The RJafroc Book

Dev P. Chakraborty, PhD

2020-08-22

Contents

Preface

- This book is currently (as of August 2020) in preparation.
- It is intended as an online update to my "physical" book (Chakraborty, 2017). Since its publication in 2017 the `RJafroc` package, on which the `R` code examples in the book depend, has evolved considerably, causing many of the examples to "break". This also gives me the opportunity to improve on the book and include additional material.

A note on the online distribution mechanism of the book

- In the hard-copy version of my book (Chakraborty, 2017) the online distribution mechanism was `BitBucket`.
- `BitBucket` allows code sharing within a *closed* group of a few users (e.g., myself and a grad 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`. All previous instructions pertaining to `BitBucket` are obsolete.
- In order to access `GitHub` material one needs to create a (free) `GitHub` account.
- Go to this link and click on `Sign Up`.

Contributing to this book

- I appreciate constructive 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`.
- When done this way, contributions from users automatically become part of the `GitHub` documentation/history of the book.

Is this book relevant to you and what are the alternatives?

- Diagnostic imaging system evaluation
- Detection
- Detection combined with localization
- Detection combined with localization and classificatin
- AI
- CV
- Alternatives

Chapter 1

Hypothesis Testing

1.1 Introduction

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

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

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

The two competing options are termed the null hypothesis (NH) and the alternative hypothesis (AH). The hypothesis testing procedure is analogous1 to the jury trial system in the US, with 20 instead of 12 jurors, with the null hypothesis being the presumption of innocence and the alternative hypothesis being the defendant is guilty, and the decision rule is to assume the defendant is innocent

unless all 20 jurors agree the defendant is guilty. If even one juror disagrees, the defendant is deemed innocent (equivalent to choosing an α – defined below - of 0.05, or 1/20).

1.2 Hypothesis testing for a single-modality single-reader ROC study

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

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

The following code chunk encodes the `Wilcoxon` function:

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

In the next code chunk we set $\mu = 1.5$ and $\sigma = 1.3$ and simulate $K_1 = 50$ non-diseased cases and $K_2 = 52$ diseased cases. For clarity I like to keep the sizes of the two arrays slightly different; this allows one to quickly check, with a glance at the `Environment` panel, that array dimensions are as expected. The `for` loop draws 50 samples from the $N(0,1)$ distribution and 52 samples from the $N(\mu, \sigma^2)$ distribution, calculates the empirical AUC using the Wilcoxon, and the process is repeated 10,000 times, the AUC values are saved to a huge array `AUC`. After exit from the `for-loop` we calculate the mean and standard deviation of the `AUC` values.

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

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

By the simple (if unimaginative) approach of sampling 10,000 times, one estimates the *population* mean and standard deviation of empirical AUC, denoted below by AUC_{pop} and σ_{AUC}, respectively. Based on the 10,000 simulations, $AUC_{pop} = 0.819178$ and σ_{AUC} =0.0417668.

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

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

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

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

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

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

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

$$Z = \frac{AUC - AUC_{pop}}{\sigma_{AUC}} \sim N(0,1)$$

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

The ubiquitous p-value is the probability that the observed magnitude of z, or larger, occurs under the null hypothesis (NH), that the true mean of Z is zero.

The p-value corresponding to an observed z of 1.0418397 is given by (as always, uppercase Z is the random variable, while lower case z is a realized value):

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

To recapitulate statistical notation, $\Pr(|Z| \geq |z| \mid Z \sim N(0,1))$ is to be parsed as $\Pr(A \mid B)$, that is, the probability $|Z| \geq |z|$ given that $Z \sim N(0,1)$. The last line in Eqn. (8.4) follows from the symmetry of the unit normal distribution, i.e., the area above 1.042 must equal the area below -1.042.

Since z is a continuous variable, the probability that a sampled value will exactly equal the observed value is zero. Therefore, one must pose the question as stated above, namely what is the probability that Z is at least as extreme as the observed value (by "extreme" I mean further from zero, in either positive or negative directions). If the observed was $z = 2.5$ then the corresponding p-value would be $2\Phi(-2.5)$=0.01242, which is smaller than 0.2975 (`2*pnorm(-2.5)` = 0.01241933). This is cited below as the "second example".

Under the zero-mean null hypothesis, the larger the magnitude of the observed value of Z, the smaller the p-value, and the more unlikely that the data supports the NH. **The p-value can be interpreted as the degree of unlikelihood that the data supports the NH.**

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

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

```
p2tailed <- pnorm(-abs(z)) + (1-pnorm(abs(z))) # p value for two-sided AH
p1tailedGT <- 1-pnorm(z) # p value for one-sided AH > 0
p1tailedLT <- pnorm(z)# p value for one-sided AH < 0
alpha  <- 0.05
```

In the first example, with observed p-value equal to 0.2975, one would not reject the null hypothesis, but in the second example, with observed p-value equal to 0.01242, one would. If the p-value is exactly 0.05 (unlikely with ROC analysis, but one needs to account for it) then one does not reject the NH. In the 20-juror analogy, of one juror insists the defendant is not guilty, then observed Pr is 0.05, and one does not reject the NH that the defendant is innocent (the double negatives, very common in statistics, can be confusing; in plain English, the defendant goes home).

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

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

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

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

1.3 Type-I errors

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

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

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

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

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

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

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

Upon exit from the for-loop, one calculates the observed Type-I error rate,

denoted `ObsvdTypeIErrRate` by summing the reject array and dividing by 2000. One calculates a 95% confidence interval for `ObsvdTypeIErrRate` based on the binomial distribution, as in Chapter 03.

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

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

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

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

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

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

If f is close to 0.05, then for 2000 trials, the 0.95 or 95% CI for f is $f \pm 0.01$, i.e., `qnorm(alpha/2) * sqrt(.05*(.95)/2000)` $= 0.009551683 \sim 0.01$.

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

1.4 One sided vs. two sided tests

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

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

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

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

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

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

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

1.5 Statistical power

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

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

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

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

A little history is due at this point. The author's first FROC study, which led to his entry into this field (Chakraborty et al., 1986), was published in Radiology in 1986 after a lot of help from a reviewer, who we (correctly) guessed was the late Prof. Charles E. Metz. Prof. Gary T. Barnes (the author's mentor at that time at the University of Alabama at Birmingham) and the author visited Prof. Charles Metz in Chicago for a day ca. 1986, to figuratively "pick Charlie's brain". Prof. Metz referred to the concept outlined in the previous paragraph, as an *ROC within an ROC*.

This curve does not summarize the result of a single ROC study. Rather it summarizes the probabilistic behavior of the two types of errors that occur when one conducts thousands of such studies, under both NH true and NH

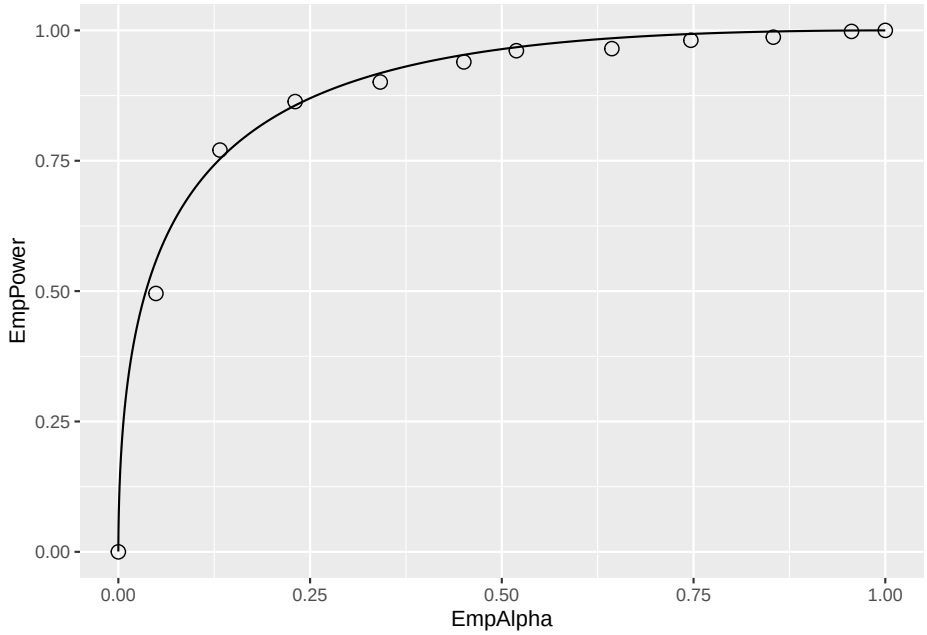
false conditions, each time with different values of α, with each trial ending in a decision to reject or not reject the null hypothesis. The long sentence is best explained with an example.

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

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

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

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



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

The action begins in line 20, which begins a for-loop in `a`, an index into `alphaArr`. Line 21 selects the appropriate value for `alpha` (0.05 on the first pass, 0.15 on the next pass, etc.). Line 22 initializes `reject[2,2000]` with zeroes, to hold the result of each trial; the first index corresponds to hypothesis `h` and the second to trial `t`. Line 23 begins a for-loop in `h`, with `h` = 1 corresponding to the NH and `h` = 2 to the AH. Line 24 begins a for-loop in `t`, the trial index. The code within this block is similar to previous examples. It simulates ratings, computes AUC, calculates the p-value, and saves a rejection of the NH as a one at the appropriate array location `reject[h,t]`. Lines

32 – 33 calculate the empirical α and empirical power for each value of α in `alphaArr`. After padding the ends with zero and ones (the trivial points), the remaining lines plot the "ROC within an ROC".

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

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

1.5.1 Factors affecting statistical power

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

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

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

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

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

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

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

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

The ES for the code above is 0.08 (in AUC units). It should be obvious that if effect size is zero, then power equals α. This is because then there is no distinction between the null and alternative hypotheses conditions. Conversely, as effect size increases, statistical power increases, the limiting value being unity, when every trial results in rejection of the null hypothesis. The reader should experiment with different values of `muAH` to be convinced of the truth of these statements.

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

```
#> pop NH mean AUC =  0.8594882 , pop NH sigma AUC =  0.02568252
```

```
#> num. non-diseased images =  100 num. diseased images =  104
#> obsvdPower =  0.313
#> 95% confidence interval =  0.2926772 0.3333228
#> Effect Size =  0.08000617 0
```

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

- Alpha: Statistical power depends on *alpha*. return the sample size to the original values . The results below are for two runs of the code, the first with the original value , set at line 16, the second with :

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

Decreasing α results in decreased statistical power.

1.6 Comments

The Wilcoxon statistic was used to estimate the area under the ROC curve. One could have used the binormal model, introduced in Chapter 06, to obtain maximum likelihood estimates of the area under the binormal model fitted ROC curve. The reasons for choosing the simpler empirical area are as follows. (1) With continuous ratings and 102 operating points, the area under the empirical ROC curve is expected to be a close approximation to the fitted area. (2) With maximum likelihood estimation, the code would be more complex – in addition to the fitting routine one would require a binning routine and that would introduce yet another variable in the analysis, namely the number of bins and how the bin boundaries were chosen. (3) The maximum likelihood fitting code can sometimes fail to converge, while the Wilcoxon method is always guaranteed to yield a result. The non-convergence issue is overcome by modern methods of curve fitting described in later chapters. (4) The aim was to provide an understanding of null hypothesis testing and statistical power without being bogged down in the details of curve fitting.

1.7 Why alpha is chosen to be 5%

One might ask why α is traditionally chosen to be 5%. It is not a magical number, rather a cost benefit tradeoff. Choosing too small a value of α would

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

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

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

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

1.8 Discussion

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

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

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

As noted in the introduction, comparing a single reader's performance to a specified value is not a clinically interesting problem. The next two chapters describe methods for significance testing of multiple-reader multiple-case (MRMC) ROC datasets, consisting of interpretations by a group of readers of a common set of cases in typically two modalities. It turns out that the analyses yield variability estimates that permit sample size calculation. After all, sample size calculation is all about estimation of variability, the denominator of the z-statistic, i.e., Eqn. (8.3), in the context of this chapter. The formulae will look more complex, as interest is not in determining the standard deviation of AUC, but in the standard deviation of the inter-modality reader-averaged AUC difference. However, the basic concepts remain the same.

1.9 References

Chapter 2

Background on Dorfman Berbaum Metz (DBM) Analysis

2.1 Introduction

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

This chapter describes the first truly successful method of analyzing MRMC ROC data, namely the Dorfman-Berbaum-Metz (DBM) method (Dorfman et al., 1992). The other method, due to Obuchowski and Rockette (Obuchowski and Rockette, 1995), is the subject of Chapter 10. Both methods have been

substantially improved by Hillis (Hillis et al., 2008; Hillis, 2007, 2014). Hence the title of this chapter: "Dorfman Berbaum Metz Hillis (DBM) Analysis". It is not an overstatement that ROC analysis came of age with the methods described in this chapter. Prior to the techniques described here, one knew of the existence of sources of variability affecting a measured AUC value, as discussed in (book) Chapter 07, but then-known techniques (Swets and Pickett, 1982) for estimating the corresponding variances and correlations were impractical.

2.1.1 Historical background

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

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

based on few values, the estimate was inaccurate, and the already case-starved original dataset made it difficult to estimate AUCs for the subsets; moreover, the division into subsets was at the discretion of the researcher, and therefore unlikely to be reproduced by others. Estimating within-reader variability required re-reading the entire case set, or at least a part of it. ROC studies have earned a deserved reputation for taking much time to complete, and having to re-read a case set was not a viable option. [Historical note: the author recalls a barroom conversation with Dr. Thomas Mertelmeir after the conclusion of an SPIE meeting ca. 2004, where Dr. Mertelmeir commiserated mightily, over several beers, about the impracticality of some of the ROC studies required of imaging device manufacturers by the FDA.]

2.1.2 The Wagner analogy

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

2.1.3 The shortage of numbers to analyze and a pivotal breakthrough

The basic issue was that the calculation of AUC reduces the relatively large number of ratings of a set of non-diseased and diseased cases to a single number. For example, after completion of an ROC study with 5 readers and 100 non-diseased and 100 diseased cases interpreted in two treatments, the data is reduced to just 10 numbers, i.e., five readers times two treatments. It is difficult to perform statistics with so few numbers. The author recalls a conversation with Prof. Kevin Berbaum at a Medical Image Perception Society meeting in Tucson, Arizona, ca. 1997, in which he described the basic idea that forms the subject of this chapter. Namely, using the jackknife pseudovalues (to be defined

below) as individual case-level figures of merit. This, of course, greatly increases the amount of data that one can work with; instead of just 10 numbers one now has 2,000 pseudovalues (2 x 5 x 200). If one assumes the pseudovalues behave essentially as case-level data, then by assumption they are independent and identically distributed, and therefore they satisfy the conditions for application of standard analysis of variance (ANOVA) techniques. [This assumption has been much criticized and is the basis for some preferring alternate approaches - but, as Hillis has stated, and I paraphrase, the pseudovalue based method "works", but lacks sufficient rigor.] The relevant paper had already been published in 1992 but other distractions and lack of formal statistical training kept the author from fully appreciating this work until later.

Although methods are available for more complex study designs including partially paired data (Metz et al., 1998; Obuchowski, 2009), I will restrict to fully paired data (i.e., each case is interpreted by all readers in all treatments). There is a long history of how this field has evolved and the author cannot do justice to all methods that are currently available. Some of the methods (Toledano, 2003; Ishwaran and Gatsonis, 2000; Toledano and Gatsonis, 1996) have the advantage that they can handle explanatory variables (termed covariates) that could influence performance, e.g., years of experience, types of cases, etc. Other methods are restricted to specific choices of FOM. Specifically, the probabilistic approach (Clarkson et al., 2006; Kupinski et al., 2006; Gallas et al., 2007; Gallas, 2006) is restricted to the empirical AUC under the ROC curve, and therefore are not applicable to other FOMs, e.g., parametrically fitted ROC AUCs or, more importantly, to location specific paradigm FOMs. Instead, the author will focus on methods for which software is readily available (i.e., freely on websites), which have been widely used (the method that the author is about to describe has been used in several hundred publications) and validated via simulations, and which apply to any scalar figure of merit, and therefore widely applicable, even to location specific paradigms.

