The RJafroc Quick Start Book

Dev P. Chakraborty, PhD

2021-12-10

Contents

Preface

- This book is currently (as of November 2021) 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.
- The physical book chapters are referred to as *book-chapters*, to distinguish them from the chapters in this online book.

TBA How much finished

10%

The pdf file of the book

Go here and then click on `Download` to get the `RJafrocQuickStart.pdf` file.

The html version of the book

Go here to view the `html` version of the book.

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`.

Structure of the book

The book is divided into parts as follows:

- Part I: Quick Start: intended for existing Windows `JAFROC` users who are seeking a quick-and-easy transition from Windows `JAFROC` to `RJafroc`.
- Part II: ROC paradigm: this covers the basics of the ROC paradigm
- Part III: Significance Testing: The general procedure used to determine the significance level, and associated statistics, of the observed difference in figure of merit between pairs of treatments or readers
- Part IV: FROC paradigm: TBA

Contributing to this book

I appreciate constructive feedback on this document. To do this raise an `Issue` on the `GitHub` interface. Click on the `Issues` tab under `dpc10ster/RJafrocQuickStart`, 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 classification
- Optimization of Artificial Intelligence (AI) algorithms

- CV
- Alternatives

ToDos TBA

- Check Bamber theorem derivation.
- Parts labeled TBA and TODOLAST need to be updated on final revision.
- Change third person to first person in references to myself.

Chapters needing heavy edits

- 12-froc.
- 13-froc-empirical.
- 13-froc-empirical-examples.

Shelved vs. removed vs. parked folders needing heavy edits

- replace functions with ; eg. erf and exp in all of document
- Also for TPF, FPF etc.
- Temporarily shelved 17c-rsm-evidence.Rmd in removed folder
- Now 17-b is breaking; possibly related to changes in RJafroc: had to do with recent changes to RJafroc code - RSM_xFROC etc requiring intrinsic parameters; fixed 17-b
- parked has dependence of ROC/FROC performance on threshold

Coding aids

- sprintf("%.4f", proper formatting of numbers
- OpPtStr(, do:
- kbl(dfA, caption = "...", booktabs = TRUE, escape = FALSE) %>% collapse_rows(columns = c(1, 3), valign = "middle") %>% kable_styling(latex_options = c("basic", "scale_down", "HOLD_position"), row_label_position = "c")
- "'{r, attr.source = ".numberLines"}
- kbl(x12, caption = "Summary of optimization results using wAFROC-AUC.", booktabs = TRUE, escape = FALSE) %>% collapse_rows(columns = c(1), valign = "middle") %>% kable_styling(latex_options = c("basic", "scale_down", "HOLD_position"), row_label_position = "c")

- $\exp(-\lambda')$ space before dollar sign generates a pdf error
- FP errors generated by GitHub actions due to undefined labels: Error: Error: pandoc version 1.12.3 or higher is required and was not found (see the help page ?rmarkdown::pandoc_available). In addition: Warning message: In verify_rstudio_version() : Please install or upgrade Pandoc to at least version 1.17.2; or if you are using RStudio, you can just install RStudio 1.0+. Execution halted

Quick Start

Chapter 1

Help

1.1 TBA How much finished

40% (need to add images for one reader; add one-modality dataset)

1.2 Getting help on the software

- If you have installed `RJafroc` from `GitHub`:
  - ?`RJafroc-package` (RStudio will auto complete …)
  - Scroll down all the way and click on `Index`
- Regardless of where you installed from use the `RJafroc` help site:
  - RJafroc help site
  - Look under `References`
  - For example, for help on the function `PlotEmpiricalOperatingCharacteristics`:
  - PlotEmpiricalOperatingCharacteristics

1.3 References

Chapter 2

JAFROC ROC data format

2.1 TBA How much finished

80% (need to add images for one reader; add one-modality dataset)

2.2 Introduction

- JAFROC data format is named after the file format adopted circa. 2006 for the input Excel file to Windows JAFROC software.
- The purpose of this chapter is to explain the data format of this file.
- Reading this file into a dataset object suitable for `RJafroc` analysis is the subject of the next chapter.
- Background on observer performance methods are in my book (Chakraborty, 2017).
- I will start with Receiver Operating Characteristic (ROC) data (Metz, 1978) as this is by far the simplest paradigm.
- In the ROC paradigm the observer assigns a rating to each image. A rating is an ordered numeric label, and, in our convention, higher values represent greater certainty or **confidence level** for presence of disease. With human observers, a 5 (or 6) point rating scale is typically used, with 1 representing highest confidence for *absence* of disease and 5 (or 6) representing highest confidence for *presence* of disease. Intermediate values represent intermediate confidence levels for presence or absence of disease.
- Note that location information, if applicable, associated with the disease, is not collected.
- There is no restriction to 5 or 6 ratings. With algorithmic observers, e.g., computer aided detection (CAD) algorithms, the rating could be a

floating point number and have infinite precision. All that is required is that higher values correspond to greater confidence in presence of disease.
- The above is termed a *positive-directed* rating scale. If lower numbers correspond to greater confidence, termed a negative-directed rating scale, a simple transformation to $max(rating) - rating + 1$, where $max(rating)$ is the maximum rating, over all readers, modalities and cases, will convert a negative-directed rating scale to a positive directed rating scale.

2.3 Note to existing users

- The Excel file format has recently undergone changes, involving three additional columns in the `Truth` worksheet.
- `RJafroc` will work with old format Excel files as the additional columns are ignored.
- Reasons for the change will become clearer in later chapters [1].

