replications) is absent, or, more accurately, the within-reader variance σ_w^2 is absorbed into the case sampling variance σ_c^2 as the two are inseparable); the term σ_c^2 is what we are calling Var; and $\sigma_c^2 r_2$ in OR paper is what we are calling Cov_2. ■

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

One needs to replace σ_R^2 in Eqn. (10.64) with an expected value. Again, rather than attempt to derive the following equation, it is shown how it follows from the existing literature[7]. We start with Table I ibid: this is a table of expected means squares for the OR model, analogous to 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:

$$E\left(MSR\right) = \boldsymbol{\sigma}_R^2 + Var - Cov_2 \quad . \qquad \textbf{(10.66)}$$

Substituting Eqn. (10.66) in Eqn. (10.64) yields,

$$\sigma_{\theta_\bullet}^2 = \frac{1}{J}\left(E\left(MSR\right) + JCov_2\right). \qquad \textbf{(10.67)}$$

An estimate of *MSR* is given by (from here on it is understood that MSR is an *estimate*, i.e., the circumflex notation is suppressed; the same is true for Cov_2):

$$MSR \equiv \widehat{MSR} = \frac{1}{J-1}\sum_{j=1}^{J}\left(\theta_j - \theta_\bullet\right)^2 \qquad \textbf{(10.68)}$$

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

$$\sigma_{\theta_\bullet}^2 = \frac{1}{J}\left(MSR + J\max\left(Cov_2, 0\right)\right). \qquad \textbf{(10.69)}$$

The observed value of the t-statistic (the subscript emphasizes that this statistic applies to the single treatment analysis) for testing the NH is:

$$t_{I=1} \equiv \frac{\mu - \mu_0}{\sigma_{\theta_\bullet}} = \left(\theta_\bullet - \mu_0\right)\sqrt{\frac{J}{\left(MSR + J\max\left(Cov_2, 0\right)\right)}} \qquad \textbf{(10.70)}$$

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

$$t_{I=1} = \frac{\mu - \mu_0}{\sigma_{\theta_\bullet}} \sim t_{df_H^{I=1}}. \qquad \textbf{(10.71)}$$

In the above equation, Hillis single-treatment degree of freedom $df_H^{I=1}$ is defined by[7]:

$$df_H^{I=1} = \left[\frac{MSR + J\max\left(J\widehat{Cov_2}, 0\right)}{MSR}\right]^2\left(J-1\right) \qquad \textbf{(10.72)}$$

---

[d] Since one is dealing with estimates of a random variable, it is possible for the estimate of Cov_2 to be negative.

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:

$$p = \mathrm{P}\left(t > \left|t_{I=1}\right| \mid t \sim t_{df_H^{I=1}}\right) \quad . \qquad \textbf{(10.73)}$$

Therefore, a $100(1-\alpha)\%$ confidence interval for $\theta_{\cdot} - \mu_0$ is:

$$\left(\theta_{\cdot} - \mu_0\right) \pm t_{\alpha/2;df_H^{I=1}} \sigma_{\theta_{\cdot}} = \left(\theta_{\cdot} - \mu_0\right) \pm t_{\alpha/2;df_H^{I=1}} \sqrt{\frac{1}{J}\left(MSR + \max\left(JCov_2, 0\right)\right)} \quad . \qquad \textbf{(10.74)}$$

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.

10.1 Code output

```
> source('~/Desktop/book3/03 B Statistics of ROC analysis/B10 ORH
Analysis/software/mainSingleTreatment.R')
data file =  CXRinvisible3-20mm.xlsx
The NH is that thetaDot = mu0, where thetaDot=  0.5383 and mu0 =  0.5834
The mean FOM for the anal2zed treatment is: 0.5383
The 95 % CI for the preceding value is: ( 0.4931 , 0.5834 )
The t-statistic and p-value to test H0: (analyzed treatment = standard) are: -2.166 , and  0.05
The difference in reader averaged analyzed treatment minus standard =  -0.04514
The 95 % CI of the preceding value is ( -0.09028 , 8.362e-07 )
```

For this dataset the NH that reader-averaged AUC in treatment 1 equals 0.583422 (the latter value was deliberately chosen to demonstrate that when the p-value is 0.05 the 95% confidence interval "touches" zero) is (just) not rejected and the p-value is (just) greater than 0.05. Change the comparison value at line 24 to 0.6 and **`Source`** the code. Now the NH is rejected with p-value = 0.0113 and the 95% confidence interval for the difference FOM is (-0.107, -0.017).