2.1.4 Organization of the chapter

The organization of the chapter is as follows. The concepts of reader and case populations, introduced in (book) Chapter 07, are recapitulated. A distinction is made between *fixed* and *random* factors – statistical terms with which one must become familiar. Described next are three types of analysis that are possible with MRMC data, depending on which factors are regarded as random and which as fixed. The general approach to the analysis is described. Two methods of analysis are possible: the jackknife pseudovalue-based approach detailed in this chapter and an alternative approach is detailed in Chapter 10. The Dorfman-Berbaum-Metz (DBM) model for the jackknife pseudovalues is described that incorporates different sources of variability and correlations possible with MRMC data. Calculation of ANOVA-related quantities, termed mean squares, from the pseudovalues, are described followed by the significance test-

ing procedure for testing the null hypothesis of no treatment effect. A relevant distribution used in the analysis, namely the F-distribution, is illustrated with R examples. The decision rule, i.e., whether to reject the NH, calculation of the ubiquitous p-value, confidence intervals and how to handle multiple treatments is illustrated with two datasets, one an older ROC dataset that has been widely used to demonstrate advances in ROC analysis, and the other a recent dataset involving evaluation of digital chest tomosynthesis vs. conventional chest imaging. The approach to validation of DBM analysis is illustrated with an R example. The chapter concludes with a section on the meaning of the pseudovalues. The intent is to explain, at an intuitive level, why the DBM method "works", even though use of pseudovalues has been questioned3 at the conceptual level. For organizational reasons and space limitations, details of the software are relegated to Online Appendices, but they are essential reading, preferably in front of a computer running the online software that is part of this book. The author has included material here that may be obvious to statisticians, e.g., an explanation of the Satterthwaite approximation, but are expected to be helpful to others from non-statistical backgrounds.

2.2 Random and fixed factors

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

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

This brings out an essential difference between how cases, readers and treatments have to be regarded in the variability estimation procedure. Cases and readers are usually regarded as random factors (there has to be at least one random factor – if not, there are no sources of variability and nothing to apply statistics to!), while treatments are regarded as fixed factors. The random factors contribute stochastic (i.e., random) variability, but the fixed factors do not, rather they contribute constant shifts in performance. The terms fixed and random factors are used in this specific sense, and are derived, in turn, from ANOVA methods in statistics10,25. With two or more treatments, there are shifts in performance of treatments relative to each other, that one seeks to assess the significance of against a background of noise contributed by the random factors. If the shifts are sufficiently large compared to the noise, then one can state, with some certainty, that they are real. Quantifying the last statement uses the methods of hypothesis testing introduced in Chapter 1 or Chapter [Hypothesis Testing].

2.3 Reader and case populations and data correlations

As discussed in (book) §7.2, conceptually there is a reader-population, generally modeled as a normal distribution $\theta_j \sim N\left(\theta_{\bullet\{1\}}, \sigma^2_{br+wr}\right)$, describing the variation of skill-level of readers. The notation closely follows that in the cited section, the only change being that the binormal model estimate A_z has been replaced by a generic FOM, denoted θ. Each reader j is characterized by a different value of θ_j, $j = 1, 2, ...J$ and one can conceptually think of a bell-shaped curve with variance σ^2_{br+wr} describing between-reader variability of the readers. A large variance implies large spread in reader skill levels.

Likewise, there is a case-population, also modeled as a normal distribution, describing the variations in difficulty levels of the patients. One actually has two unit-variance distributions, one per diseased state, characterized by a separation parameter and conceptually an easy case set has a larger than usual separation parameter while a difficult case set has a smaller than usual separation parameter. The distribution of the separation parameter can be modeled as a bell-shaped curve $\theta_{\{c\}} \sim N\left(\theta_{\{\bullet\}}, \sigma^2_{cs+wr}\right)$ with variance σ^2_{cs+wr} describing the variations in difficulty levels of different case samples. Note the need for the case-set index, introduced in Chapter 07, to specify the separation parameter for a specific case-set (in principle a j-index is also needed as one cannot have an interpretation without a reader; for now it is suppressed; one can think of the stated equation as applying to the average reader). A small variance σ^2_{cs} implies the different case sets have similar difficulty levels while a larger variance would imply a larger spread in difficulty levels.

Anytime one has a common random component to two measurements, the measurements are correlated. In the Wagner analogy, the common component is

the random height, as a function of time, of a wave, which contributes the same amount to both height measurements (since the boats are adjacent). Since the readers interpret a common case set in all treatments one needs to account for various types of correlations that are potentially present. These occur due to the various types of pairings that can occur with MRMC data, where each pairing implies the presence of a common component to the measurements: (a) the same reader interpreting the same cases in different treatments, (b) different readers interpreting the same cases in the same treatment and (c) different readers interpreting the same cases in different treatments. These pairings are more clearly elucidated in (book) Chapter 10. The current chapter uses jackknife pseudovalue based analysis to model the variances and the correlations. Hillis has shown that the two approaches are essentially equivalent (Hillis et al., 2008).

2.4 Three types of analyses

MRMC analysis attempts to draw conclusions regarding the significances of inter-treatment shifts in performance. Ideally a conclusion (i.e., a difference is significant: yes/no; the "yes" applies if the p-value is less than alpha) should generalize to the respective populations from which the random samples were obtained. In other words, the idea is to generalize from the observed samples to the underlying populations. Three types of analyses are possible depending on which factor(s) one regards as random and which as fixed: random-reader random-case (RRRC), fixed-reader random-case (FRRC) and random-reader fixed-case (RRFC). If a factor is regarded as random, then the conclusion of the study applies to the population from which the levels of the factor were sampled. If a factor is regarded as fixed, then the conclusion applies only to the specific levels of the sampled factor. For example, if reader is regarded as a random factor, the conclusion generalizes to the reader population from which the readers used in the study were obtained. If reader is regarded as a fixed factor, then the conclusion applies to the specific readers that participated in the study. Regarding a factor as fixed effectively "freezes out" the sampling variability of the population and interest then centers only on the specific levels of the factor used in the study. For fixed reader analysis, conclusions about the significances of differences between pairs of readers are allowed; these are not allowed if reader is treated as a random factor. Likewise, treating case as a fixed factor means the conclusion of the study is specific to the case-set used in the study.

2.5 General approach

This section provides an overview of the steps involved in analysis of MRMC data. Two approaches are described in parallel: a figure of merit (FOM) derived

jackknife pseudovalue based approach, detailed in this chapter and an FOM based approach, detailed in the next chapter. The analysis proceeds as follows:

1. A FOM is selected: *the selection of FOM is the single-most critical aspect of analyzing an observer performance study.* The selected FOM is denoted θ. To keep the notation reasonably compact the usual circumflex "hat" symbol used previously to denote an estimate is suppressed. The FOM has to be an objective scalar measure of performance with larger values characterizing better performance. [The qualifier "larger" is trivially satisfied; if the figure of merit has the opposite characteristic, a sign change is all that is needed to bring it back to compliance with this requirement.] Examples are empirical AUC, the binormal model-based estimate A_z , other advance method based estimates of AUC, sensitivity at a predefined value of specificity, etc. An example of a FOM requiring a sign-change is FPF at a specified TPF, where smaller values signify better performance.
2. For each treatment i and reader j the figure of merit θ_{ij} is estimated from the ratings data. Repeating this over all treatments and readers yields a matrix of observed values θ_{ij}. This is averaged over all readers in each treatment yielding $\theta_{i\bullet}$. The observed effect-size ES_{obs} is defined as the difference between the reader-averaged FOMs in the two treatments, i.e., $ES_{obs} = \theta_{2\bullet} - \theta_{1\bullet}$. While extensible to more than two treatments, the explanation is more transparent by restricting to two modalities.
3. If the magnitude of ES_{obs} is "large" one has reason to suspect that there might indeed be a significant difference in AUCs between the two treatments, where significant is used in the sense of (book) Chapter 08. Quantification of this statement, specifically how large is "large", requires the conceptually more complex steps described next.

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

1. Depending on the selected method of modeling the data (pseudovalue vs. FOM) a statistical model is used which includes parameters modeling the true values in each treatment, and expected variations due to different variability components in the model, e.g., between-reader variability, case-sampling variability, interactions (e.g., modeling the possibility that the random effect of a given reader could be treatment dependent) and the presence of correlations (between pseudovalues or FOMs) because of the pairings inherent in the interpretations.
2. In RRRC analysis one accounts for randomness in readers and cases. In FRRC analysis one regards reader as a fixed factor. In RRFC analysis one regards case as a fixed factor. The statistical model depends on the type of analysis.

3. The parameters of the statistical model are estimated from the observed data.
4. The estimates are used to infer the statistical distribution of the observed effect size, ES_{obs}, regarded as a realization of a random variable, under the null hypothesis (NH) that the true effect size is zero.
5. Based on this statistical distribution, and assuming a two-sided test, the probability (this is the oft-quoted p-value) of obtaining an effect size at least as extreme as that actually observed, is calculated, as in Chapter 08.
6. If the p-value is smaller than a preselected value, denoted α, one declares the treatments different at the α - significance level. The quantity α is the control (or cap) on the probability of making a Type I error, defined as rejecting the NH when it is true. It is common to set $\alpha = 0.05$ but depending on the severity of the consequences of a Type I error, as discussed in (book) Chapter 08, one might consider choosing a different value. Notice that α is a pre-selected number while the p-value is a realization of a random variable.
7. For a valid statistical analysis, the empirical probability α_{emp} over many (typically 2000) independent NH datasets, that the p-value is smaller than α, should equal α to within statistical uncertainty.

2.6 Summary TBA

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

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

A third MRMC ROC method, due to Clarkson, Kupinski and Barrett19,20, implemented in open-source JAVA software by Gallas and colleagues22,44 (http:

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

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

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

2.7 References

Chapter 3

Significance Testing using the DBM Method

DBM = Dorfman Berbaum Metz

3.1 The DBM sampling model

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

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

$$Y'_{ijk} = K\theta_{ij} - (K-1)\theta_{ij(k)} \tag{3.1}$$

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

Recall from book Chapter 07 that the jackknife is a way of teasing out the case-dependence: the left hand side of Equation (3.1) has a case index k, with k running from 1 to K, where K is the total number of cases: $K = K_1 + K_2$.

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

$$Y_{ijk} = Y'_{ijk} + \left(\theta_{ij} - Y'_{ij\bullet}\right) \tag{3.2}$$

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

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

$$Y_{ij\bullet} = Y'_{ij\bullet} + \left(\theta_{ij} - Y'_{ij\bullet}\right) = \theta_{ij} \tag{3.3}$$

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

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

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

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

$$Y_{n(ijk)} = \mu + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + (\tau RC)_{ijk} + \epsilon_{n(ijk)} \tag{3.4}$$

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

$$\sum_{i=1}^{I} \tau_i = 0 \Rightarrow \tau_\bullet = 0 \tag{3.5}$$

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

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

The parameter τ_i is estimated as follows:

$$Y_{ijk} \equiv Y_{1(ijk)}\tau_i = Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet} \tag{3.6}$$

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

3.1.1 Explanation of terms in the model

The right hand side of Eqn. (3.1) consists of one fixed and 7 random effects. The current analysis assumes readers and cases as random factors (RRRC), so by definition R_j and C_k are random effects, and moreover, any term that includes a random factor is a random effect; for example, $(\tau R)_{ij}$ is a random effect because it includes the R factor. Here is a list of the random terms:

$$R_j, C_k, (\tau R)_{ij}, (\tau C)_{ik}, (RC)_{jk}, (\tau RC)_{ijk}, \epsilon_{ijk} \tag{3.7}$$

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

$$\left.\begin{aligned} R_j &\sim N\left(0, \sigma_R^2\right) \\ C_k &\sim N\left(0, \sigma_C^2\right) \\ (\tau R)_{ij} &\sim N\left(0, \sigma_{\tau R}^2\right) \\ (\tau C)_{ik} &\sim N\left(0, \sigma_{\tau C}^2\right) \\ (RC)_{jk} &\sim N\left(0, \sigma_{RC}^2\right) \\ (\tau RC)_{ijk} &\sim N\left(0, \sigma_{\tau RC}^2\right) \\ \epsilon_{ijk} &\sim N\left(0, \sigma_\epsilon^2\right) \end{aligned}\right\} \tag{3.8}$$

Equation (3.8) defines the meanings of the variance components appearing in Equation (3.7). One could have placed a Y subscript (or superscript) on each

of the variances, as they describe fluctuations of the pseudovalues, not FOM values. However, this tends to clutter the notation. So here is the convention:

Unless explicitly stated otherwise, all variance symbols in this chapter refer to pseudovalues. Another convention: $(\tau R)_{ij}$ is *not* the product of the treatment and reader factors, rather it is a single factor, namely the treatment-reader factor with IJ levels, subscripted by the index ij and similarly for the other product-like terms in Equation (3.8).

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

The variances defined in (3.8) are collectively termed *variance components*. Specifically, they are jackknife pseudovalue variance components, to be distinguished from figure of merit (FOM) variance components to be introduced in TBA Chapter 10. They are in order: $\sigma_R^2, \sigma_C^2 \sigma_{\tau R}^2, \sigma_{\tau C}^2, \sigma_{RC}^2, \sigma_{\tau RC}^2, \sigma_\epsilon^2$. They have the following meanings.

- The term σ_R^2 is the variance of readers that is independent of treatment or case, which are modeled separately. It is not to be confused with the terms σ_{br+wr}^2 and σ_{cs+wr}^2 used in §9.3, which describe the variability of θ measured under specified conditions. [A jackknife pseudovalue is a weighted difference of FOM like quantities, TBA (3.1). Its meaning will be explored later. For now, *a pseudovalue variance is distinct from a FOM variance*.]
- The term σ_C^2 is the variance of cases that is independent of treatment or reader.
- The term $\sigma_{\tau R}^2$ is the treatment-dependent variance of readers that was excluded in the definition of σ_R^2. If one were to sample readers and treatments for the same case-set, the net variance would be $\sigma_R^2 + \sigma_{\tau R}^2 + \sigma_\epsilon^2$.
- The term $\sigma_{\tau C}^2$ is the treatment-dependent variance of cases that was excluded in the definition of σ_C^2. So, if one were to sample cases and treatments for the same readers, the net variance would be $\sigma_C^2 + \sigma_{\tau C}^2 + \sigma_\epsilon^2$.
- The term σ_{RC}^2 is the treatment-independent variance of readers and cases that were excluded in the definitions of σ_R^2 and σ_C^2. So, if one were to sample readers and cases for the same treatment, the net variance would be $\sigma_R^2 + \sigma_C^2 + \sigma_{RC}^2 + \sigma_\epsilon^2$.
- The term $\sigma_{\tau RC}^2$ is the variance of treatments, readers and cases that were excluded in the definitions of all the preceding terms in TBA (3.1). So, if one were to sample treatments, readers and cases the net variance would be $\sigma_R^2 + \sigma_C^2 + \sigma_{\tau C}^2 + \sigma_{RC}^2 + \sigma_{\tau RC}^2 + \sigma_\epsilon^2$.
- The last term, σ_ϵ^2 describes the variance arising from different replications of the study using the same treatments, readers and cases. Measuring this variance requires repeating the study several (N) times with the same

treatments, readers and cases, and computing the variance of $Y_{n(ijk)}$, where the additional n-index refers to true replications, $n = 1, 2, ..., N$.

$$\sigma_\epsilon^2 = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{k}\frac{1}{N-1}\sum_{n=1}^{N}\left(Y_{n(ijk)} - Y_{\bullet(ijk)}\right)^2 \tag{3.9}$$

The right hand side of TBA (3.1) is the variance of $Y_{n(ijk)}$, for specific ijk, with respect to the replication index n, averaged over all ijk. In practice $N = 1$ (i.e., there are no replications) and this variance cannot be estimated (it would imply dividing by zero). It has the meaning of *reader inconsistency*, usually termed *within-reader* variability. As will be shown later, the presence of this inestimable term does not limit ones ability to perform significance testing on the treatment effect without having to replicate the whole study, as implied in earlier work (Obuchowski and Rockette, 1995).

An equation like TBA (3.1) is termed a *linear model* with the left hand side, the pseudovalue "observations", modeled by a sum of fixed and random terms. Specifically it is a *mixed model*, because the right hand side has both fixed and random effects. Statistical methods have been developed for analysis of such linear models. One estimates the terms on the right hand side of TBA (3.1), it being understood that for the random effects, one estimates the variances of the zero-mean normal distributions, TBA (3.1)Eqn. (9.7), from which the samples are obtained (by assumption).