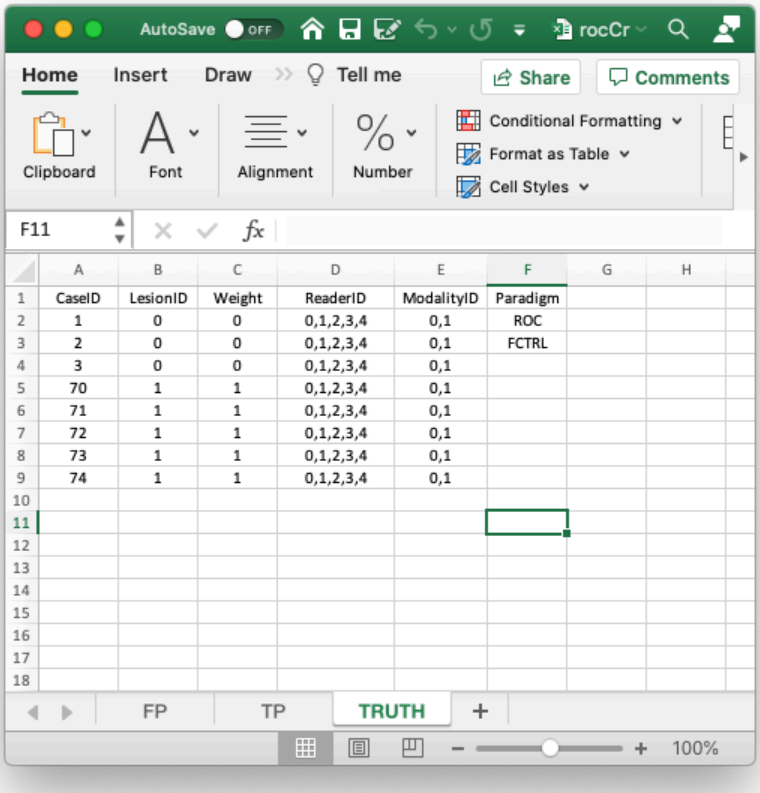
2.4 Contents of Excel file

- The illustrations in this chapter correspond to Excel file `R/quick-start/rocCr.xlsx` in the project directory [2]. This is termed a *toy file*, i.e., an artificial small dataset created to illustrate essential features of the data format.
- The Excel file has three worksheets: `Truth`, `NL` (or `FP`) and `LL` (or `TP`).

[1] They are needed for generalization to other data collection paradigms and for better data entry error control

[2] To access files one needs to `fork` the repository, which creates, on your computer, a copy of all files used to create this document

2.5 The Truth worksheet



	A	B	C	D	E	F
1	CaseID	LesionID	Weight	ReaderID	ModalityID	Paradigm
2	1	0	0	0,1,2,3,4	0,1	ROC
3	2	0	0	0,1,2,3,4	0,1	FCTRL
4	3	0	0	0,1,2,3,4	0,1	
5	70	1	1	0,1,2,3,4	0,1	
6	71	1	1	0,1,2,3,4	0,1	
7	72	1	1	0,1,2,3,4	0,1	
8	73	1	1	0,1,2,3,4	0,1	
9	74	1	1	0,1,2,3,4	0,1	

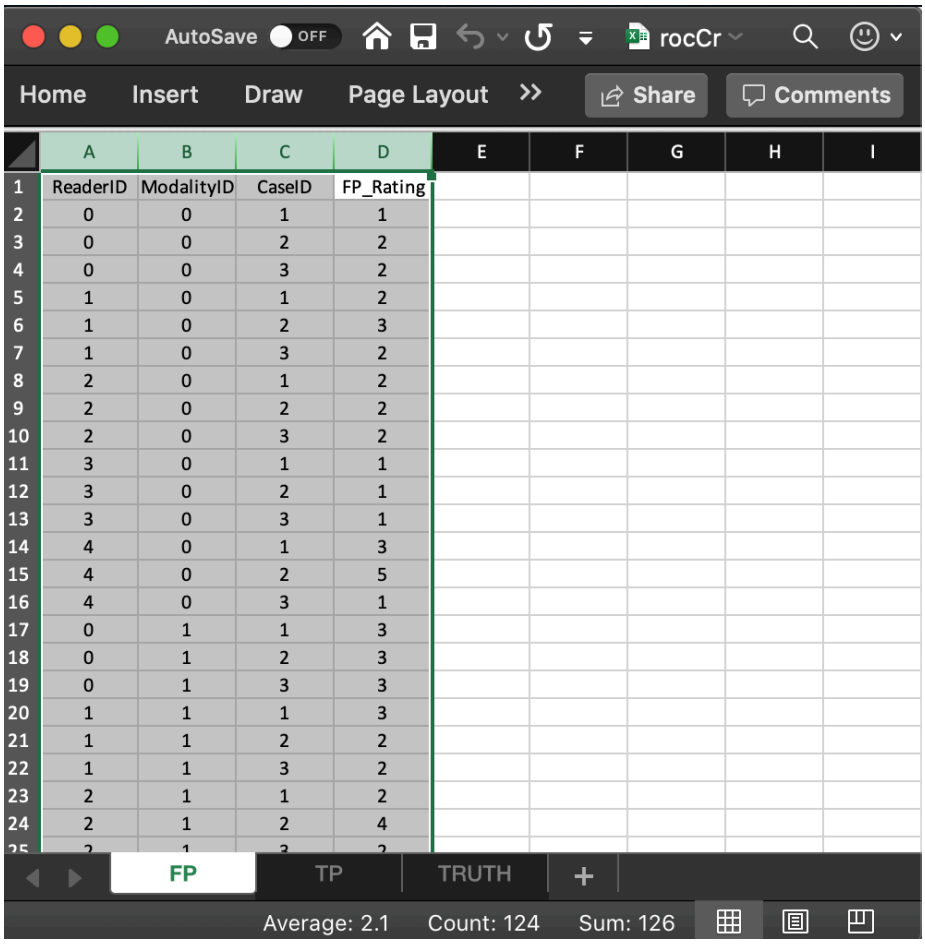
- The `Truth` worksheet contains 6 columns: `CaseID`, `LesionID`, `Weight`, `ReaderID`, `ModalityID` and `Paradigm`.
- The first five columns contain as many rows as there are cases (images) in the dataset.
- `CaseID`: **unique integers**, one per case, representing the cases in the dataset.
- `LesionID`: integers 0 or 1, with each 0 representing a non-diseased case and each 1 representing a diseased case.
- In the current dataset, the non-diseased cases are labeled `1`, `2` and `3`, while the diseased cases are labeled `70`, `71`, `72`, `73` and `74`. The values do not have to be consecutive integers; they need not be ordered; the only requirement is that they be **unique integers**.
- `Weight`: A floating point value, typically filled in with 0 or 1; this field is

not used for ROC data.