An application of this method to comparing the performance of a group of radiologists to computer-aided diagnosis (CAD) on the same set of images is presented in **Chapter 22**.

10.8: 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 Eqn. (10.26) 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 paper[6]. 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 by Dorfman Berbaum and Metz[1] and the subsequent one by Obuchowski and Rockette[6] 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 [17-20]. 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, and for which software is readily available.

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

10.9: References

1. Dorfman DD, Berbaum KS, Metz CE. ROC characteristic rating analysis: Generalization to the Population of Readers and Patients with the Jackknife method. *Invest Radiol.* 1992;27(9):723-731.
2. Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data. *Statistics in Medicine.* 2005;24(10):1579-1607.
3. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC studies. *Statistics in Medicine.* 2007;26:596-619.
4. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. *Acad Radiol.* 2008;15(5):647-661.

5. Hajian-Tilaki KO, Hanley JA, Joseph L, Collet JP. Extension of receiver operating characteristic analysis to data concerning multiple signal detection tasks. *Acad Radiol.* 1997;4:222-229.
6. Obuchowski NA, Rockette HE. Hypothesis Testing of the Diagnostic Accuracy for Multiple Diagnostic Tests: An ANOVA Approach with Dependent Observations. *Communications in Statistics: Simulation and Computation.* 1995;24:285-308.
7. Hillis SL. A marginal‐mean ANOVA approach for analyzing multireader multicase radiological imaging data. *Statistics in medicine.* 2014;33(2):330-360.
8. Larsen RJ, Marx ML. *An Introduction to Mathematical Statistics and Its Applications.* 3rd ed. Upper Saddle River, NJ: Prentice-Hall Inc; 2001.
9. Strang G. *Linear Algebra and its Applications.* 4 ed. Stamford, CT: Cengage Learning; 2005.
10. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap.* Vol 57. Boca Raton: Chapman & Hall/CRC; 1993.
11. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics.* 1988;44:837-845.
12. Dobbins JT, McAdams HP, Sabol JM, et al. Multi-Institutional Evaluation of Digital Tomosynthesis, Dual-Energy Radiography, and Conventional Chest Radiography for the Detection and Management of Pulmonary Nodules. Radiology. *Radiology.* 2016;000(000):(in press).
13. Obuchowski NA. Sample size calculations in studies of test accuracy. *Statistical Methods in Medical Research.* 1998;7(4):371-392.
14. Obuchowski NA. Sample Size Tables For Receiver Operating Characteristic Studies. *Am J Roentgenol.* 2000;175(3):603-608.
15. Roe CA, Metz CE. Variance-Component Modeling in the Analysis of Receiver Operating Characteristic Index Estimates. *Acad Radiol.* 1997;4(8):587-600.
16. Chakraborty DP, Zhai X. RJafroc 1.0.0: Modeling, Analysis, Validation and Visualization of Observer Performance Studies in Diagnostic Radiology. 2018; https://cran.r-project.org/web/packages/RJafroc/.
17. Clarkson E, Kupinski MA, Barrett HH. A Probabilistic Model for the MRMC Method, Part 1: Theoretical Development. *Academic Radiology.* 2006;13(11):1410-1421.
18. Kupinski MA, Clarkson E, Barrett HH. A Probabilistic Model for the MRMC Method, Part 2: Validation and Applications. *Academic Radiology.* 2006;13(11):1422-1430.
19. Gallas BD. One-Shot Estimate of MRMC Variance: AUC. *Academic Radiology.* 2006;13(3):353-362.
20. Gallas BD, Bandos A, Samuelson FW, Wagner RF. A Framework for Random-Effects ROC Analysis: Biases with the Bootstrap and Other Variance Estimators. *Communications in Statistics - Theory and Methods.* 2009;38(15):2586 - 2603.