Estimating the fixed effects is trivial. The term μ is estimated by averaging the left hand side of TBA (3.1)Eqn. (9.4) over all three indices (since $N = 1$): $\mu = Y_{\bullet\bullet\bullet}$

Because of the way the treatment effect is defined, TBA (3.1) Eqn. (9.5), averaging, which involves summing, over the treatment-index i, yields zero, and all of the remaining random terms yield zero upon averaging, because they are individually sampled from zero-mean normal distributions. To estimate the treatment effect one takes the difference $\tau_i = Y_{i\bullet\bullet} - \mu$.

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

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

3.1.3 Definitions of mean-squares

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

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

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

$$\left.\begin{aligned}
\text{MST} &= \frac{JK\sum_{i=1}^{I}(Y_{i\bullet\bullet}-Y_{\bullet\bullet\bullet})^2}{I-1} \\
\text{MSR} &= \frac{IK\sum_{j=1}^{J}(Y_{\bullet j\bullet}-Y_{\bullet\bullet\bullet})^2}{J-1} \\
\text{MS(C)} &= \frac{IJ\sum_{k=1}^{K}(Y_{\bullet\bullet k}-Y_{\bullet\bullet\bullet})^2}{K-1} \\
\text{MSTR} &= \frac{K\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij\bullet}-Y_{i\bullet\bullet}-Y_{\bullet j\bullet}+Y_{\bullet\bullet\bullet})^2}{(I-1)(J-1)} \\
\text{MSTC} &= \frac{J\sum_{i=1}^{I}\sum_{k=1}^{K}(Y_{i\bullet k}-Y_{i\bullet\bullet}-Y_{\bullet\bullet k}+Y_{\bullet\bullet\bullet})^2}{(I-1)(K-1)} \\
\text{MSRC} &= \frac{I\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{\bullet jk}-Y_{\bullet j\bullet}-Y_{\bullet\bullet k}+Y_{\bullet\bullet\bullet})^2}{(J-1)(K-1)} \\
\text{MSTRC} &= \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk}-Y_{ij\bullet}-Y_{i\bullet k}-Y_{\bullet jk}+Y_{i\bullet\bullet}+Y_{\bullet j\bullet}+Y_{\bullet\bullet k}-Y_{\bullet\bullet\bullet})^2}{(I-1)(J-1)K-1)}
\end{aligned}\right\} \quad (3.10)$$

Note the absence of MSE, corresponding to the ϵ term on the right hand side of (3.10). With only one observation per treatment-reader-case combination, MSE cannot be estimated; it effectively gets absorbed into the $MSTRC$ term.

3.2 Expected values of mean squares

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

> experimental units within a level of a random factor. Tables 1 and 2 show the ANOVA tables with expected mean squares for the unrestricted formulation."
>
> — (Dorfman et al., 1995)

The *observed* mean squares defined in Equation (3.10) can be calculated directly from the *observed* pseudovalues. The next step in the analysis is to obtain expressions for their *expected* values in terms of the variances defined in (3.10). Assuming no replications, i.e., $N = 1$, the expected mean squares are as follows, Table Table 3.1; understanding how this table is derived, would lead the author well outside his expertise and the scope of this book; suffice to say that these are *unconstrained* estimates (as summarized in the quotation above) which are different from the *constrained* estimates appearing in the original DBM publication (Dorfman et al., 1992).

Table 3.1: Unconstrained expected values of mean-squares, as in (Dorfman et al., 1995)

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

- In Table 3.1 the following notation is used as a shorthand:

$$\sigma^2_\tau = \frac{1}{I-1}\sum_{i=1}^{I}\left(Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet}\right)^2 \tag{3.11}$$

Since treatment is a fixed effect, the variance symbol σ^2_τ, which is used for notational consistency in Table 3.1, could cause confusion. The right hand side "looks like" a variance, indeed one that could be calculated for just two treatments but, of course, random sampling from a *distribution of treatments* is not the intent of the notation.

3.3 Random-reader random-case (RRRC) analysis

Both readers and cases are regarded as random factors. The expected mean squares in Table Table 3.1 are variance-like quantities; specifically, they are weighted linear combinations of the variances appearing in (3.8). For single factors the column headed "degrees of freedom" (df) is one less than the number of levels of the corresponding factor; estimating a variance requires first estimating the mean, which imposes a constraint, thereby decreasing df by one. For interaction terms, df is the product of the degrees of freedom for the individual factors. As an example, the term $(\tau RC)_{ijk}$ contains three individual factors, and therefore $df = (I-1)(J-1)(K-1)$. The number of degrees of freedom can be thought of as the amount of information available in estimating a mean square. As a special case, with no replications, the ϵ term has zero df as $N - 1 = 0$. With only one observation $Y_{1(ijk)}$ there is no information to estimate the variance corresponding to the ϵ term. To estimate this term one needs to replicate the study several times – each time the same readers interpret the same cases in all treatments – a very boring task for the reader and totally unnecessary from the researcher's point of view.

3.3.1 Calculation of mean squares: an example

- We choose `dataset02` to illustrate calculation of mean squares for pseudovalues. This is referred to in the book as the "VD" dataset (Van Dyke et al., 1993). It consists of 114 cases, 45 of which are diseased, interpreted in two treatments by five radiologists using the ROC paradigm.
- The first line computes the pseudovalues using the `RJafroc` function `UtilPseudoValues()`, and the second line extracts the numbers of treatments, readers and cases. The following lines calculate, using Equation (3.10) the mean-squares. After displaying the results of the calculation, the results are compared to those calculated by the `RJafroc` function `UtilMeanSquares()`.

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

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

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

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

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

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

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

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

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

```
}
msRC <- msRC * I/((J - 1) * (K - 1))

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

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

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

3.3.2 Significance testing

If the NH of no treatment effect is true, i.e., if $\sigma^2_\tau = 0$, then according to Table 3.1 the following holds (the last term in the row labeled T in Table 3.1 drops out):

$$E\left(MST \mid NH\right) = \sigma^2_\epsilon + \sigma^2_{\tau RC} + K\sigma^2_{\tau R} + J\sigma^2_{\tau C} \tag{3.12}$$

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

$$\begin{aligned} &E\left(MSTR\right) + E\left(MSTC\right) - E\left(MSTRC\right) \\ &= (\sigma^2_\epsilon + \sigma^2_{\tau RC} + K\sigma^2_{\tau R}) + (\sigma^2_\epsilon + \sigma^2_{\tau RC} + J\sigma^2_{\tau C}) - (\sigma^2_\epsilon + \sigma^2_{\tau RC}) \\ &= \sigma^2_\epsilon + \sigma^2_{\tau RC} + J\sigma^2_{\tau C} + K\sigma^2_{\tau R} \\ &= E\left(MST \mid NH\right) \end{aligned} \tag{3.13}$$

Therefore, under the NH, the ratio:

$$\frac{E\left(MST \mid NH\right)}{E\left(MSTR\right)+E\left(MSTC\right)-E\left(MSTRC\right)}=1 \tag{3.14}$$

In practice, one does not know the expected values – that would require averaging each of these quantities, regarded as random variables, over their respective distributions. Therefore, one defines the following statistic, denoted F_{DBM}, using the observed values of the mean squares, calculated almost trivially as in the previous example, using their definitions in Equation (3.10):

$$F_{DBM}=\frac{MST}{MSTR+MSTC-MSTRC} \tag{3.15}$$

F_{DBM} is a realization of a random variable. A non-zero treatment effect, i.e., $\sigma_\tau^2>0$, will cause the ratio to be larger than one, because $E\left(MST\right)$ will be larger, see row labeled T in Table 3.1. Therefore values of $F_{DBM}>1$ will tend to reject the NH. Drawing on a theorem from statistics (Larsen and Marx, 2001), under the NH the ratio of two independent mean squares is distributed as a (central) F-statistic with degrees of freedom corresponding to those of the mean squares forming the numerator and denominator of the ratio (Theorem 12.2.5 in "An Introduction to Mathematical Statistics and Its Applications"). To perform hypothesis testing one needs the distribution, under the NH, of the statistic defined by Eqn. (3.15). This is completely analogous to Chapter 08 where knowledge of the distribution of AUC under the NH enabled testing the null hypothesis that the observed value of AUC equals a pre-specified value.

Under the NH the left hand side of by (9.18), i.e., $F_{DBM|NH}$, is distributed according to the F-distribution characterized by two numbers:

- A numerator degrees of freedom (ndf) – determined by the degrees of freedom of the numerator MST of the ratio comprising the F-statistic, i.e., I–1, and
- A denominator degrees of freedom (ddf) - determined by the degrees of freedom of the denominator of the ratio comprising the F-statistic, to be described below.

Summarizing,

$$\left.\begin{array}{r} F_{DBM|NH}\sim F_{\mathrm{ndf,ddf}} \\ \mathrm{ndf}=I-1 \end{array}\right\} \tag{3.16}$$

The next topic is estimating ddf.

3.3.3 The Satterthwaite approximation

The denominator of the F-ratio is $MSTR + MSTC - MSTRC$. This is not a *simple* mean square (I am using terminology in the Satterthwaite papers - he means any mean square defined by equations such as in Equation (3.10)). Rather it is a *linear combination of mean squares* (with coefficients 1, 1 and -1), and the resulting value could even be negative leading to a negative $F_{DBM|NH}$, which is an illegal value for a sample from an F-distribution (a ratio of two variances). In 1941 Satterthwaite (Satterthwaite, 1941, 1946) proposed an approximate degree of freedom for a linear combination of simple mean square quantities. TBA Online Appendix 9.A explains the approximation in more detail. The end result is that the mean square quantity described in Equation (3.15) has an approximate degree of freedom defined by (this is called the *Satterthwaite's approximation*):

$$ddf_{Sat} = \frac{(MSTR + MSTC - MSTRC)^2}{\left(\frac{MSTR^2}{(I-1)(J-1)} + \frac{MSTC^2}{(I-1)(K-1)} + \frac{MSTRC^2}{(I-1)(J-1)(K-1)}\right)} \tag{3.17}$$

The subscript Sat is for Satterthwaite. From Equation (3.17) it should be fairly obvious that in general ddf_{Sat} is not an integer. To accommodate possible negative estimates of the denominator of Equation (3.17), the original DBM method (Dorfman et al., 1992) proposed, depending on the signs of $\sigma^2_{\tau R}$ and $\sigma^2_{\tau C}$, four expressions for the F-statistic and corresponding expressions for ddf. Rather than repeat them here, since they have been superseded by the method described below, the interested reader is referred to Eqn. 6 and Eqn. 7 in Reference (Hillis et al., 2008).

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

$$F_{DBM} = \frac{MST}{MSTR + \max(MSTC - MSTRC, 0)} \tag{3.18}$$

Now the denominator cannot be negative. One can think of the F-statistic F_{DBM} as a signal-to-noise ratio like quantity, with the difference that both numerator and denominator are variance like quantities. If the "variance" represented by the treatment effect is larger than the variance of the noise tending to mask the treatment effect, then F_{DBM} tends to be large, which makes the observed treatment "variance" stand out more clearly compared to the noise, and the NH is more likely to be rejected. Hillis in (Hillis et al., 2005) has shown that the left hand side of Equation (3.18) is distributed as an F-statistic with ndf $= I - 1$ and denominator degrees of freedom ddf_H defined by:

$$ddf_H = \frac{(MSTR + \max(MSTC - MSTRC, 0))^2}{\mathrm{MSTR}^2}(I-1)(J-1) \tag{3.19}$$

Summarizing,

$$F_{DBM} \sim F_{\text{ndf},\text{ddf}_H} \text{ndf} = I - 1 \tag{3.20}$$

Instead of 4 rules, as in the original DBM method, the Hillis modification involves just one rule, summarized by Equations (3.19) through (3.20). Moreover, the F-statistic is constrained to non-negative values. Using simulation testing (Hillis et al., 2008) he has been shown that the modified DBM method has better null hypothesis behavior than the original DBM method. The latter tended to be too conservative, typically yielding Type I error rates smaller than the expected 5% for $\alpha = 0.05$.

3.3.4 Decision rules, p-value and confidence intervals

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

$$\Pr\left(F \leq F_{1-\alpha,\text{ndf},\text{ddf}_H} \mid F \sim F_{\text{ndf},\text{ddf}_H}\right) = 1-\alpha \tag{3.21}$$

The critical value $F_{1-\alpha,\text{ndf},\text{ddf}_H}$ increases as α decreases. The value of α, generally chosen to be 0.05, termed the *nominal* α, is fixed. The decision rule is that if $F_{DBM} > F_{1-\alpha,\text{ndf},\text{ddf}_H}$ one rejects the NH and otherwise one does not. It follows, from the definition of F_{DBM}, Equation (3.18), that rejection of the NH is more likely to occur if: * F_{DBM} is large, which occurs if MST is large, meaning the treatment effect is large * $MSTR + \max(MSTC - MSTRC, 0)$ is small, see comments following TBA (3.1) Eqn. (9.23). * α is large: for then $F_{1-\alpha,\text{ndf},\text{ddf}_H}$ decreases and is more likely to be exceeded by the observed value of F_{DBM}. * ndf is large: the more the number of treatment pairings, the greater the chance that at least one pairing will reject the NH. This is one reason sample size calculations are rarely conducted for more than 2-treatments. * ddf_H is large: this causes the critical value to decrease, see below, and is more likely to be exceeded by the observed value of F_{DBM}.

3.3.4.1 p-value of the F-test

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

$$p = \Pr\left(F > F_{DBM} \mid F \sim F_{\text{ndf},\text{ddf}_H}\right) \tag{3.22}$$

3.3.4.2 Confidence intervals for inter-treatment FOM differences

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

$$CI_{1-\alpha} = (Y_{i\bullet\bullet} - Y_{i'\bullet\bullet}) \pm t_{\alpha/2;\mathrm{ddf}_H} \sqrt{\frac{2}{JK}\left(MSTR + \max\left(MSTC - MSTRC, 0\right)\right)} \tag{3.23}$$

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

$$\Pr\left(T > t_{\alpha/2;\mathrm{ddf}_H}\right) = \alpha/2 \tag{3.24}$$

Since centered pseudovalues were used:

$$(Y_{i\bullet\bullet} - Y_{i'\bullet\bullet}) = (\theta_{i\bullet} - \theta_{i'\bullet}) \tag{3.25}$$

Therefore, Equation (3.23) can be rewritten:

$$CI_{1-\alpha} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2;\mathrm{ddf}_H} \sqrt{\frac{2}{JK}\left(MSTR + \max\left(MSTC - MSTRC, 0\right)\right)} \tag{3.26}$$

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

- $F_{DBM} > F_{1-\alpha,\mathrm{ndf},\mathrm{ddf}_H}$
- $p < \alpha$
- $CI_{1-alpha}$ excludes zero

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

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

Line 1 defines α. Line 2 forms a data frame from previously calculated mean-squares. Line 3 calculates the denominator appearing in Equation (3.18). Line 4 computes the observed value of F_{DBM}, namely the ratio of the numerator and denominator in Equation (3.18). Line 5 sets ndf to $I-1$. Line 6 computes ddf_H. Line 7 computes the critical value of the F-distribution $F_{crit} \equiv F_{\text{ndf},\text{ddf}_H}$. Line 8 calculates the p-value, using the definition Equation (3.22). Line 9 prints out the just calculated quantities. The next line uses the `RJafroc` function `StSignificanceTesting()` and the 2nd last line prints out corresponding `RJafroc`-computed quantities. Note the correspondences between the values just computed and those provide by `RJafroc`. Note that the FOM difference is not significant at the 5% level of significance as $p > \alpha$. The last line shows that F_{DBM} does not exceed F_{crit}. The two rules are equivalent.