- `ReaderID`: a **comma-separated** listing of reader labels, each represented by a **unique integer**, that have interpreted the case. In the example shown below each cell has the value `0, 1, 2, 3, 4` meaning that each of these readers has interpreted all cases (hence the "factorial" design).
  - **With multiple readers each cell in this column has to be text formatted as otherwise Excel will not accept it.**
  - Select the worksheet, then `Format` - `Cells` - `Number` - `Text` - `OK`.
- `ModalityID`: a comma-separated listing of modalities, each represented by a **unique integer**, that are applied to each case. In the example each cell has the value `0, 1`.
  - **With multiple modalities each cell has to be text formatted as otherwise Excel will not accept it.**
  - Format the cells as described above.
- `Paradigm`: this column contains two cells, `ROC` and `factorial`. It informs the software that this is an ROC dataset, and the design is factorial, meaning each reader has interpreted each case in each modality.
- There are 5 diseased cases in the dataset (the number of 1's in the `LesionID` column of the `Truth` worksheet).
- There are 3 non-diseased cases in the dataset (the number of 0's in the `LesionID` column).
- There are 5 readers in the dataset (each cell in the `ReaderID` column contains the string `0, 1, 2, 3, 4`).
- There are 2 modalities in the dataset (each cell in the `ModalityID` column contains the string `0, 1`).

2.6 The false positive (FP) ratings

These are found in the `FP` or `NL` worksheet, see below.

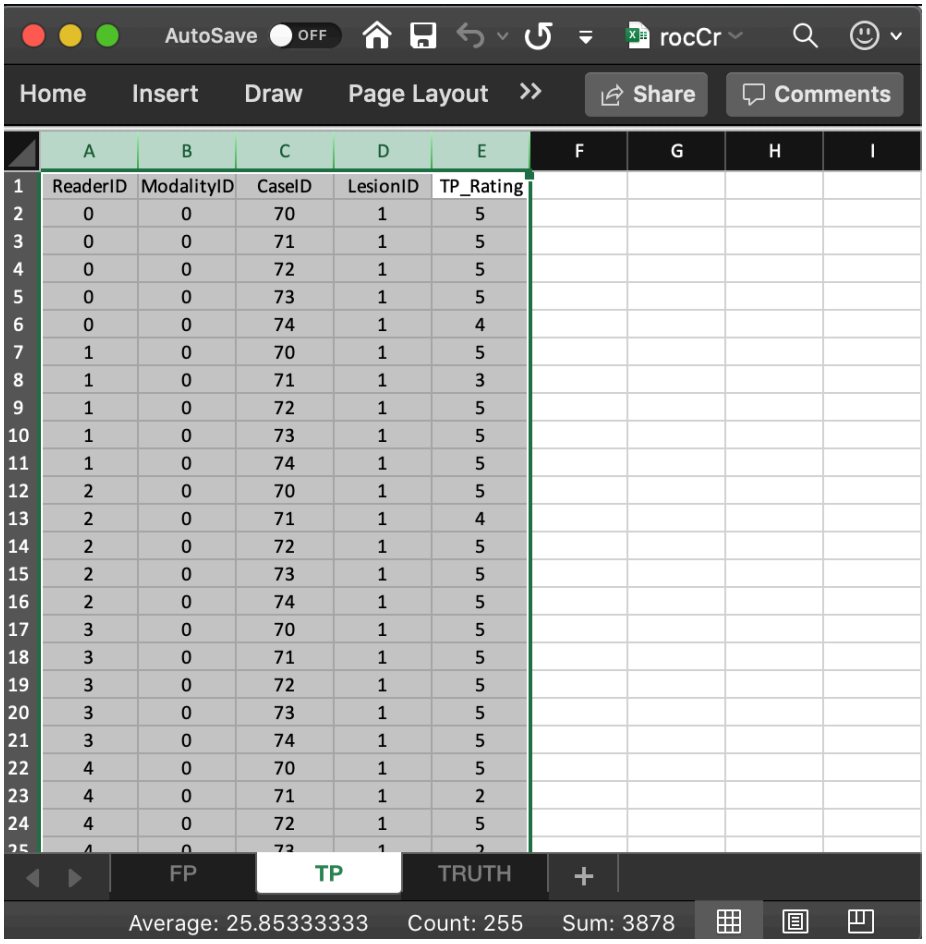


	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	FP_Rating					
2	0	0	1	1					
3	0	0	2	2					
4	0	0	3	2					
5	1	0	1	2					
6	1	0	2	3					
7	1	0	3	2					
8	2	0	1	2					
9	2	0	2	2					
10	2	0	3	2					
11	3	0	1	1					
12	3	0	2	1					
13	3	0	3	1					
14	4	0	1	3					
15	4	0	2	5					
16	4	0	3	1					
17	0	1	1	3					
18	0	1	2	3					
19	0	1	3	3					
20	1	1	1	3					
21	1	1	2	2					
22	1	1	3	2					
23	2	1	1	2					
24	2	1	2	4					

- It consists of 4 columns, each of length 30 (# of modalities X number of readers X number of non-diseased cases).
- `ReaderID`: the reader labels: `0`, `1`, `2`, `3` and `4`. Each reader label occurs 6 times (# of modalities X number of non-diseased cases).
- `ModalityID`: the modality or treatment labels: `0` and `1`. Each label occurs 15 times (# of readers X number of non-diseased cases).
- `CaseID`: the case labels for non-diseased cases: `1`, `2` and `3`. Each label occurs 10 times (# of modalities X # of readers).
- The label of a diseased case cannot occur in the FP worksheet. If it does the software generates an error.
- `FP_Rating`: the floating point ratings of non-diseased cases. Each row of this worksheet contains a rating corresponding to the values of `ReaderID`, `ModalityID` and `CaseID` for that row.