```
alpha <- 0.05
retMS <- data.frame("msT" = msT, "msR" = msR, "msC" = msC,
                    "msTR" = msTR, "msTC" = msTC,
                    "msRC" = msRC, "msTRC" = msTRC)
F_DBM_den <- retMS$msTR+max(retMS$msTC - retMS$msTRC,0)
F_DBM <- retMS$msT / F_DBM_den
ndf <- (I-1)
ddf_H <- (F_DBM_den^2/retMS$msTR^2)*(I-1)*(J-1)
FCrit <- qf(1 - alpha, ndf, ddf_H)
pValueH <- 1 - pf(F_DBM, ndf, ddf_H)
data.frame("F_DBM" = F_DBM, "ddf_H"= ddf_H, "pValueH" = pValueH) # Line 9
#>      F_DBM    ddf_H    pValueH
#> 1 4.456319 15.25967 0.05166569
retRJafroc <- StSignificanceTesting(dataset02,
                                    FOM = "Wilcoxon",
                                    method = "DBM")
data.frame("F_DBM" = retRJafroc$RRRC$FTests$FStat[1],
           "ddf_H"= retRJafroc$RRRC$FTests$DF[2],
           "pValueH" = retRJafroc$RRRC$FTests$p[1])
#>       F_DBM     ddf_H     pValueH
#> 1 4.4563187 15.259675 0.051665686
F_DBM > FCrit
#> [1] FALSE
```

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

Line 1 computes the FOM matrix using function `UtilFigureOfMerit`. The next 9 lines compute the treatment FOM differences. The next line `nDiffs` (for

"number of differences") evaluates to 1, as with two treatments, there is only one difference. The next line initializes `CI_DIFF_FOM_RRRC`, which stands for "confidence intervals, FOM differences, for RRRC analysis". The next 8 lines evaluate, using Equation (3.26), and prints the lower value, the mid-point and the upper value of the confidence interval. Finally, these values are compared to those yielded by `RJafroc`. The FOM difference is not significant, whether viewed from the point of view of the F-statistic not exceeding the critical value, the observed p-value being larger than alpha or the 95% CI for the FOM difference including zero.

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

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

3.4.1 The non-centrality parameter

In the significance-testing procedure just described, the relevant distribution was that of the F-statistic when the NH is true, Equation (3.20). *For sample size estimation, one needs to know the distribution of the statistic when the NH is false.* In the latter condition (i.e., the AH) the observed F-statistic, defined by Equation (3.15), is distributed as a *non-central* F-distribution $F_{\text{ndf},\text{ddf}_H,\Delta}$ with *non-centrality parameter* Δ:

$$F_{DBM|AH} \sim F_{\text{ndf},ddf_H,\Delta} \tag{3.27}$$

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

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

The parameters σ_τ^2, $\sigma_{\tau R}^2$ and $\sigma_{\tau C}^2$ appearing in this equation are identical to three of the six variances describing the DBM model, Equation (3.4). The estimates of $\sigma_{\tau R}^2$ and/or $\sigma_{\tau C}^2$ can turn out to be negative (if either of these parameters is close to zero, an estimate from a small pilot study can be negative). To avoid a possibly negative denominator, (Hillis and Berbaum, 2004) suggest the following modifications (see sentence following Eqn. 4 in cited paper):

$$\Delta = \frac{JK\sigma_\tau^2}{\sigma_\epsilon^2 + \max(K\sigma_{\tau R}^2, 0) + \max(J\sigma_{\tau C}^2, 0)} \tag{3.28}$$

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

$$d = \theta_{1\bullet} - \theta_{2\bullet} \tag{3.29}$$

For two treatments, since the individual treatment effects must be the negatives of each other (because they sum to zero, see (3.5)), it follows that:

$$\sigma_\tau^2 = \frac{d^2}{2} \tag{3.30}$$

Therefore, for two treatments the numerator of the expression for Δ is $JKd^2/2$. Dividing numerator and denominator of Equation (3.28) by K, one gets the final expression for Δ, as coded in `RJafroc`, namely:

$$\Delta = \frac{Jd^2/2}{\max(\sigma_{\tau R}^2, 0) + (\sigma_\epsilon^2 + \max(J\sigma_{\tau C}^2, 0))/K} \tag{3.31}$$

The variances, σ_τ^2, $\sigma_{\tau R}^2$ and $\sigma_{\tau C}^2$, appearing in Equation (3.31), can be calculated from the observed mean squares using the following equations, see (Hillis and Berbaum, 2004) Eqn. 4,

$$\left.\begin{aligned} \sigma_\epsilon^2 &= \mathrm{MSTRC}^* \\ \sigma_{\tau R}^2 &= \frac{\mathrm{MSTR}^* - \mathrm{MSTRC}^*}{K^*} \\ \sigma_{\tau C}^2 &= \frac{\mathrm{MSTC}^* - \mathrm{MSTRC}^*}{J^*} \end{aligned}\right\} \tag{3.32}$$

- Here the asterisk is used to (consistently) denote quantities, including the mean squares, pertaining to the *pilot* study.
- In particular, J^* and K^* denote the numbers of readers and cases, respectively, *in the pilot study*, while J and K, appearing elsewhere, for example in Equation (3.31), are the corresponding numbers for the *planned or pivotal study*.
- The three variances, determined from the pilot study via Equation (3.32), are assumed to apply unchanged to the pivotal study (as they are sample-size independent parameters of the DBM model).

3.4.2 The denominator degrees of freedom

- (The numerator degrees of freedom of the non-central F distribution is always unity.) It remains to calculate the appropriate denominator degrees of freedom for the pivotal study. This is denoted df_2, to distinguish it from ddf_H, where the latter applies to the pilot study as in Equation (3.19).
- The starting point is Equation (3.19) with the left hand side replaced by df_2, and with the emphasis that *all quantities appearing in it apply to the pivotal study*.
- The mean squares appearing in Equation (3.19) can be related to the variances by an equation analogous to Equation (3.32), except that, again, all quantities in it apply to the *pivotal* study (note the absence of asterisks):

$$\left.\begin{aligned} \sigma_\epsilon^2 &= MSTRC \\ \sigma_{\tau R}^2 &= \frac{MSTR - MSTRC}{K} \\ \sigma_{\tau C}^2 &= \frac{MSTC - MSTRC}{J} \end{aligned}\right\} \tag{3.33}$$

Substituting from Equation (3.33) into Equation (3.19) with the left hand side replaced by df_2, and dividing numerator and denominator by K^2, one has the final expression as coded in `RJafroc`:

$$df_2 = \frac{(\max(\sigma_{\tau R}^2, 0) + (\max(J\sigma_{\tau C}^2, 0) + \sigma_\epsilon^2)/K)^2}{(\max(\sigma_{\tau R}^2, 0) + \sigma_\epsilon^2/K)^2}(J-1) \tag{3.34}$$

3.4.3 Example of sample size estimation, RRRC generalization

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

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

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

```
#RRRC
J <- 10; K <- 163
den <- max(VarTR, 0) + (VarErr + J * max(VarTC, 0)) / K
deltaRRRC <- (d^2 * J/2) / den
df2 <- den^2 * (J - 1) / (max(VarTR, 0) + VarErr / K)^2
fvalueRRRC <- qf(1 - alpha, 1, df2)
Power <- 1-pf(fvalueRRRC, 1, df2, ncp = deltaRRRC)
data.frame("J"= J,  "K" = K, "fvalueRRRC" = fvalueRRRC, "df2" = df2, "deltaRRRC" = deltaRRRC, "Po
#>    J   K fvalueRRRC       df2 deltaRRRC  PowerRRRC
#> 1 10 163  3.9930236 63.137871 8.1269825 0.80156249
```

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

The extension to FRRC generalization is as follows. One sets $\sigma^2_R = 0$ and $\sigma^2_{\tau R} = 0$ in the DBM model (3.4). The F-statistic for testing the NH and its distribution under the NH is:

$$F = \frac{\text{MST}}{\text{MSTC}} \sim F_{I-1,(I-1)(K-1)} \tag{3.35}$$

The NH is rejected if the observed value of F exceeds the critical value defined by $F_{\alpha,I-1,(I-1)(K-1)}$. For two modalities the denominator degrees of freedom is $df_2 = K - 1$. The expression for the non-centrality parameter follows from (3.31) upon setting $\sigma^2_{\tau R} = 0$.

$$\Delta = \frac{Jd^2/2}{(\sigma^2_\epsilon + \max(J\sigma^2_{\tau C}, 0))/K} \tag{3.36}$$

These equations are coded in the following code-chunk:

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

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

The extension to RRFC generalization is as follows. One sets $\sigma^2_C = 0$ and $\sigma^2_{\tau C} = 0$ in the DBM model (3.4). The F-statistic for testing the NH and its distribution under the NH is:

$$F = \frac{\text{MST}}{\text{MSTR}} \sim F_{I-1,(I-1)(J-1)} \tag{3.37}$$

The NH is rejected if the observed value of F exceeds the critical value defined by $F_{\alpha,I-1,(I-1)(J-1)}$. For two modalities the denominator degrees of freedom is $df_2 = J-1$. The expression for the non-centrality parameter follows from (3.31) upon setting $\sigma^2_{\tau C} = 0$.

$$\Delta = \frac{Jd^2/2}{\max(\sigma^2_{\tau R}, 0) + \sigma^2_{\epsilon}/K} \tag{3.38}$$

These equations are coded in the following code-chunk:

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

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

3.7 Summary TBA

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

DBM has been used in several hundred ROC studies (Prof. Kevin Berbaum, private communication ca. 2010). While the method allows generalization of a study finding, e.g., rejection of the NH, to the population of readers and cases, the author believes this is sometimes taken too literally. If a study is done at a single hospital, then the radiologists tend to be more homogenous as compared to sampling radiologists from different hospitals. This is because close

interactions between radiologists at a hospital tend to homogenize reading styles and performance. A similar issue applies to patient characteristics, which are also expected to vary more between different geographical locations than within a given location served by the hospital. This means is that single hospital study based p-values may tend to be biased downwards, declaring differences that may not be replicable if a wider sampling "net" were used using the same sample size. The price paid for a wider sampling net is that one must use more readers and cases to achieve the same sensitivity to genuine treatment effects, i.e., statistical power (i.e., there is no "free-lunch").

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

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

is out dated; the parameter values are based on datasets then available (prior to 1997). Medical imaging technology has changed substantially in the intervening decades. d Finally, the methodology used to arrive at the proposed parameter values is not clearly described. Needed is a more realistic simulator, incorporating knowledge from alternative ROC models and paradigms that is calibrated, by a clearly defined method, to current datasets.

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

3.8 Things for me to think about

3.8.1 Expected values of mean squares

Assuming no replications the expected mean squares are as follows, Table Table 3.1; understanding how this table is derived, would lead the author well outside his expertise and the scope of this book; suffice to say that these are *unconstrained* estimates (as summarized in the quotation above) which are different from the *constrained* estimates appearing in the original DBM publication (Dorfman et al., 1992), Table 9.2; the differences between these two types of estimates is summarized in (Dorfman et al., 1995). For reference, Table 9.3 is the table published in the most recent paper that I am aware of (Hillis, 2014). All three tables are different! **In this chapter I will stick to Table Table 3.1 for the subsequent development.**

Table 3.2: Table 9.1 Unconstrained expected values of mean-squares, as in (Dorfman et al., 1995)

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

Table 3.3: Table 9.2 Constrained expected values of mean-squares, as in (Dorfman et al., 1992)

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

Table 3.4: Table 9.3 As in Hillis "marginal-means ANOVA paper" (Hillis, 2014)

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

3.9 References

Chapter 4

DBM method special cases

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

4.1 Fixed-reader random-case (FRRC) analysis

The model is the same as in TBA (3.1) Eqn. (9.4) except one puts $\sigma_R^2 = \sigma_{\tau R}^2 = 0$ in Table Table 3.1. The appropriate test statistic is:

$$\frac{E\left(MST\right)}{E\left(MSTC\right)} = \frac{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2 + JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2} \tag{4.1}$$

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

$$\frac{E\left(MST\right)}{E\left(MSTC\right)} = 1 \tag{4.2}$$

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

$$F_{DBM|R} = \frac{MST}{MSTC} \tag{4.3}$$

The observed value $F_{DBM|R}$ (the Roe-Metz notation (Roe and Metz, 1997) is used which indicates that the factor appearing to the right of the vertical bar is regarded as fixed) is distributed as an F-statistic with ndf = I–1 and $ddf = (I-1)(K-1)$; the degrees of freedom follow from the rows labeled T and TC in TBA Table Table 3.1. Therefore, the distribution of the observed

value is (no Satterthwaite approximation needed this time as both numerator and denominator are simple mean-squares):

$$F_{DBM|R} \sim F_{I-1,(I-1)(K-1)} \tag{4.4}$$

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

$$F_{DBM|R} > F_{1-\alpha,I-1,(I-1)(K-1)} \tag{4.5}$$

The p-value of the test is the probability that a random sample from the F-distribution TBA (3.1) Eqn. (9.39), exceeds the observed value:

$$p = \Pr\left(F > F_{DBM|R} \mid F \sim F_{I-1,(I-1)(K-1)}\right) \tag{4.6}$$

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

$$CI_{1-\alpha} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,(I-1)(K-1)}\sqrt{2\frac{MST}{JK}} \tag{4.7}$$

4.1.1 Single-reader multiple-treatment analysis

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

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

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

theta <- as.matrix(UtilFigureOfMerit(dataset02, FOM = "Wilcoxon"))
theta_i_dot <- array(dim = I)
```

```
for (i in 1:I) theta_i_dot[i] <- mean(theta[i,])

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

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

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

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

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

4.2 Random-reader fixed-case (RRFC) analysis

The model is the same as in TBA (3.1) Eqn. (9.4) except one puts $\sigma_C^2 = \sigma_{\tau C}^2 = 0$ in Table Table 3.1. It follows that:

$$\frac{E(MST)}{E(MSTR)} = \frac{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2 + JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2} \tag{4.8}$$

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

$$\frac{E(MST)}{E(MSTR)} = 1 \tag{4.9}$$

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

$$F_{DBM|C} \sim \frac{MST}{MSTR} \tag{4.10}$$

The observed value $F_{DBM|C}$ is distributed as an F-statistic with $ndf = I$–1 and $ddf = (I-1)(J-1)$, see rows labeled T and TR in Table Table 3.1.

$$F_{DBM|C} \sim F_{I-1,(I-1)(J-1))} \tag{4.11}$$

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

$$F_{DBM|C} > F_{1-\alpha,I-1,(I-1)(J-1))} \tag{4.12}$$

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

$$p = \Pr\left(F > F_{DBM|C} \mid F \sim F_{I-1,(I-1)(J-1)}\right) \tag{4.13}$$

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

$$CI_{1-\alpha} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,(I-1)(J-1)}\sqrt{2\frac{MSTR}{JK}} \tag{4.14}$$

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

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

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

4.3 References

Chapter 5

Introduction to the Obuchowski Rockette (OR) formulation of significance testing

5.1 Introduction

This chapter starts with a gentle introduction to the Obuchowski and Rockette method (alternatively, this chapter could be titled "An introduction to covariance matrices"). The reason is that the method was rather opaque to me, and I suspect most non-statistician users. Part of the problem, in my opinion, is the notation, namely lack of the *case-set* index $\{c\}$. While this may seem like a trivial point to statisticians, it did present a conceptual problem for me. A key difference of the Obuchowski and Rockette method from DBM is in how the error term is modeled by a non-diagonal covariance matrix. Therefore, the structure of the covariance matrix is examined in some detail.

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

5.2 Single-reader multiple-treatment OR 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. This is because reader is a random factor while treatment is a fixed factor.

The figure of merit θ is modeled as:

$$\theta_{i\{c\}} = \mu + \tau_i + \epsilon_{i\{c\}} \tag{5.1}$$

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

Eqn. (5.1) 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:

$$\sum_{i=1}^{I} \tau_i = 0 \tag{5.2}$$

The left hand side of Eqn. (5.1) 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.

In my opinion, the case-set index is essential for clarity; without it θ_i is a fixed quantity - the figure of merit estimate for treatment i - lacking an index allowing for sampling related variability. Obuchowski and Rockette define a *k-index*, the "k^{th} repetition of the study involving the same diagnostic test, reader and patient (sic)". In my 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 (Swets and Pickett, 1982). As noted earlier, within-reader variability cannot be "turned off" and affects the interpretations of all case-sets.

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

Finally, and this is where I had some difficulty understanding what was going on, Eqn. (5.1) has an additive random error term $\epsilon_{i\{c\}}$ whose sampling behavior is described by a multivariate normal distribution with an I-dimensional zero mean vector and an $I \times I$ dimensional covariance matrix Σ:

$$\epsilon_{i\{c\}} \sim N_I\left(\vec{0}, \Sigma\right) \tag{5.3}$$

Here N_I is the I-variate normal distribution (i.e., each sample yields I random numbers). For the single-reader model Eqn. (5.1), the covariance matrix has the following structure :

$$\Sigma_{ii'} = Cov\left(\epsilon_{i\{c\}}, \epsilon_{i'\{c\}}\right) = \begin{cases} Var & (i = i') \\ Cov_1 & (i \neq i') \end{cases} \tag{5.4}$$

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:

$$\Sigma = \begin{pmatrix} Var & Cov_1 & ... & Cov_1 & Cov_1 \\ Cov_1 & Var & ... & Cov_1 & Cov_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Cov_1 & Cov_1 & ... & Var & Cov_1 \\ Cov_1 & Cov_1 & ... & Cov_1 & Var \end{pmatrix} \tag{5.5}$$

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.

5.2.1 Definitions of covariance and correlation

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

$$Cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i - x_\bullet)(y_i - y_\bullet)}{N-1} = E(XY) - E(X) - E(Y) \tag{5.6}$$

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:

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

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

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

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

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

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

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

It has the property:

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

5.2.2 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:

$$Var(X-Y) = Var(X) + Var(Y) - 2Cov(X,Y) = 2(Var - Cov) \qquad (5.7)$$

The right hand side specializes to the OR single-reader multiple-treatment model where the variances (for different treatments) are equal and likewise the covariances in Eqn. (5.5) 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, only elements at and above the diagonal are shown):

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & & \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & & & \sigma^2 & \rho_1\sigma^2 \\ & & & & \sigma^2 \end{bmatrix} \tag{5.8}$$

5.2.3 Estimating the variance-covariance matrix

An unbiased estimate of the covariance matrix Eqn. (5.4) follows from:

$$\Sigma_{ii'} \mid_{ps} = \frac{1}{C-1} \sum_{c=1}^{C} \left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right) \left(\theta_{i'\{c\}} - \theta_{i'\{\bullet\}}\right) \tag{5.9}$$

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

$$\mathrm{Var}_i \mid_{ps} = \frac{1}{C-1} \sum_{c=1}^{C} \left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)^2 \tag{5.10}$$

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

Sampling different case-sets, as required by Eqn. (5.9), is unrealistic. In reality one has $C = 1$, i.e., a single dataset. Therefore, direct application of this formula is impossible. However, as seen when this situation was encountered before in (book) Chapter 07, one uses resampling methods to realize, for example, different bootstrap samples, which are resampling-based "stand-ins" for actual case-sets. If B is the total number of bootstraps, then the estimation formula is:

$$\Sigma_{ii'} \mid_{bs} = \frac{1}{B-1} \sum_{b=1}^{B} \left(\theta_{i\{b\}} - \theta_{i\{\bullet\}}\right) \left(\theta_{i'\{b\}} - \theta_{i'\{\bullet\}}\right) \tag{5.11}$$

Eqn. (5.11), the bootstrap method of estimating the covariance matrix, is a direct translation of Eqn. (5.9). Alternatively, one could have used the jackknife FOM values $\theta_{i(k)}$, i.e., the figure of merit with a case k removed, repeated for all k, to estimate the covariance matrix:

$$\Sigma_{ii'} \mid_{jk} = \frac{(K-1)^2}{K} \left[\frac{1}{K-1} \sum_{k=1}^{K} \left(\theta_{i(k)} - \theta_{i(\bullet)}\right) \left(\theta_{i'(k)} - \theta_{i'(\bullet)}\right) \right] \tag{5.12}$$

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

Note the subtle difference in usage of ellipses and parentheses between Eqn. (5.9) and Eqn. (5.12). In the former, the subscript $\{c\}$ denotes a set of K cases while in the latter, (k) denotes the original case set with case k removed, leaving $K-1$ cases. There is a similar subtle difference in usage of ellipses and parentheses between Eqn. (5.11) and Eqn. (5.12). The subscript enclosed in parenthesis, i.e., (k), denotes the FOM with case k removed, while in the bootstrap equation one uses the ellipses (curly brackets) $\{b\}$ to denote the b^{th} bootstrap *case-set*, i.e., a whole set of K_1 non-diseased and K_2 diseased cases, sampled with replacement from the original dataset.

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

5.2.4 The variance inflation factor

In Eqn. (5.12), the expression for the jackknife covariance estimate contains a *variance inflation factor*:

$$\frac{(K-1)^2}{K} \tag{5.13}$$

This factor multiplies the traditional expression for the covariance, shown in square brackets in Eqn. (5.12). It is only needed for the jackknife estimate. The bootstrap and the DeLong estimate, see next, do not require this factor.

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

5.2.5 Meaning of the covariance matrix in Eqn. (5.5)

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

$$\sigma_i^2 \equiv Var_i = \frac{1}{C-1}\sum_{c=1}^{C}\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right) \tag{5.14}$$

The process is repeated for all treatments and the I-variance values are averaged. This is the final estimate of Var appearing in Eqn. (5.3).

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

$$\mathrm{Cov}_{1;ii'} = \frac{1}{C-1}\sum_{c=1}^{C}\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i'\{c\}} - \theta_{i'\{\bullet\}}\right) \tag{5.15}$$

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

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

$$\rho_{1;ii'} = \frac{1}{C-1}\sum_{c=1}^{C}\frac{\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i'\{c\}} - \theta_{i'\{\bullet\}}\right)}{\sigma_i\sigma_{i'}} \tag{5.16}$$

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

To summarize, the covariance matrix can be estimated using the jackknife or the bootstrap, or, in the special case of the empirical AUC figure of merit, the DeLong method can be used. In (book) Chapter 07, these three methods were described in the context of estimating the *variance* of AUC. Eqn. (5.11) and Eqn. (5.12) 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).

5.2.6 Code illustrating the covariance matrix

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

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

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

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

The following notation is used in the code below:

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

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