2.7 The true positive (TP) ratings

These are found in the `TP` or `LL` worksheet, see below.



	A	B	C	D	E
1	ReaderID	ModalityID	CaseID	LesionID	TP_Rating
2	0	0	70	1	5
3	0	0	71	1	5
4	0	0	72	1	5
5	0	0	73	1	5
6	0	0	74	1	4
7	1	0	70	1	5
8	1	0	71	1	3
9	1	0	72	1	5
10	1	0	73	1	5
11	1	0	74	1	5
12	2	0	70	1	5
13	2	0	71	1	4
14	2	0	72	1	5
15	2	0	73	1	5
16	2	0	74	1	5
17	3	0	70	1	5
18	3	0	71	1	5
19	3	0	72	1	5
20	3	0	73	1	5
21	3	0	74	1	5
22	4	0	70	1	5
23	4	0	71	1	2
24	4	0	72	1	5

- It consists of 5 columns, each of length 50 (# of modalities X number of readers X number of diseased cases).
- `ReaderID`: the reader labels: `0`, `1`, `2`, `3` and `4`. Each reader label occurs 10 times (# of modalities X number of diseased cases).
- `ModalityID`: the modality or treatment labels: `0` and `1`. Each label occurs 25 times (# of readers X number of diseased cases).
- `LesionID`: For an ROC dataset this column contains fifty 1's (each diseased case has one lesion).
- `CaseID`: the case labels for non-diseased cases: `70`, `71`, `72`, `73` and `74`. Each label occurs 10 times (# of modalities X # of readers). For an ROC dataset the label of a non-diseased case cannot occur in the TP worksheet.

If it does the software generates an error.

- `TP_Rating`: the floating point ratings of diseased cases. Each row of this worksheet contains a rating corresponding to the values of `ReaderID`, `ModalityID`, `LesionID` and `CaseID` for that row.

2.8 A single reader dataset

```
rocCr1R <- "R/quick-start/rocCr1R.xlsx"
x <- DfReadDataFile(rocCr1R, newExcelFileFormat = TRUE)
str(x)
#> List of 3
#>  $ ratings     :List of 3
#>   ..$ NL   : num [1:2, 1, 1:8, 1] 2 3 3 2 2 ...
#>   ..$ LL   : num [1:2, 1, 1:5, 1] 5 5 3 3 5 5 5 5 5 5
#>   ..$ LL_IL: logi NA
#>  $ lesions     :List of 3
#>   ..$ perCase: int [1:5] 1 1 1 1 1
#>   ..$ IDs    : num [1:5, 1] 1 1 1 1 1
#>   ..$ weights: num [1:5, 1] 1 1 1 1 1
#>  $ descriptions:List of 7
#>   ..$ fileName     : chr "rocCr1R"
#>   ..$ type         : chr "ROC"
#>   ..$ name         : logi NA
#>   ..$ truthTableStr: num [1:2, 1, 1:8, 1:2] 1 1 1 1 1 1 NA NA NA NA ...
#>   ..$ design       : chr "FCTRL"
#>   ..$ modalityID   : Named chr [1:2] "0" "1"
#>   .. ..- attr(*, "names")= chr [1:2] "0" "1"
#>   ..$ readerID     : Named chr "1"
#>   .. ..- attr(*, "names")= chr "1"
```

2.9 References

Chapter 3

Reading the Excel data file

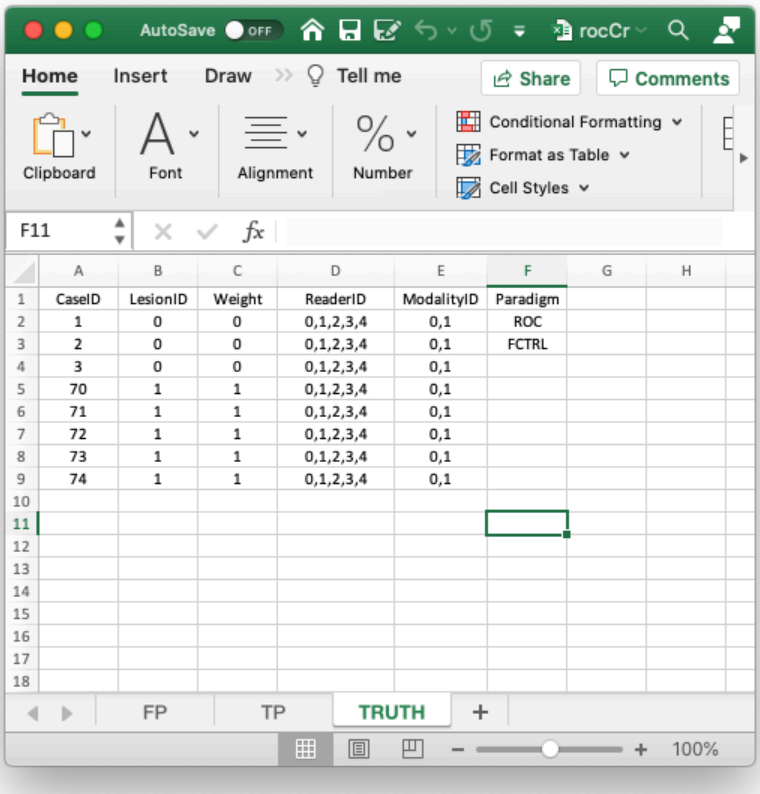
3.1 TBA How much finished

90%

3.2 Introduction

In the previous chapter I described the format of the Excel file `R/quick-start/rocCr.xlsx` corresponding to a small factorial ROC dataset. Described here is how to read this file in order to create an `RJafroc` dataset. It introduces the `RJafroc` function `DfReadDataFile()`. Also shown are the correspondences between values in the Excel file and the dataset object.