```
ret1 <- VarCov1_Jk(zik1, zik2)
Var <- ret1$Var
Cov1 <- ret1$Cov1 # use these (i.e., jackknife) as default values in subsequent code
data.frame ("Cov1_jk" = Cov1, "Var_jk" = Var)
#>        Cov1_jk       Var_jk
#> 1 0.0003734661 0.0006989006

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

```
            "Var_rj_jk" = ret4$VarCom["Var", "Estimates"])
#>     Cov1_rj_jk    Var_rj_jk
#> 1 0.0003734661 0.0006989006
```

Note that the estimates are identical and that the 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 `UtilORVarComponentsFactorial`:

```
ret5 <- UtilORVarComponentsFactorial(
  rocData1R, FOM = "Wilcoxon",
  covEstMethod = "bootstrap", nBoots = 2000, seed = 100)
data.frame ("Cov_rj_bs" = ret5$VarCom["Cov1", "Estimates"],
            "Var_rj_bs" = ret5$VarCom["Var", "Estimates"])
#>      Cov_rj_bs    Var_rj_bs
#> 1 0.0003466804 0.0006738506
```

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

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

```
mtrxDLStr <- VarCovMtrxDLStr(rocData1R)
ret3 <- VarCovs(mtrxDLStr)
data.frame ("Cov_dl" = ret3$cov1, "Var_dl" = ret3$var)
#>          Cov_dl       Var_dl
#> 1 0.0003684357 0.0006900766

ret5 <- UtilORVarComponentsFactorial(
  rocData1R, FOM = "Wilcoxon", covEstMethod = "DeLong")
data.frame ("Cov_rj_dl" = ret5$VarCom["Cov1", "Estimates"],
            "Var_rj_dl" = ret5$VarCom["Var", "Estimates"])
#>       Cov_rj_dl    Var_rj_dl
#> 1 0.0003684357 0.0006900766
```

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

5.2.7 Significance testing

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

$$MS(T) = \frac{1}{I-1}\sum_{i=1}^{I}(\theta_i - \theta_\bullet)^2 \tag{5.17}$$

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

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

$$\chi_{1\mathrm{R}} \equiv \frac{(I-1)MS(T)}{Var - Cov_1} \sim \chi^2_{I-1} \tag{5.18}$$

Eqn. (5.18) is from §5.4 in (Hillis, 2007) with two covariance terms "zeroed out" because they are multiplied by $J-1=0$ (since we are restricting to $J=1$).

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

$$F_{1\mathrm{R}} \equiv \frac{MS(T)}{Var - Cov_1} \sim F_{I-1,\infty} \tag{5.19}$$

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

$$\Pr(d \leq D_{1-\alpha} \mid d \sim D) = 1-\alpha \tag{5.20}$$

With definition Eqn. (5.20), 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):

$$\frac{\chi^2_{1-\alpha,I-1}}{I-1} = F_{1-\alpha,I-1,\infty} \tag{5.21}$$

Eqn. (5.21) 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
```

5.2.8 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:

$$p \equiv \Pr(f > F_{1R} \mid f \sim F_{I-1,\infty}) \tag{5.22}$$

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

$$CI_{1-\alpha,1R} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,\infty}\sqrt{2(Var - Cov_1)} \tag{5.23}$$

Comparing Eqn. (5.23) to Eqn. (5.7) 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.]

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

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

Shown below are results obtained using `RJafroc` function `StSignificanceTesting` with `analysisOption = "FRRC"` for `method` = "DBM" (which uses the jackknife), and for OR using 3 different ways of estimating the covariance matrix for the one-reader analysis (i.e., Cov_1 and Var).

```
ret1 <- StSignificanceTesting(
  rocData1R,FOM = "Wilcoxon", method = "DBM", analysisOption = "FRRC")
data.frame("DBM:F" = ret1$FRRC$FTests["Treatment", "FStat"],
           "DBM:ddf" = ret1$FRRC$FTests["Treatment", "DF"],
           "DBM:P-val" = ret1$FRRC$FTests["Treatment", "p"])
#>       DBM.F DBM.ddf  DBM.P.val
#> 1 1.2201111       1 0.27168532

ret2 <- StSignificanceTesting(
  rocData1R,FOM = "Wilcoxon", method = "OR", analysisOption = "FRRC")
data.frame("ORJack:Chisq" = ret2$FRRC$FTests["Treatment", "Chisq"],
           "ORJack:ddf" = ret2$FRRC$FTests["Treatment", "DF"],
           "ORJack:P-val" = ret2$FRRC$FTests["Treatment", "p"])
#>   ORJack.Chisq ORJack.ddf ORJack.P.val
#> 1    1.2201111          1   0.26933885

ret3 <- StSignificanceTesting(
  rocData1R,FOM = "Wilcoxon", method = "OR", analysisOption = "FRRC",
                              covEstMethod = "DeLong")
data.frame("ORDeLong:Chisq" = ret3$FRRC$FTests["Treatment", "Chisq"],
           "ORDeLong:ddf" = ret3$FRRC$FTests["Treatment", "DF"],
           "ORDeLong:P-val" = ret3$FRRC$FTests["Treatment", "p"])
#>   ORDeLong.Chisq ORDeLong.ddf ORDeLong.P.val
#> 1      1.2345017            1     0.26653335

ret4 <- StSignificanceTesting(
  rocData1R,FOM = "Wilcoxon", method = "OR", analysisOption = "FRRC",
                              covEstMethod = "bootstrap")
```

```
data.frame("ORBoot:Chisq" = ret4$FRRC$FTests["Treatment", "Chisq"],
           "ORBoot:ddf" = ret4$FRRC$FTests["Treatment", "DF"],
           "ORBoot:P-val" = ret4$FRRC$FTests["Treatment", "p"])
#>   ORBoot.Chisq ORBoot.ddf ORBoot.P.val
#> 1    1.2311613          1   0.26718131
```

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

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

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

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

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

trtDiff <- array(dim = c(I,I))
for (i1 in 1:(I-1)) {
  for (i2 in (i1+1):I) {
    trtDiff[i1,i2] <- theta_i[i1]- theta_i[i2]
  }
}
trtDiff <- trtDiff[!is.na(trtDiff)]
nDiffs <- I*(I-1)/2
CI_DIFF_FOM_1RMT <- array(dim = c(nDiffs, 3))
for (i in 1 : nDiffs) {
  CI_DIFF_FOM_1RMT[i,1] <- trtDiff[i] + qt(alpha/2,  df = Inf)*sqrt(2*(Var - Cov1))
  CI_DIFF_FOM_1RMT[i,2] <- trtDiff[i]
  CI_DIFF_FOM_1RMT[i,3] <- trtDiff[i] + qt(1-alpha/2,df = Inf)*sqrt(2*(Var - Cov1))
  print(data.frame("theta_1" = theta_i[1],
                   "theta_2" = theta_i[2],
                   "Var" = Var,
                   "Cov1" = Cov1,
                   "MS_T" = MS_T,
                   "F_1R" = F_1R,
                   "pValue" = pValue,
                   "Lower" = CI_DIFF_FOM_1RMT[i,1],
```

```
                  "Mid" = CI_DIFF_FOM_1RMT[i,2],
                  "Upper" = CI_DIFF_FOM_1RMT[i,3]))
}
#>      theta_1    theta_2           Var          Cov1          MS_T      F_1R
#> 1 0.91964573 0.94782609 0.00069890056 0.0003734661 0.00039706618 1.2201111
#>       pValue        Lower          Mid       Upper
#> 1 0.26933885 -0.078183215 -0.028180354 0.021822507
```

The following shows the corresponding output of `RJafroc`.

```
ret_rj <- StSignificanceTesting(
  rocData1R, FOM = "Wilcoxon", method = "OR", analysisOption = "FRRC")
print(data.frame("theta_1" = ret_rj$FOMs$foms[1,1],
                 "theta_2" = ret_rj$FOMs$foms[2,1],
                 "Var" = ret_rj$ANOVA$VarCom["Var", "Estimates"],
                 "Cov1" = ret_rj$ANOVA$VarCom["Cov1", "Estimates"],
                 "MS_T" = ret_rj$ANOVA$TRanova[1,3],
                 "Chisq_1R" = ret_rj$FRRC$FTests["Treatment","Chisq"],
                 "pValue" = ret_rj$FRRC$FTests["Treatment","p"],
                 "Lower" = ret_rj$FRRC$ciDiffTrt[1,"CILower"],
                 "Mid" = ret_rj$FRRC$ciDiffTrt[1,"Estimate"],
                 "Upper" = ret_rj$FRRC$ciDiffTrt[1,"CIUpper"]))
#>      theta_1    theta_2           Var          Cov1          MS_T  Chisq_1R
#> 1 0.91964573 0.94782609 0.00069890056 0.0003734661 0.00039706618 1.2201111
#>       pValue        Lower          Mid       Upper
#> 1 0.26933885 -0.078183215 -0.028180354 0.021822507
```

The first-principles and the `RJafroc` values agree exactly with each other [for $I = 2$, the F and chisquare statistics are identical]. This above code also shows how to extract the different estimates (Var, Cov_1, etc.) from the object `ret_rj` returned by `RJafroc`. Specifically,

- Var: ret_rj$ANOVA$VarCom[“Var”, “Estimates”]
- Cov1: ret_rj$ANOVA$VarCom[“Cov1”, “Estimates”]
- Chisquare-statistic: ret_rj$FRRC$FTests[“Treatment”,“Chisq”]
- df: ret_rj$FRRC$FTests[1,“DF”]
- p-value: ret_rj$FRRC$FTests[“Treatment”,“p”]
- CI Lower: ret_rj$FRRC$ciDiffTrt[1,“CILower”]
- Mid Value: ret_rj$FRRC$ciDiffTrt[1,“Estimate”]
- CI Upper: ret_rj$FRRC$ciDiffTrt[1,“CIUpper”]

5.2.9.1 Jumping ahead

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

`analysisOption` = "RRRC"]:

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

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

5.3 Multiple-reader multiple-treatment OR 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. (5.1), one needs an additional j index to denote reader dependence of the figure of merit, and additional terms to model reader and treatment-reader variability, and the error term needs to be modified to account for the additional random (i.e., reader) factor.

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

$$\theta_{ij\{c\}} = \mu + \tau_i + R_j + (\tau R)_{ij} + \epsilon_{ij\{c\}} \tag{5.24}$$

- The fixed treatment effect τ_i is subject to the usual constraint, Eqn. (5.2).
- The first two terms on the right hand side of Eqn. (5.24) have their usual meanings: a constant term μ representing performance averaged over treatments and readers, and a treatment effect τ_i (i = 1,2, ..., I).

- The next two terms are, by assumption, mutually independent random samples specified as follows:
  - R_j denotes the random treatment-independent figure-of-merit contribution of reader j (j = 1,2, ..., J), modeled by a zero-mean normal distribution with variance σ_R^2;
  - $(\tau R)_{ij}$ denotes the treatment-dependent random contribution of reader j in treatment i, modeled as a sample from a zero-mean normal distribution with variance $\sigma_{\tau R}^2$. There could be a perceived notational clash with similar variance component terms defined for the DBM model – except in that case they applied to pseudovalues. The meaning should be clear from the context.
- Summarizing:

$$\begin{cases} R_j \sim N(0, \sigma_R^2) \\ \tau R \sim N(0, \sigma_{\tau R}^2) \end{cases} \tag{5.25}$$

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):

$$\mu = \theta_{\bullet\bullet\{1\}} \tag{5.26}$$

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

$$\tau_i = \theta_{i\bullet\{1\}} - \theta_{\bullet\bullet\{1\}} \tag{5.27}$$

The τ_i estimates obey the constraint Eqn. (5.2). For example, with two treatments, the values of τ_i must be the negatives of each other: $\tau_1 = -\tau_2$.

The error term on the right hand side of Eqn. (5.24) 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,

$$\epsilon_{ij\{c\}} \sim N_{IJ}(\vec{0}, \Sigma) \tag{5.28}$$

Here N_{IJ} is the IJ-variate normal distribution, $\vec{0}$ is the zero-vector with length IJ, denoting the vector-mean of the distribution. The counterpart of the variance, namely the covariance matrix Σ of the distribution, is defined by 4 parameters, Var, Cov_1, Cov_2, Cov_3, defined as follows:

$$Cov(\epsilon_{ij\{c\}}, \epsilon_{i'j'\{c\}}) = \left\{ \begin{array}{l} Var\ (i = i', j = j') \\ Cov1\ (i \neq i', j = j') \\ Cov2\ (i = i', j \neq j') \\ Cov3\ (i \neq i', j \neq j') \end{array} \right\} \tag{5.29}$$

Apart from fixed effects, the model implied by Eqn. (5.24) and Eqn. (5.29) 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 DBM model, which should not be a surprise since one is modeling the data with equivalent models. The Obuchowski and Rockette model Eqn. (5.24) "looks" simpler because four covariance terms are encapsulated in the ϵ term. As with the singe-reader multiple-treatment model, the covariance matrix is assumed to be independent of treatment or reader.

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

5.3.1 Structure of the covariance matrix

To understand the structure of this matrix, recall that the diagonal elements of a square covariance matrix are the 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 6×6 covariance matrix would look like this:

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

Shown in each cell of the matrix is a pair of ij-values, serving as column indices, followed by a pair of ij-values serving as row indices, and a comma separates the pairs. For example, the first column is labeled by (11,xx), where xx depends on the row. The second column is labeled (12,xx), the third column is labeled (13,xx), and the remaining columns are successively labeled (21,xx), (22,xx)

and (23,xx). Likewise, the first row is labeled by (yy,11), where yy depends on the column. The following rows are labeled (yy,12), (yy,13), (yy,21), (yy,22)and (yy,23). Note that the reader index increments faster than the treatment index.

The diagonal elements are evidently those cells where the row and column index-pairs are equal. These are (11,11), (12,12), (13,13), (21,21), (22,22) and (23,23). According to Eqn. (5.29) these cells represent Var.

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

According to Eqn. (5.29) cells with different treatment indices but identical reader indices represent Cov_1. As an example, cell (21,11) has the same reader indices, namely reader 1, but different treatment indices, namely 2 and 1, so it is 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, cells with identical treatment indices but different reader indices represent 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, cells with different treatment indices and different reader indices represent 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 Var is estimated as follows:

$$Var = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})^2 \right\rangle_{ij} \quad \epsilon_{ij\{c\}} \sim N_{IJ}(\vec{0}, \Sigma) \tag{5.30}$$

Of course, in practice one would use the bootstrap or the jackknife as a stand-in for the c-index (with the understanding that if the jackknife is used, then a variance inflation factor has to be included on the right hand side of Eqn. (5.30). Notice that the left-hand-side of Eqn. (5.30) 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:

$$Cov = \begin{cases} Cov_1 = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})(\theta_{i'j\{c\}} - \theta_{i'j\{\bullet\}}) \right\rangle_{ii',jj} \\ Cov_2 = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})(\theta_{ij'\{c\}} - \theta_{ij'\{\bullet\}}) \right\rangle_{ii,jj'} \\ Cov_3 = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})(\theta_{i'j'\{c\}} - \theta_{i'j'\{\bullet\}}) \right\rangle_{ii',jj'} \end{cases} \tag{5.31}$$

In Eqn. (5.31) *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 $\langle\rangle_{ii',jj'}$.

5.3.2 Physical meanings of the covariance terms

The meanings of the different terms follow a similar description to that given in Eqn. 5.3.1. The diagonal term Var is the variance of the figures-of-merit when reader j interprets different case-sets $\{c\}$ in treatment i. Each case-set yields a number $\theta_{ij\{c\}}$ and the variance of the C numbers, averaged over the $I \times J$ treatments and readers, is Var. It captures the total variability due to varying difficulty levels of the case-sets, inter-reader and within-reader variability.

It is easier to see the physical meanings of Cov_1, Cov_2, Cov_3 if one starts with the correlations.

- $\rho_{1;ii'jj}$ is the correlation of the figures-of-merit when reader j interprets case-sets in different treatments i, i'. Each case-set, starting with $c = 1$, yields two numbers $\theta_{ij\{1\}}$ and $\theta_{i'j\{1\}}$. The correlation of the two pairs of C-length arrays, averaged over all pairings of different treatments and same readers, is ρ_1. The correlation exists due to the common contribution of the shared reader. When the common variation is large, the two arrays become more correlated and ρ_1 approaches unity. If there is no common variation, the two arrays become independent, and ρ_1 equals zero. Converting from correlation to covariance, see Eqn. (5.8), one has $Cov_1 < Var$.

- $\rho_{2;iijj'}$ is the correlation of the figures-of-merit values when different readers j, j' interpret the same case-sets in the same treatment i. As before this yields two C-length arrays, whose correlation, upon averaging over all distinct treatment pairings and same readers, yields ρ_2. If one assumes that common variation between different-reader same-treatment FOMs is smaller than the common variation between same-reader different-treatment FOMs, then ρ_2 will be smaller than ρ_1. This is equivalent to stating that readers agree more with themselves in different treatments than they do with other readers in the same treatment. Translating to covariances, one has $Cov_2 < Cov_1 < Var$.

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

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

$$Cov_3 \leq Cov_2 \leq Cov_1 \leq Var \tag{5.32}$$

5.4 Discussion/Summary/1

5.5 References

Chapter 6

Obuchowski Rockette (OR) Analysis

6.1 Introduction

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

Hillis' has proposed that a method based on an earlier publication (Obuchowski and Rockette, 1995), which does not depend on pseudovalues, is preferable from both conceptual and practical points of view. This chapter is named "OR Analysis", where OR stands for Obuchowski and Rockette. The OR method has advantages in being able to handle more complex study designs (Hillis, 2014) that are addressed in subsequent chapters, and applications to other FOMs (e.g., the FROC paradigm uses a rather different FOM from empirical ROC-AUC) are best performed with the OR method.

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

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

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

6.2 Random-reader random-case (RRRC) analysis

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

$$NH : \tau_i = 0 \;\; (i = 1, 2, ..., I) \tag{6.1}$$

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

$$F_{ORH} = \frac{MS(T)}{MS(TR) + J \max(Cov_2 - Cov_3, 0)} \tag{6.2}$$

Eqn. (6.2) incorporates Hillis' modification of the original OR F-statistic. The modification ensures that the constraint Eqn. (5.32) is always obeyed and also avoids a possibly negative (and hence illegal) F-statistic. The relevant mean squares are defined by (note that these are calculated using *FOM* values, not *pseudovalues*):

$$\left.\begin{aligned} MS(T) &= \tfrac{J}{I-1}\textstyle\sum_{i=1}^{I}(\theta_{i\bullet} - \theta_{\bullet\bullet})^2 \\ MS(R) &= \tfrac{I}{J-1}\textstyle\sum_{j=1}^{J}(\theta_{\bullet j} - \theta_{\bullet\bullet})^2 \\ MS(TR) &= \tfrac{1}{(I-1)(J-1)}\textstyle\sum_{i=1}^{I}\sum_{j=1}^{J}(\theta_{ij} - \theta_{i\bullet} - \theta_{\bullet j} + \theta_{\bullet\bullet}) \end{aligned}\right\} \tag{6.3}$$

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

$$F_{OR} = \frac{MS(T)}{MS(TR) + J(Cov_2 - Cov_3)} \tag{6.4}$$

Note that Eqn. (6.4) lacks the constraint, subsequently proposed by Hillis, which ensures that the denominator cannot be negative. The following distribution was proposed for the test statistic.

$$F_{ORH} \sim F_{ndf,ddf} \tag{6.5}$$

The original degrees of freedom were defined by:

$$\text{ndf} = I - 1 \text{ddf} = (I-1) \times (J-1) \tag{6.6}$$

It turns out that the Obuchowski-Rockette test statistic is very conservative, meaning it is highly biased against rejecting the null hypothesis (the data simulator used in their validation did not detect this behavior). Because of this the predicted sample sizes tended to be quite large. In this connection I have two informative anecdotes.

6.2.1 Two 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 version of Eqn. (6.2) with the *ddf* as originally suggested by Obuchowski and Rockette, predicted such high number of readers and cases that he was doubtful about the chances of anyone conducting a practical ROC study!

- The second story is that the author once conducted NH simulations using a Roe-Metz simulator 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 denominator degrees of freedom *ddf*, see next, substitution of which magically solved the problem, i.e., the simulations rejected the null hypothesis 5% of the time.

6.2.2 Hillis ddf

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

$$\mathrm{ddf}_H = \frac{\left[MS(TR) + J\max(Cov_2 - Cov_3, 0)\right]^2}{\frac{\left[MS(TR)\right]^2}{(I-1)(J-1)}} \tag{6.7}$$

If $Cov_2 < Cov_3$ (which is the *exact opposite* of the expected ordering, Eqn. (5.32)), this reduces to $(I-1)\times(J-1)$, the value originally proposed by Obuchowski and Rockette. With Hillis' proposed changes, under the null hypothesis the observed statistic F_{ORH}, defined in Eqn. (6.2), is distributed as an F-statistic with ndf $= I-1$ and $ddf = ddf_H$ degrees of freedom (Hillis et al., 2005; Hillis, 2007; Hillis et al., 2008):

$$F_{ORH} \sim F_{ndf,ddf_H} \tag{6.8}$$

If the expected ordering is true, i.e., $Cov_2 > Cov_3$, which is the more likely situation, then ddf_H is *larger* than $(I-1)\times(J-1)$, i.e., the Obuchowski-Rockette's *ddf*, and the p-value decreases, i.e., there is a larger probability of rejecting the NH. The modified OR method is more likely to have the correct NH behavior, i.e, it will reject the NH 5% of the time when alpha is set to 0.05 (statisticians refer to this as "the 5% test"). This has been confirmed in simulation testing (Hillis et al. (2008)).

6.2.3 Decision rule, p-value and confidence interval

The critical value of the F-statistic for rejection of the null hypothesis is $F_{1-\alpha,ndf,ddf_H}$, 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 Eqn. (6.2):

- Rejection of the NH is more likely if $MS(T)$ increases, meaning the treatment effect is larger;
- $MS(TR)$ is smaller meaning there is less contamination of the treatment effect by treatment-reader variability;
- The greater of $Cov2$ or $Cov3$, which is usually $Cov2$, decreases, meaning there is less "noise" in the measurement due to between-reader variability. Recall that Cov_2 involves different-reader same-treatment pairings.
- α increases, meaning one is allowing a greater probability of Type I errors;

- ndf increases, as this lowers the critical value of the F-statistic. With more treatment pairings, the chance that at least one paired-difference will reject the NH is larger.
- ddf_H increases, as this lowers the critical value of the F-statistic.

The p-value of the test is the probability, under the NH, that an equal or larger value of the F-statistic than F_{OR} 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_{OR}:

$$p = \Pr(F > F_{OR} \mid F \sim F_{ndf,ddf_H}) \tag{6.9}$$

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

$$CI_{1-\alpha,RRRC} = \theta_{i\bullet} - \theta_{i'\bullet} \pm t_{\alpha/2,ddf_H} \sqrt{\frac{2}{J}(MS(TR) + J\max(Cov_2 - Cov_3, 0))} \tag{6.10}$$

6.3 Fixed-reader random-case (FRRC) analysis

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

$$F_{OR|R} = \frac{MS(T)}{Var - Cov_1 + (J-1)\max(Cov_2 - Cov_3, 0)} \tag{6.11}$$

$F_{OR|R}$ is distributed as an F-statistic with:

$$\left.\begin{array}{c} ndf = I - 1 \\ ddf = \infty \\ F_{OR|R} \sim F_{ndf,ddf} \end{array}\right\} \tag{6.12}$$

Alternatively, as with Eqn. (5.18),

$$(I-1)F_{OR|R} \sim \chi^2_{I-1}$$

For $J = 1$, Eqn. (6.11) reduces to Eqn. (5.19).

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.,:

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

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 Eqn. (6.11):

$$p = \Pr(F > F_{OR|R} \mid F \sim F_{I-1,\infty}) \tag{6.13}$$

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

$$CI_{1-\alpha,FRRC} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,\infty} \sqrt{\frac{2}{J}(Var - Cov_1 + (J-1)\max(Cov_2 - Cov_3, 0))} \tag{6.14}$$

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. (6.13) is less than α.
- The $(1-\alpha)$ confidence interval does not include zero.

Notice that for J = 1, Eqn. (6.14) reduces to Eqn. (5.23).

6.4 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:

$$F_{OR|C} = \frac{MS(T)}{MS(TR)} \tag{6.15}$$

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

$$\left.\begin{array}{c} ndf = I-1 \\ ddf = (I-1)(J-1) \\ F_{OR|C} \sim F_{ndf,ddf} \end{array}\right\} \tag{6.16}$$

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:

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

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

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

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

$$CI_{1-\alpha,RRFC} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,(I-1)(J-1)} \sqrt{\frac{2}{J} MS(TR)} \tag{6.17}$$

6.5 Discussion/Summary/4

6.6 References

Chapter 7

Coding illustrations of the OR method

7.1 Introduction

It is time to reinforce the formulae with examples. We illustrate for the VanDyke dataset, i.e., `dataset02`.

7.1.1 Calculate figures of merit

```
foms <- UtilFigureOfMerit(dataset02, FOM = "Wilcoxon")
print(foms)
#>           rdr0      rdr1      rdr2      rdr3      rdr4
#> trt0 0.9196457 0.8587762 0.9038647 0.9731079 0.8297907
#> trt1 0.9478261 0.9053140 0.9217391 0.9993559 0.9299517
```

7.1.2 Calculate variance covariance and mean-squares

```
ret <- UtilORVarComponentsFactorial(dataset02, FOM = "Wilcoxon", covEstMethod = "jackknife")
print(ret)
#> $TRanova
#>              SS DF           MS
#> T  0.004796171  1 0.0047961705
#> R  0.015344800  4 0.0038362000
#> TR 0.002204122  4 0.0005510306
```

```
#>
#> $VarCom
#>          Estimates      Rhos
#> VarR  0.0015349993        NA
#> VarTR 0.0002004025        NA
#> Cov1  0.0003466137 0.4320314
#> Cov2  0.0003440748 0.4288668
#> Cov3  0.0002390284 0.2979333
#> Var   0.0008022883        NA
#>
#> $IndividualTrt
#>      DF  msREachTrt   varEachTrt  cov2EachTrt
#> trt0  4 0.003082629 0.0010141028 0.0004839618
#> trt1  4 0.001304602 0.0005904738 0.0002041879
#>
#> $IndividualRdr
#>      DF   msTEachRdr   varEachRdr  cov1EachRdr
#> rdr0  1 0.0003970662 0.0006989006 3.734661e-04
#> rdr1  1 0.0010828854 0.0011060528 7.601598e-04
#> rdr2  1 0.0001597470 0.0008423434 3.553224e-04
#> rdr3  1 0.0003444784 0.0001505777 1.083399e-06
#> rdr4  1 0.0050161160 0.0012135668 2.430368e-04
```

7.1.3 Calculate F-statistic

From the previous chapter, the F-statistic is calculated using:

$$F_{ORH} = \frac{MS(T)}{MS(TR) + J\max(Cov_2 - Cov_3, 0)} \tag{7.1}$$

- MS(T) is in `ret$TRanova["T", "MS"]`, whose value is 0.0047962.
- MS(TR) is in `ret$TRanova["TR", "MS"]`, whose value is 5.5103062×10^{-4}.
- The value of `J`, the number of readers, is the length of the second dimension of `dataset02$ratings$NL[1,,1,1]`, which is 5.
- The value of `Cov2` is in `ret$VarCom["Cov2", "Estimates"]`, whose value is 3.4407483×10^{-4}.
- The value of `Cov3` is in `ret$VarCom["Cov3", "Estimates"]`, whose value is 2.3902837×10^{-4}.