3.3 The structure of an ROC dataset



In the following code chunk the second statement reads the Excel file using the function `DfReadDataFile()` and saves it to object `x`. The third statement shows the structure of `x`.

```
rocCr <- "R/quick-start/rocCr.xlsx"
x <- DfReadDataFile(rocCr, newExcelFileFormat = TRUE)
str(x)
#> List of 3
#>  $ ratings     :List of 3
#>   ..$ NL   : num [1:2, 1:5, 1:8, 1] 1 3 2 3 2 2 1 2 3 2 ...
#>   ..$ LL   : num [1:2, 1:5, 1:5, 1] 5 5 5 5 5 5 5 5 5 5 ...
#>   ..$ LL_IL: logi NA
#>  $ lesions     :List of 3
#>   ..$ perCase: int [1:5] 1 1 1 1 1
```

```
#>   ..$ IDs    : num [1:5, 1] 1 1 1 1 1
#>   ..$ weights: num [1:5, 1] 1 1 1 1 1
#>  $ descriptions:List of 7
#>   ..$ fileName     : chr "rocCr"
#>   ..$ type         : chr "ROC"
#>   ..$ name         : logi NA
#>   ..$ truthTableStr: num [1:2, 1:5, 1:8, 1:2] 1 1 1 1 1 1 1 1 1 1 ...
#>   ..$ design       : chr "FCTRL"
#>   ..$ modalityID   : Named chr [1:2] "0" "1"
#>   .. ..- attr(*, "names")= chr [1:2] "0" "1"
#>   ..$ readerID     : Named chr [1:5] "0" "1" "2" "3" ...
#>   .. ..- attr(*, "names")= chr [1:5] "0" "1" "2" "3" ...
```

- In the above code chunk flag `newExcelFileFormat` is set to `TRUE` as otherwise columns D - F in the `Truth` worksheet are ignored and the dataset is assumed to be factorial, with `dataType` "automatically" determined from the contents of the FP and TP worksheets. [1]
- Flag `newExcelFileFormat = FALSE`, the default, is for compatibility with older JAFROC format Excel files, which did not have columns D - F in the `Truth` worksheet. Its usage is deprecated.
- The dataset object `x` is a `list` variable with 3 members: `ratings`, `lesions` and `descriptions`.
- The `x$ratings` member contains 3 sub-lists.
  - The `x$ratings$NL` member, with dimension [2, 5, 8, 1], contains the ratings of normal cases. The first dimension (2) is the number of treatments, the second (5) is the number of readers and the third (8) is the total number of cases. For ROC datasets the fourth dimension is always unity. The five extra values [2] in the third dimension, which are filled with `NAs`, are needed for compatibility with FROC datasets.
  - The `x$ratings$LL`, with dimension [2, 5, 5, 1], contains the ratings of abnormal cases. The third dimension (5) corresponds to the 5 diseased cases.
  - The `x$ratings$LL_IL` member, equal to NA'; this member is there for compatibility with LROC data, `_IL` denotes incorrect-localizations.
- The `x$lesions` member contains 3 sub-lists.
  - The `x$lesions$perCase` member is a vector with 5 ones representing the 5 diseased cases in the dataset.
  - The `x$lesions$IDs` member is an array with 5 ones.
  - The `x$lesions$weights` member is an array with 5 ones.

[1] The assumptions underlying the "automatic" determination could be defeated by data entry errors.

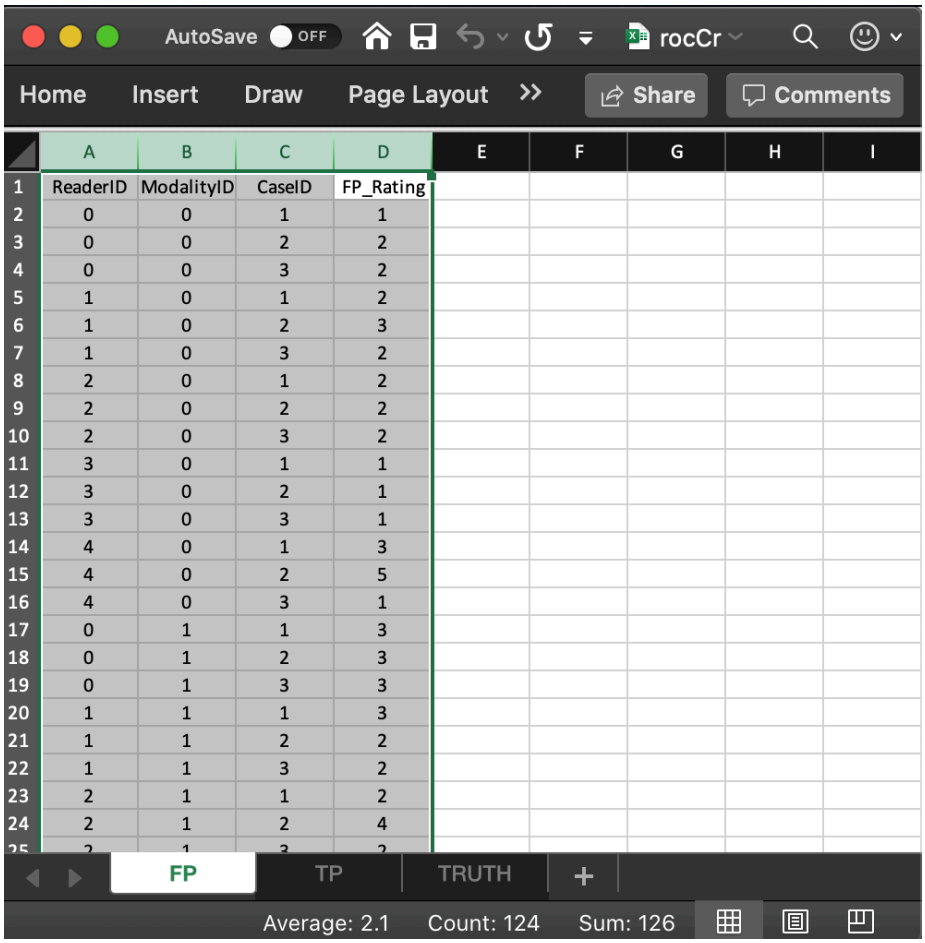
[2] with only 3 non-diseased cases why does one need 8 values?