Applying Eqn. (7.1) one gets (`den` is the denominator in Eqn. (7.1)):

```
J <- length(dataset02$ratings$NL[1,,1,1])
den <- ret$TRanova["TR", "MS"] + J* max(ret$VarCom["Cov2", "Estimates"] - ret$VarCom["Cov3", "Est
F_ORH <- ret$TRanova["T", "MS"]/den
print(F_ORH)
#> [1] 4.456319
```

7.1.4 Calculate ddf_H

From the previous chapter, the Hillis `ddf` is calculated using:

$$\text{ddf}_H = \frac{[MS(TR) + J\max(Cov_2 - Cov_3, 0)]^2}{\frac{[MS(TR)]^2}{(I-1)(J-1)}} \tag{7.2}$$

The numerator of `ddf` is seen to be identical to `den^2`. Therefore, the implementation is as follows:

```
I <- length(dataset02$ratings$NL[,1,1,1])
ddf <- den^2*(I-1)*(J-1)/(ret$TRanova["TR", "MS"])^2
print(ddf)
#> [1] 15.25967
```

7.1.5 Calculate the p-value

This is the probability that a sample from $F_{I-1,ddf}$ exceeds the observed value of the statistic, `F_ORH` $= 4.4563187$. This is calculated as follows:

```
p <- 1 - pf(F_ORH, I - 1, ddf)
print(p)
#> [1] 0.05166569
```

The difference is not significant at $\alpha = 0.05$.

7.1.6 Calculate confidence intervals for reader-averaged inter-treatment differences

Since `I` $= 2$, their is only one difference in reader-averaged FOMs, namely, the first treatment minus the second:

```
trtMeans <- rowMeans(foms)
trtMeanDiffs <- trtMeans[1] - trtMeans[2]
print(trtMeanDiffs)
#>        trt0
#> -0.04380032
```

From the previous chapter, the $(1-\alpha)$ confidence interval for $\theta_{i\bullet} - \theta_{i'\bullet}$ is given by (the average is over the reader index):

$$CI_{1-\alpha,RRRC} = \theta_{i\bullet} - \theta_{i'\bullet} \pm t_{\alpha/2,ddf_H}\sqrt{\frac{2}{J}(MS(TR) + J\max(Cov_2 - Cov_3, 0))} \tag{7.3}$$

The expression inside the square-root symbol is `2/J*den`. The implementation follows:

```
alpha <- 0.05
stdErr <- sqrt(2 * den/J)
t_crit <- abs(qt(alpha/2, ddf))
CI <- c(trtMeanDiffs - t_crit*stdErr, trtMeanDiffs + t_crit*stdErr)
names(CI) <- c("Lower", "Right")
print(CI)
#>         Lower         Right
#> -0.0879594986  0.0003588544
```

The confidence interval, shown as `[Lower, Right]`, includes zero, which confirms that the reader-averaged FOM difference between treatments is not significant.

7.2 Discussion/Summary/5

7.3 References

Chapter 8

Sample size estimation for ROC studies DBM method

8.1 Introduction

The question addressed here is "how many readers and cases", usually abbreviated to "sample-size", should one employ to conduct a "well-planned" ROC study. The reasons for the quotes around "well-planned" will shortly become clear. If cost were no concern, the reply would be: "as many readers and cases as one can get". There are other causes affecting sample-size, e.g., the data collection paradigm and analysis, however, this chapter is restricted to the MRMC ROC data collection paradigm, with data analyzed by the DBM method described in a previous chapter. The next chapter will deal with data analyzed by the OR method.

It turns out that provided one can specify conceptually valid effect-sizes between different paradigms (i.e., in the same "units"), the methods described in this chapter are extensible to other paradigms; see TBA Chapter 19 for sample size estimation for FROC studies. *For this reason it is important to understand the concepts of sample-size estimation in the simpler ROC context.*

For simplicity and practicality, this chapter, and the next, is restricted to analysis of two-treatment data ($I = 2$). The purpose of most imaging system assessment studies is to determine, for a given diagnostic task, whether radiologists perform better using a new treatment over the conventional treatment, and whether the difference is statistically significant. Therefore, the two-treatment case is the most common one encountered. While it is possible to extend the methods to more than two treatments, the extensions are not, in my opinion, clinically interesting.

Assume the figure of merit (FOM) θ is chosen to be the area AUC under the ROC

curve (empirical or fitted is immaterial as far as the formulae are concerned; however, the choice will affect statistical power). The statistical analysis determines the significance level of the study, i.e., the probability or p-value for incorrectly rejecting the null hypothesis (NH) that the two θs are equal: $NH : \theta_1 = \theta_2$, where the subscripts refer to the two treatments and the bullet represents the average over the reader index. If the p-value is smaller than a pre-specified α, typically set at 5%, one rejects the NH and declares the treatments different at the α significance level. Statistical power is the probability of correctly rejecting the null hypothesis when the alternative hypothesis $AH : \theta_1 \neq \theta_2$ is true, (TBA Chapter 08).

The value of the *true* difference between the treatments, known as the *true effect-size* is, of course, unknown. If it were known, there would be no need to conduct the ROC study. One would simply adopt the treatment with the higher θ. Sample-size estimation involves making an educated guess regarding the true effect-size , called the *anticipated effect size*, and denoted by d. To quote Harold Kundel (ICRU, 1996): "any calculation of power amounts to specification of the anticipated effect-size". Increasing the anticipated effect size will increase statistical power but may represent an unrealistic expectation of the true difference between the treatments, in the sense that it overestimates the ability of technology to achieve this much improvement. Conversely, an unduly small d might be clinically insignificant, besides requiring a very large sample-size to achieve sufficient statistical power.

Statistical power depends on the magnitude of d divided by the standard deviation $\sigma(d)$ of d, i.e. $D = \frac{|d|}{\sigma(d)}$. The sign is relevant as it determines whether the project is worth pursuing at all (see TBA §11.8.4). The ratio is termed (Cohen, 1988) Cohen's D. When this signal-to-noise-ratio-like quantity is large, statistical power approaches 100%. Reader and case variability and data correlations determine $\sigma(d)$. No matter how small the anticipated d, as long as it is finite, then, using sufficiently large numbers of readers and cases $\sigma(d)$ can be made sufficiently small to achieve near 100% statistical power. Of course, a very small effect-size may not be clinically significant. There is a key difference between *statistical significance* and *clinical significance*. An effect-size in AUC units could be so small, e.g., 0.001, as to be clinically insignificant, but by employing a sufficiently large sample size one could design a study to detect this small - and clinically meaningless - difference with near unit probability, i.e., high statistical power.

What determines clinical significance? A small effect-size, e.g., 0.01 AUC units, could be clinically significant if it applies to a large population, where the small benefit in detection rate is amplified by the number of patients benefiting from the new treatment. In contrast, for an "orphan" disease, i.e., one with very low prevalence, an effect-size of 0.05 might not be enough to justify the additional cost of the new treatment. The improvement might have to be 0.1 before it is worth it for a new treatment to be brought to market. One hates to monetize life and death issues, but there is no getting away from it, as cost/benefit issues de-

termine clinical significance. The arbiters of clinical significance are engineers, imaging scientists, clinicians, epidemiologists, insurance companies and those who set government health care policies. The engineers and imaging scientists determine whether the effect-size the clinicians would like is feasible from technical and scientific viewpoints. The clinician determines, based on incidence of disease and other considerations, e.g., altruistic, malpractice, cost of the new device and insurance reimbursement, what effect-size is justifiable. Cohen has suggested that d values of 0.2, 0.5, and 0.8 be considered small, medium, and large, respectively, but he has also argued against their indiscriminate usage. However, after a study is completed, clinicians often find that an effect-size that biostatisticians label as small may, in certain circumstances, be clinically significant and an effect-size that they label as large may in other circumstances be clinically insignificant. Clearly, this is a complex issue. Some suggestions on choosing a clinically significant effect size are made in (TBA §11.12).

Having developed a new imaging modality the R&D team wishes to compare it to the existing standard with the short-term goal of making a submission to the FDA to allow them to perform pre-market testing of the device. The long-term goal is to commercialize the device. Assume the R&D team has optimized the device based on physical measurements, (TBA Chapter 01), perhaps supplemented with anecdotal feedback from clinicians based on a few images. Needed at this point is a pilot study. A pilot study, conducted with a relatively small and practical sample size, is intended to provide estimates of different sources of variability and correlations. It also provides an initial estimate of the effect-size, termed the *observed effect-size*, d. Based on results from the pilot the sample-size tools described in this chapter permit estimation of the numbers of readers and cases that will reduce $\sigma(d)$ sufficiently to achieve the desired power for the larger "pivotal" study. [A distinction could be made in the notation between observed and anticipated effect sizes, but it will be clear from the context. Later, it will be shown how one can make an educated guess about the anticipated effect size from an observed effect size.]

This chapter is concerned with multiple-reader MRMC studies that follow the fully crossed factorial design meaning that each reader interprets a common case-set in all treatments. Since the resulting pairings (i.e., correlations) tend to decrease $\sigma(d)$ (since the variations occur in tandem, they tend to cancel out in the difference, see (TBA Chapter 09, Introduction), for Dr. Robert Wagner's sailboat analogy) it yields more statistical power compared to an unpaired design, and consequently this design is frequently used. Two sample-size estimation procedures for MRMC are the Hillis-Berbaum method (Hillis and Berbaum, 2004) and the Obuchowski-Rockette (Obuchowski, 1998) method. With recent work by Hillis, the two methods have been shown to be substantially equivalent.

This chapter will focus on the DBM approach. Since it is based on a standard ANOVA model, it is easier to extend the NH testing procedure described in Chapter 09 to the alternative hypothesis, which is relevant for sample size estimation. [TBA Online Appendix 11.A shows how to translate the DBM formulae

to the OR method (Hillis et al., 2011).]

Given an effect-size, and choosing this wisely is the most difficult part of the process, the method described in this chapter uses pseudovalue variance components estimated by the DBM method to predict sample-sizes (i.e., different combinations of numbers of readers and cases) necessary to achieve a desired power.

8.2 Statistical Power

The concept of statistical power was introduced in [TBA Chapter 08] but is worth repeating. There are two possible decisions following a test of a null hypothesis (NH): reject or fail to reject the NH. Each decision is associated with a probability on an erroneous conclusion. If the NH is true and one rejects it, the probability of the ensuing Type-I error is denoted α. If the NH is false and one fails to reject it, the probability of the ensuing Type II- error is denoted β. Statistical power is the complement of β, i.e.,

$$Power = 1 - \beta \tag{8.1}$$

Typically, one aims for $\beta = 0.2$ or less, i.e., a statistical power of 80% or more. Like $\alpha = 0.05$, this is a *convention* and more nuanced cost-benefit considerations may cause the researcher to adopt a different value.

8.2.1 Observed vs. anticipated effect-size

Assuming no other similar studies have already been conducted with the treatments in question, the observed effect-size, although "merely an estimate", is the best information available at the end of the pilot study regarding the value of the true effect-size. From the two previous chapters one knows that the significance testing software will report not only the observed effect-size, but also a 95% confidence interval associate with it. It will be shown later how one can use this information to make an educated guess regarding the value of the anticipated effect-size.

8.2.2 Dependence of statistical power on estimates of model parameters

Examination of the expression for , Eqn. (11.5), shows that statistical power increases if: * The numerator is large. This occurs if: (a) the anticipated effect-size d is large. Since effect-size enters as the *square*, TBA Eqn. (11.8), it is has a particularly strong effect; (b) If $J \times K$ is large. Both of these results

should be obvious, as a large effect size and a large sample size should result in increased probability of rejecting the NH. * The denominator is small. The first term in the denominator is $(\sigma_\epsilon^2 + \sigma_{\tau RC}^2)$. These two terms cannot be separated. This is the residual variability of the jackknife pseudovalues. It should make sense that the smaller the variability, the larger is the non-centrality parameter and the statistical power. * The next term in the denominator is $K\sigma_{\tau R}^2$, the treatment-reader variance component multiplied by the total number of cases. The reader variance σ_R^2 has no effect on statistical power, because it has an equal effect on both treatments and cancels out in the difference. Instead, it is the treatment-reader variance σ_R^2 that contributes "noise" tending to confound the estimate of the effect-size. * The variance components estimated by the ANOVA procedure are realizations of random variables and as such subject to noise (there actually exists a beast such as variance of a variance). The presence of the K term, usually large, can amplify the effect of noise in the estimate of σ_R^2, making the sample size estimation procedure less accurate. * The final term in the denominator is $J\sigma_{\tau C}^2$. The variance σ_C^2 has no impact on statistical power, as it cancels out in the difference. The treatment-case variance component introduces "noise" into the estimate of the effect size, thereby decreasing power. Since it is multiplied by J, the number of readers, and typically $J << K$, the error amplification effect on accuracy of the sample size estimate is not as bad as with the treatment-reader variance component. * Accuracy of sample size estimation, essentially estimating confidence intervals for statistical power, is addressed in (Chakraborty, 2010).

8.2.3 Formulae for random-reader random-case (RRRC) sample size estimation

8.2.4 Significance testing

8.2.5 p-value and confidence interval

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

Having performed a pilot study and planning to perform a pivotal study, sample size estimation follows the following procedure, which assumes that both reader and case are treated as random factors. Different formulae, described later, apply when either reader or case is treated as a fixed factor.

- Perform DBM analysis on the pilot data. This yields the observed effect size as well as estimates of all relevant variance components and mean squares appearing in TBA Eqn. (11.5) and Eqn. (11.7).
- This is the difficult but critical part: make an educated guess regarding the effect-size, d, that one is interested in "detecting" (i.e., hoping to reject the

NH with probability $1-\beta$). The author prefers the term "anticipated" effect-size to "true" effect-size (the latter implies knowledge of the true difference between the modalities which, as noted earlier, would obviate the need for a pivotal study).

- Two scenarios are considered below. In the first scenario, the effect-size is assumed equal to that observed in the pilot study, i.e., $d = d_{obs}$.
- In the second, so-called "best-case" scenario, one assumes that the anticipate value of d is the observed value plus two-sigma of the confidence interval, in the correct direction, of course, i.e., $d = |d_{obs}| + 2\sigma$. Here σ is one-fourth the width of the 95% confidence interval for d_{obs}. Anticipating more than 2σ greater than the observed effect-size would be overly optimistic. The width of the CI implies that chances are less than 2.5% that the anticipated value is at or beyond the overly optimistic value. These points will become clearer when example datasets are analyzed below.
- Calculate statistical power using the distribution implied by Eqn. (11.4), to calculate the probability that a random value of the relevant F-statistic will exceed the critical value, as in §11.3.2.
- If power is below the desired or "target" power, one tries successively larger value of J and / or K until the target power is reached.

8.3 Formulae for fixed-reader random-case (FRRC) sample size estimation

It was shown in TBA §9.8.2 that for fixed-reader analysis the non-centrality parameter is defined by:

$$\Delta = \frac{JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2} \tag{8.2}$$

The sampling distribution of the F-statistic under the AH is:

$$F_{AH|R} \equiv \frac{MST}{MSTC} \sim F_{I-1,(I-1)(K-1),\Delta} \tag{8.3}$$

8.3.1 Formulae for random-reader fixed-case (RRFC) sample size estimation

It is shown in TBA §9.9 that for fixed-case analysis the non-centrality parameter is defined by:

$$\Delta = \frac{JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2} \tag{8.4}$$

Under the AH, the test statistic is distributed as a non-central F-distribution as follows:

$$F_{AH|C} \equiv \frac{MST}{MSTR} \sim F_{I-1,(I-1)(J-1),\Delta} \tag{8.5}$$

8.3.2 Fixed-reader random-case (FRRC) analysis TBA

It is a realization of a random variable, so one has some leeway in the choice of anticipated effect size - more on this later. Here J^* and K^* refer to the number of readers and cases in the *pilot* study.

8.3.3 Random-reader fixed-case (RRFC) analysis

8.3.4 Single-treatment multiple-reader analysis

8.4 Discussion/Summary/2

8.5 References

Chapter 9

Sample size estimation for ROC studies OR method

9.1 Introduction

9.2 Statistical Power

$$Power = 1 - \beta \tag{9.1}$$

9.2.1 Sample size estimation for random-reader random-cases

For convenience the OR model is repeated below with the case-set index suppressed:

$$Y_{n(ijk)} = \mu + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + (\tau RC)_{ijk} + \epsilon_{n(ijk)} \tag{9.2}$$

As usual, the treatment effects τ_i are subject to the constraint that they sum to zero. The observed effect size (a random variable) is defined by:

$$d = \theta_{1\bullet} - \theta_{2\bullet} \tag{9.3}$$

It is a realization of a random variable, so one has some leeway in the choice of anticipated effect size. In the significance-testing procedure described in TBA Chapter 09 interest was in the distribution of the F-statistic when the NH is

true. For sample size estimation, one needs to know the distribution of the statistic when the NH is false. It was shown that then the observed F-statistic TBA Eqn. (9.35) is distributed as a non-central F-distribution $F_{ndf,ddf,\Delta}$ with non-centrality parameter Δ:

$$F_{DBM|AH} \sim F_{ndf,ddf,\Delta} \tag{9.4}$$

The non-centrality parameter was defined, Eqn. TBA (9.34), by:

$$\Delta = \frac{JK\sigma^2_{Y;\tau}}{\left(\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC}\right) + K\sigma^2_{Y;\tau R} + J\sigma^2_{Y;\tau C}} \tag{9.5}$$

To minimize confusion, this equation has been rewritten here using the subscript Y to explicitly denote pseudo-value derived quantities (in TBA Chapter 09 this subscript was suppressed.

The estimate of $\sigma^2_{Y;\tau C}$ can turn out to bee negative. To avoid a negative denominator, Hillis suggests the following modification:

$$\Delta = \frac{JK\sigma^2_{Y;\tau}}{\left(\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC}\right) + K\sigma^2_{Y;\tau R} + \max\left(J\sigma^2_{Y;\tau C}, 0\right)} \tag{9.6}$$

This expression depends on three variance components, $(\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC})$ - the two terms are inseparable - $\sigma^2_{Y;\tau R}$ and $\sigma^2_{Y;\tau C}$. The ddf term appearing in TBA Eqn. (11.4) was defined by TBA Eqn. (9.24) - this quantity does not change between NH and AH:

$$ddf_H = \frac{\left[MSTR + \max(MSTR - MSTRC, 0)\right]^2}{\frac{[MSTR]^2}{(I-1)(J-1)}} \tag{9.7}$$

The mean squares in this expression can be expressed in terms of the three variance-components appearing in TBA Eqn. (11.6). Hillis and Berbaum (Hillis and Berbaum, 2004) have derived these expression and they will not be repeated here (Eqn. 4 in the cited reference). RJafroc implements a function to calculate the mean squares, `UtilMeanSquares()`, which allows ddf to be calculated using Eqn. TBA (11.7). The sample size functions in this package need only the three variance-components (the formula for ddf_H is implemented internally).

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

$$\sigma^2_{Y;\tau} = \frac{d^2}{2} \tag{9.8}$$

9.2.2 Dependence of statistical power on estimates of model parameters

Examination of the expression for , Eqn. (11.5), shows that statistical power increases if: * The numerator is large. This occurs if: (a) the anticipated effect-size d is large. Since effect-size enters as the *square*, TBA Eqn. (11.8), it is has a particularly strong effect; (b) If $J \times K$ is large. Both of these results should be obvious, as a large effect size and a large sample size should result in increased probability of rejecting the NH. * The denominator is small. The first term in the denominator is $\left(\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC}\right)$. These two terms cannot be separated. This is the residual variability of the jackknife pseudovalues. It should make sense that the smaller the variability, the larger is the non-centrality parameter and the statistical power. * The next term in the denominator is $K\sigma^2_{Y;\tau R}$, the treatment-reader variance component multiplied by the total number of cases. The reader variance $\sigma^2_{Y;R}$ has no effect on statistical power, because it has an equal effect on both treatments and cancels out in the difference. Instead, it is the treatment-reader variance $\sigma^2_{Y;R}$ that contributes "noise" tending to confound the estimate of the effect-size. * The variance components estimated by the ANOVA procedure are realizations of random variables and as such subject to noise (there actually exists a beast such as variance of a variance). The presence of the K term, usually large, can amplify the effect of noise in the estimate of $\sigma^2_{Y;R}$, making the sample size estimation procedure less accurate. * The final term in the denominator is $J\sigma^2_{Y;\tau C}$. The variance $\sigma^2_{Y;C}$ has no impact on statistical power, as it cancels out in the difference. The treatment-case variance component introduces "noise" into the estimate of the effect size, thereby decreasing power. Since it is multiplied by J, the number of readers, and typically $J << K$, the error amplification effect on accuracy of the sample size estimate is not as bad as with the treatment-reader variance component. * Accuracy of sample size estimation, essentially estimating confidence intervals for statistical power, is addressed in (Chakraborty, 2010).

9.2.3 Formulae for random-reader random-case (RRRC) sample size estimation

9.2.4 Significance testing

9.2.5 p-value and confidence interval

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

Having performed a pilot study and planning to perform a pivotal study, sample size estimation follows the following procedure, which assumes that both reader

and case are treated as random factors. Different formulae, described later, apply when either reader or case is treated as a fixed factor.

- Perform OR analysis on the pilot data. This yields the observed effect size as well as estimates of all relevant variance components and mean squares appearing in TBA Eqn. (11.5) and Eqn. (11.7).
- This is the difficult but critical part: make an educated guess regarding the effect-size, d, that one is interested in "detecting" (i.e., hoping to reject the NH with probability $1-\beta$). The author prefers the term "anticipated" effect-size to "true" effect-size (the latter implies knowledge of the true difference between the modalities which, as noted earlier, would obviate the need for a pivotal study).
- Two scenarios are considered below. In the first scenario, the effect-size is assumed equal to that observed in the pilot study, i.e., $d = d_{obs}$.
- In the second, so-called "best-case" scenario, one assumes that the anticipate value of d is the observed value plus two-sigma of the confidence interval, in the correct direction, of course, i.e., $d = |d_{obs}| + 2\sigma$. Here σ is one-fourth the width of the 95% confidence interval for d_{obs}. Anticipating more than 2σ greater than the observed effect-size would be overly optimistic. The width of the CI implies that chances are less than 2.5% that the anticipated value is at or beyond the overly optimistic value. These points will become clearer when example datasets are analyzed below.
- Calculate statistical power using the distribution implied by Eqn. (11.4), to calculate the probability that a random value of the relevant F-statistic will exceed the critical value, as in §11.3.2.
- If power is below the desired or "target" power, one tries successively larger value of J and / or K until the target power is reached.

9.3 Formulae for fixed-reader random-case (FRRC) sample size estimation

It was shown in TBA §9.8.2 that for fixed-reader analysis the non-centrality parameter is defined by:

$$\Delta = \frac{JK\sigma^2_{Y;\tau}}{\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC} + J\sigma^2_{Y;\tau C}} \tag{9.9}$$

The sampling distribution of the F-statistic under the AH is:

$$F_{AH|R} \equiv \frac{MST}{MSTC} \sim F_{I-1,(I-1)(K-1),\Delta} \tag{9.10}$$

9.3.1 Formulae for random-reader fixed-case (RRFC) sample size estimation

It is shown in TBA §9.9 that for fixed-case analysis the non-centrality parameter is defined by:

$$\Delta = \frac{JK\sigma^2_{Y;\tau}}{\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC} + K\sigma^2_{Y;\tau R}} \tag{9.11}$$

Under the AH, the test statistic is distributed as a non-central F-distribution as follows:

$$F_{AH|C} \equiv \frac{MST}{MSTR} \sim F_{I-1,(I-1)(J-1),\Delta} \tag{9.12}$$

9.3.2 Example 1

In the first example the Van Dyke dataset is regarded as a pilot study. Two implementations are shown, a direct application of the relevant formulae, including usage of the mean squares, which in principle can be calculated from the three variance-components. This is then compared to the `RJafroc` implementation.

Shown first is the "open" implementation.

```
alpha <- 0.05;cat("alpha = ", alpha, "\n")
#> alpha =  0.05
rocData <- dataset02 # select Van Dyke dataset
retDbm <- StSignificanceTesting(dataset = rocData, FOM = "Wilcoxon", method = "DBM")
varYTR <- retDbm$ANOVA$VarCom["VarTR","Estimates"]
varYTC <- retDbm$ANOVA$VarCom["VarTC","Estimates"]
varYEps <- retDbm$ANOVA$VarCom["VarErr","Estimates"]
effectSize <- retDbm$FOMs$trtMeanDiffs["trt0-trt1","Estimate"]
cat("effect size = ", effectSize, "\n")
#> effect size =  -0.043800322

#RRRC
J <- 10; K <- 163
ncp <- (0.5*J*K*(effectSize)^2)/(K*varYTR+max(J*varYTC,0)+varYEps)
MS <- UtilMeanSquares(rocData, FOM = "Wilcoxon", method = "DBM")
ddf <- (MS$msTR+max(MS$msTC-MS$msTRC,0))^2/(MS$msTR^2)*(J-1)
FCrit <- qf(1 - alpha, 1, ddf)
Power <- 1-pf(FCrit, 1, ddf, ncp = ncp)
data.frame("J"= J,  "K" = K, "FCrit" = FCrit, "ddf" = ddf, "ncp" = ncp, "RRRCPower" = Power)
#>    J   K     FCrit       ddf       ncp  RRRCPower
```

```
#> 1 10 163 4.1270572 34.334268 8.1269825 0.79111255

#FRRC
J <- 10; K <- 133
ncp <- (0.5*J*K*(effectSize)^2)/(max(J*varYTC,0)+varYEps)
ddf <- (K-1)
FCrit <- qf(1 - alpha, 1, ddf)
Power <- 1-pf(FCrit, 1, ddf, ncp = ncp)
data.frame("J"= J,  "K" = K, "FCrit" = FCrit, "ddf" = ddf, "ncp" = ncp, "RRRCPower" = P
#>    J   K    FCrit ddf       ncp  RRRCPower
#> 1 10 133 3.912875 132 7.9873835 0.80111671

#RRFC
J <- 10; K <- 53
ncp <- (0.5*J*K*(effectSize)^2)/(K*varYTR+varYEps)
ddf <- (J-1)
FCrit <- qf(1 - alpha, 1, ddf)
Power <- 1-pf(FCrit, 1, ddf, ncp = ncp)
data.frame("J"= J,  "K" = K, "FCrit" = FCrit, "ddf" = ddf, "ncp" = ncp, "RRRCPower" = P
#>    J  K    FCrit ddf       ncp  RRRCPower
#> 1 10 53 5.117355   9 10.048716 0.80496663
```

For 10 readers, the numbers of cases needed for 80% power is largest (163) for RRRC and least for RRFC (53). For all three analyses, the expectation of 80% power is met - the numbers of cases and readers were chosen to achieve close to 80% statistical power. Intermediate quantities such as the critical value of the F-statistic, `ddf` and `ncp` are shown. The reader should confirm that the code does in fact implement the relevant formulae. Shown next is the `RJafroc` implementation. The relevant file is mainSsDbm.R, a listing of which follows:

9.3.3 Fixed-reader random-case (FRRC) analysis

9.3.4 Random-reader fixed-case (RRFC) analysis

9.3.5 Single-treatment multiple-reader analysis

9.4 Discussion/Summary/3

9.5 References

Chapter 10

Split Plot Study Design

10.1 References

Bibliography

Bunch, P., Hamilton, J., Sanderson, G., and Simmons, A. (1977). Free response approach to measurement and characterization of radiographic observer performance. In Gray, J. E. and Hendee, W. R., editors, *Application of Optical Instrumentation in Medicine VI*, volume 0127, pages 124 – 135. International Society for Optics and Photonics, SPIE.

Chakraborty, D., Breatnach, E., Yester, M., Soto, B., Barnes, G., and Fraser, R. (1986). Digital and conventional chest imaging: A modified ROC study of observer performance using simulated nodules. *Radiology*, 158:35–39.

Chakraborty, D. P. (2010). Prediction accuracy of a sample-size estimation method for ROC studies. *Academic radiology*, 17:628–638.

Chakraborty, D. P. (2017). *Observer Performance Methods for Diagnostic Imaging - Foundations, Modeling, and Applications with R-Based Examples*. CRC Press, Boca Raton, FL.

Clarkson, E., Kupinski, M. A., and Barrett, H. H. (2006). A probabilistic model for the MRMC method, part 1: Theoretical development. *Academic Radiology*, 13(11):1410–1421.

Cohen, J. (1988). Statistical power analysis for the social sciences. 1988. *Google Scholar*.

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.

Dorfman, D. D., Berbaum, K. S., and Lenth, R. V. (1995). Multireader, multicase receiver operating characteristic methodology: A bootstrap analysis. *Academic Radiology*, 2(7):626–633.

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

Gallas, B. D., Pennello, G. a., and Myers, K. J. (2007). Multireader multicase variance analysis for binary data. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 24(12):70–80.

Hajian-Tilaki, K. O., Hanley, J. A., Joseph, L., and Collet, J. P. (1997). Extension of receiver operating characteristic analysis to data concerning multiple signal detection tasks. *Acad Radiol*, 4:222–229.

Hillis, S., Obuchowski, N., Schartz, K., and Berbaum, K. (2005). A comparison of the dorfman-berbaum-metz and obuchowski-rockette methods for receiver operating characteristic (ROC) data. *Statistics in Medicine*, 24(10):1579–1607.

Hillis, S. L. (2007). A comparison of denominator degrees of freedom methods for multiple observer ROC studies. *Statistics in Medicine*, 26:596–619.

Hillis, S. L. (2014). A marginal-mean ANOVA approach for analyzing multireader multicase radiological imaging data. *Statistics in Medicine*, 33(2):330–360.

Hillis, S. L., Berbaum, K., and Metz, C. (2008). Recent developments in the dorfman-berbaum-metz procedure for multireader ROC study analysis. *Acad Radiol*, 15(5):647–661.

Hillis, S. L. and Berbaum, K. S. (2004). Power estimation for the dorfman-berbaum-metz method. *Acad. Radiol.*, 11(11):1260–1273.

Hillis, S. L., Obuchowski, N. A., and Berbaum, K. S. (2011). Power estimation for multireader ROC methods: An updated and unified approach. *Academic Radiology*, 18(2):129–142.

ICRU (1996). Medical imaging: the assessment of image quality. *JOURNAL OF THE ICRU*, 54(1):37–40.

Ishwaran, H. and Gatsonis, C. A. (2000). A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *The Canadian Journal of Statistics*, 28(4):731–750.

Kupinski, M. A., Clarkson, E., and Barrett, H. H. (2006). A probabilistic model for the MRMC method, part 2: Validation and applications. *Academic Radiology*, 13(11):1422–1430.

Larsen, R. J. and Marx, M. L. (2001). *An Introduction to Mathematical Statistics and Its Applications*. Prentice-Hall Inc, Upper Saddle River, NJ, 3rd edition.

Metz, C. E., Herman, B. A., and Roe, C. E. (1998). Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets. *Med Decis Making*, 18(1):110–121.

Niklason, L. T., Hickey, N. M., Chakraborty, D. P., Sabbagh, E. A., Yester, M. V., Fraser, R. G., and Barnes, G. T. (1986). Simulated pulmonary nodules: detection with dual-energy digital versus conventional radiography. *Radiology*, 160:589–593.

Obuchowski, N. (2009). Reducing the number of reader interpretations in MRMC studies. *Acad Radiol*, 16:209–217.

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.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, 6(5):309–316.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Series in Cognition and Perception. Academic Press, New York, first edition.

Toledano, A. and Gatsonis, C. (1996). Ordinal regression methodology for ROC curves derived from correlated data. *Stat Med*, 15(16):1807–1826.

Toledano, A. Y. (2003). Three methods for analyzing correlated ROC curves: A comparison in real data sets. *Statistics in Medicine*, 22(18):2919–33.

Van Dyke, C., White, R., Obuchowski, N., Geisinger, M., Lorig, R., and Meziane, M. (1993). Cine mri in the diagnosis of thoracic aortic dissection. *79th RSNA Meetings*.

Zhou, X.-H., McClish, D. K., and Obuchowski, N. A. (2009). *Statistical methods in diagnostic medicine*, volume 569. John Wiley & Sons.