  - These are irrelevant for ROC datasets. They are there for compatibility with FROC datasets.

- The `x$descriptions` member contains 7 sub-lists.

  - The `x$descriptions$fileName` member is the base name of the file that was read to create this dataset, "rocCr" in the current example, otherwise it is `NA` (the latter would apply, for example, for a simulated dataset).
  - The `x$descriptions$type` member indicates that this is an `ROC` dataset.
  - The `x$descriptions$name` member is the name of this dataset, if it is an embedded dataset, otherwise `NA`.
  - The `x$descriptions$truthTableStr` member, with dimension [2, 5, 8, 2], quantifies the structure of the dataset, as explained in TBA Vignette #3 (it is used to check for data entry errors).
  - The `x$descriptions$design` member specifies the dataset design, which is "FCTRL" in the present example (a factorial dataset).
  - The `x$descriptions$modalityID` member is a vector with two elements `"0"` and `"1"`, naming the two modalities.
  - The `x$readerID` member is a vector with five elements `"0"`, `"1"`, `"2"`, `"3"` and `"4"`, naming the five readers.

3.4 Correspondence between NL member of dataset and the FP worksheet



	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	FP_Rating					
2	0	0	1	1					
3	0	0	2	2					
4	0	0	3	2					
5	1	0	1	2					
6	1	0	2	3					
7	1	0	3	2					
8	2	0	1	2					
9	2	0	2	2					
10	2	0	3	2					
11	3	0	1	1					
12	3	0	2	1					
13	3	0	3	1					
14	4	0	1	3					
15	4	0	2	5					
16	4	0	3	1					
17	0	1	1	3					
18	0	1	2	3					
19	0	1	3	3					
20	1	1	1	3					
21	1	1	2	2					
22	1	1	3	2					
23	2	1	1	2					
24	2	1	2	4					

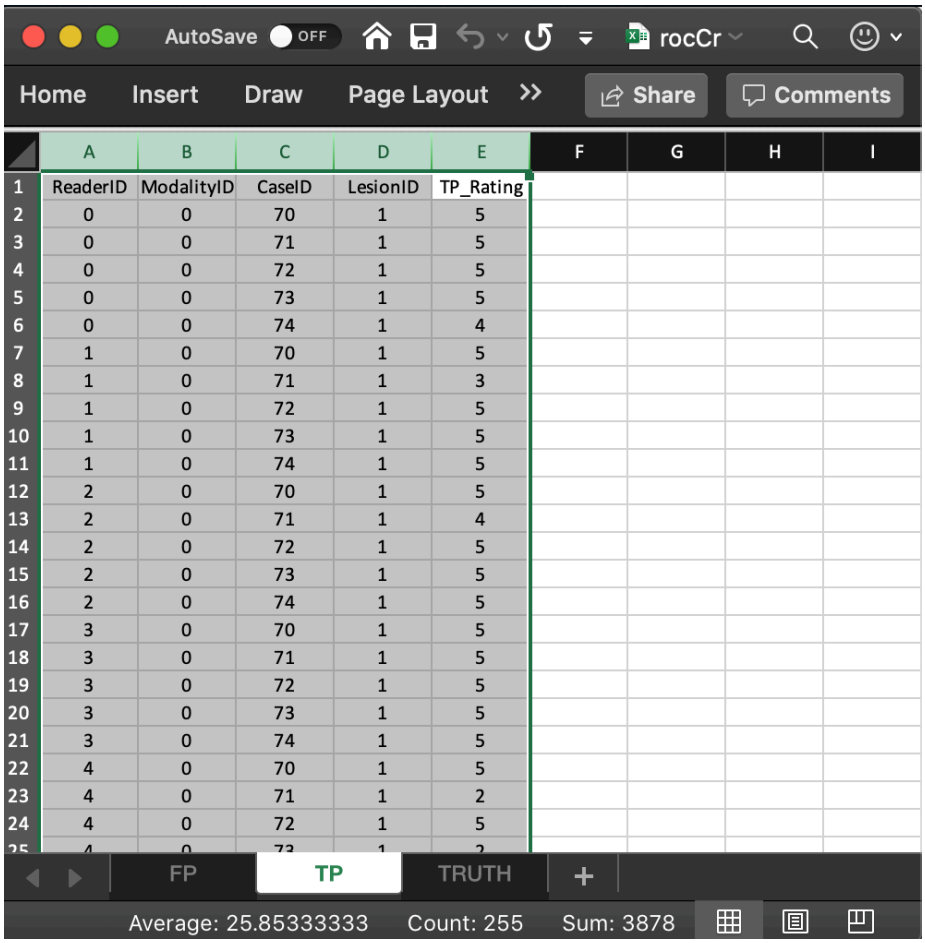
- The list member `x$ratings$NL` is an array with `dim = c(2,5,8,1)`.
  - The first dimension (2) comes from the number of modalities.
  - The second dimension (5) comes from the number of readers.
  - The third dimension (8) comes from the **total** number of cases.
  - The fourth dimension is alway 1 for an ROC dataset.
- The value of `x$ratings$NL[1,5,2,1]`, i.e., 5, corresponds to row 15 of the FP table, i.e., to `ModalityID` = 0, `ReaderID` = 4 and `CaseID` = 2.
- The value of `x$ratings$NL[2,3,2,1]`, i.e., 4, corresponds to row 24 of the FP table, i.e., to `ModalityID` 1, `ReaderID` 2 and `CaseID` 2.
- All values for case index $>$ 3 and case index $<=$ 8 are `-Inf`. For example the value of `x$ratings$NL[2,3,4,1]` is `-Inf`. This is because there are

only 3 non-diseased cases. The extra length is needed for compatibility with FROC datasets.

3.5 Case-index vs. caseID

- Regardless of what order they occur in the worksheet, the non-diseased cases are always indexed first. In the current example the case indices are 1, 2 and 3, corresponding to the three non-diseased cases with `caseIDs` equal to 1, 2 and 3.
- Regardless of what order they occur in the worksheet, in the NL array the diseased cases are always indexed after the last non-diseased case. In the current example the case indices in the `NL` array are 4, 5, 6, 7 and 8, corresponding to the five diseased cases with `caseIDs` equal to 70, 71, 72, 73, and 74. In the `LL` array they are numbered 1, 2, 3, 4 and 5, corresponding to the five diseased cases with `caseIDs` equal to 70, 71, 72, 73, and 74. Some examples follow:
- `x$ratings$NL[1,3,2,1]`, a FP rating, refers to `ModalityID` 0, `ReaderID` 2 and `CaseID` 2 (since the modality and reader IDs start with 0).
- `x$ratings$NL[2,5,4,1]`, a FP rating, refers to `ModalityID` 1, `ReaderID` 4 and `CaseID` 70, the first diseased case; this is `-Inf`.
- `x$ratings$NL[1,4,8,1]`, a FP rating, refers to `ModalityID` 0, `ReaderID` 3 and `CaseID` 74, the last diseased case; this is `-Inf`.
- `x$ratings$NL[1,3,9,1]`, a FP rating, is an illegal value, as the third index cannot exceed 8.
- `x$ratings$NL[1,3,8,2]`, a FP rating, is an illegal value, as the fourth index cannot exceed 1 for an ROC dataset.
- `x$ratings$LL[1,3,1,1]`, a TP rating, refers to `ModalityID` 0, `ReaderID` 2 and `CaseID` 70, the first diseased case.
- `x$ratings$LL[2,5,4,1]`, a TP rating, refers to `ModalityID` 1, `ReaderID` 4 and `CaseID` 73, the fourth diseased case.

3.6 Correspondence between LL member of dataset and the TP worksheet



	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	LesionID	TP_Rating				
2	0	0	70	1	5				
3	0	0	71	1	5				
4	0	0	72	1	5				
5	0	0	73	1	5				
6	0	0	74	1	4				
7	1	0	70	1	5				
8	1	0	71	1	3				
9	1	0	72	1	5				
10	1	0	73	1	5				
11	1	0	74	1	5				
12	2	0	70	1	5				
13	2	0	71	1	4				
14	2	0	72	1	5				
15	2	0	73	1	5				
16	2	0	74	1	5				
17	3	0	70	1	5				
18	3	0	71	1	5				
19	3	0	72	1	5				
20	3	0	73	1	5				
21	3	0	74	1	5				
22	4	0	70	1	5				
23	4	0	71	1	2				
24	4	0	72	1	5				

- The list member `x$ratings$LL` is an array with `dim = c(2,5,5,1)`.
  - The first dimension (2) comes from the number of modalities.
  - The second dimension (5) comes from the number of readers.
  - The third dimension (5) comes from the number of diseased cases.
  - The fourth dimension is always 1 for an ROC dataset.
- The value of `x$ratings$LL[1,1,5,1]`, i.e., 4, corresponds to row 6 of the TP table, i.e., to `ModalityID` = 0, `ReaderID` = 0 and `CaseID` = 74.
- The value of `x$ratings$LL[1,2,2,1]`, i.e., 3, corresponds to row 8 of the TP table, i.e., to `ModalityID` = 0, `ReaderID` = 1 and `CaseID` = 71.
- The value of `x$ratings$LL[1,4,4,1]`, i.e., 5, corresponds to row 21 of the TP table, i.e., to `ModalityID` = 0, `ReaderID` = 3 and `CaseID` = 74.

- The value of `x$ratings$LL[1,5,2,1]`, i.e., 2, corresponds to row 23 of the TP table, i.e., to `ModalityID` = 0, `ReaderID` = 4 and `CaseID` = 71.
- There are no `-Inf` values in `x$ratings$LL`: `any(x$ratings$LL == -Inf)` = FALSE. This is true for any ROC dataset.

3.7 References

Chapter 4

Data format and reading FROC data

4.1 TBA How much finished

90%

4.2 Introduction

In the Free-response Receiver Operating Characteristic (FROC) paradigm the observer searches each case for signs of **localized disease** and marks and rates localized regions that are sufficiently suspicious for presence of disease. FROC data consists of **mark-rating pairs**, where each mark is a localized-region that was considered sufficiently suspicious for presence of a localized lesion and the rating is it's confidence level. As in the ROC paradigm, the rating can be an integer or quasi-continuous (e.g., 0 – 100), or a floating point value, *as long as higher numbers represent greater confidence in presence of a lesion at the indicated region.* This is termed a positive-directed confidence level scheme. By adopting a proximity criterion, the investigator classifies each mark as a lesion localization (`LL`) - if it is close to a real lesion - or a non-lesion localization (`NL`) otherwise.

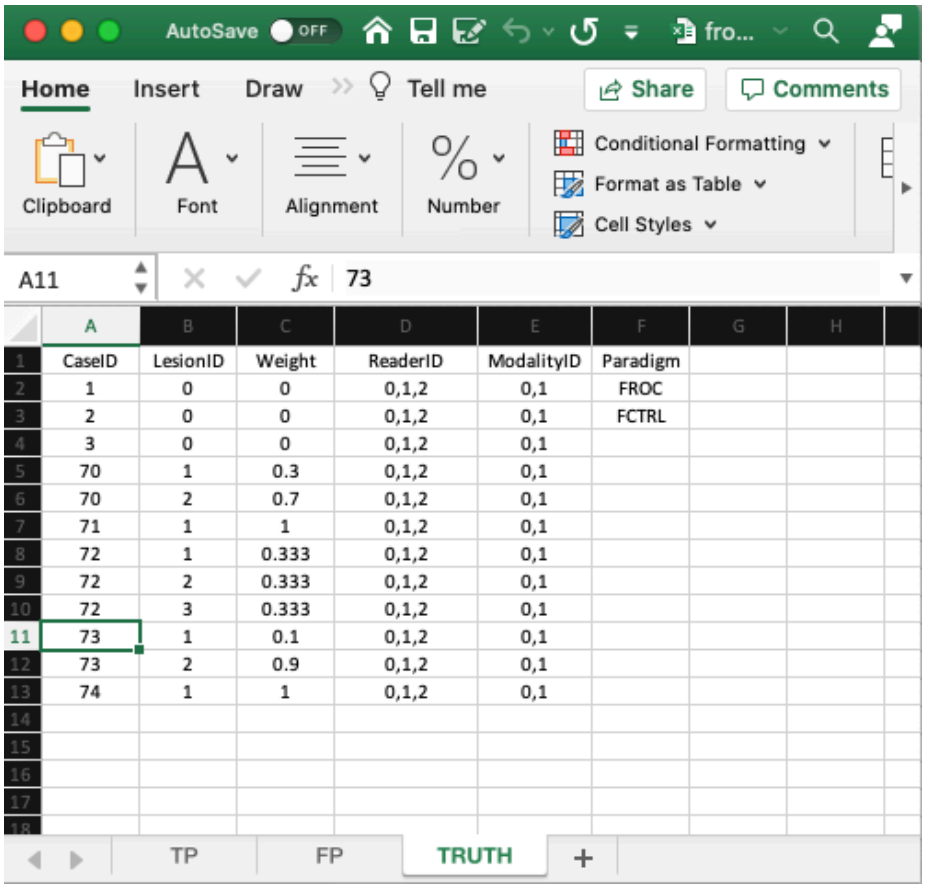
The purpose of this chapter is to:

- Explain the data format of the input Excel file for FROC datasets.
- Explain the format of the FROC dataset.
- Explain the lesion distribution array returned by `UtilLesionDistr()`.
- Explain the lesion weights array returned by `UtilLesionWeightsDistr()`.

- Details on the FROC paradigm are in my book (Chakraborty, 2017).

The chapter is illustrated with a toy data file, `R/quick-start/frocCr.xlsx` in which readers '0', '1' and '2' interpret 8 cases in two modalities, '0' and '1'. The design is 'factorial', abbreviated to `FCTRL` in the software; this is also termed a 'fully-crossed' design. The Excel file has three worksheets named `Truth`, `NL` (or `FP`) and `LL` (or `TP`).

4.3 The `Truth` worksheet



	A	B	C	D	E	F
1	CaseID	LesionID	Weight	ReaderID	ModalityID	Paradigm
2	1	0	0	0,1,2	0,1	FROC
3	2	0	0	0,1,2	0,1	FCTRL
4	3	0	0	0,1,2	0,1	
5	70	1	0.3	0,1,2	0,1	
6	70	2	0.7	0,1,2	0,1	
7	71	1	1	0,1,2	0,1	
8	72	1	0.333	0,1,2	0,1	
9	72	2	0.333	0,1,2	0,1	
10	72	3	0.333	0,1,2	0,1	
11	73	1	0.1	0,1,2	0,1	
12	73	2	0.9	0,1,2	0,1	
13	74	1	1	0,1,2	0,1	

- The `Truth` worksheet contains 6 columns: `CaseID`, `LesionID`, `Weight`, `ReaderID`, `ModalityID` and `Paradigm`.
- Since a diseased case may have more than one lesion, the first five columns contain **at least** as many rows as there are cases (images) in the dataset. There are 8 cases in the dataset and 12 rows of data, because some of the diseased cases contain more than one lesion.

- `CaseID`: unique **integers** representing the cases in the dataset: '1', '2', '3', the 3 non-diseased cases, and '70', '71', '72', '73', '74', the 5 diseased cases. The ordering of the numbers is inconsequential. [1]
- `LesionID`: integers 0, 1, 2, etc.,
  - Each 0 represents a non-diseased case,
  - Each 1 represents the *first* lesion on a diseased case, 2 the *second* lesion, if present, and so on.
  - This field is zero for non-diseased cases '1', '2', '3'.
  - For the first diseased case, i.e., '70', it is 1 for the first lesion and 2 for the second lesion.
  - For the second diseased case i.e., '71', it is 1, as this case has only one lesion.
  - For the third diseased case, i.e., '72', it is 1 for the first lesion, 2 for the second lesion and 3 for the third lesion.
  - For the fourth diseased case, i.e., '73', it is 1 for the first lesion and 2 for the second lesion.
  - For the fifth diseased case i.e., '74', it is 1, as this case has only one lesion.
- There are 3 non-diseased cases in the dataset (the number of 0's in the `LesionID` column).
- There are 5 diseased cases in the dataset (the number of 1's in the `LesionID` column).
- `Weight` or clinical importance - e.g., mortality associated with lesion:
  - non-negative floating point values
  - 0 for each non-diseased case
  - For each diseased case values that sum to unity.
  - A simple way to assign equal weights to all lesions in a case is to fill the `Weight` column with zeroes.
- `LesionID`
  - Diseased case `70` has two lesions, with `LesionID`s '1' and '2', and weights 0.3 and 0.7.
  - Diseased case `71` has one lesion, with `LesionID` = 1, and `Weight` = 1.
  - Diseased case `72` has three lesions, with `LesionID`s 1, 2 and 3 and weights 1/3 each.
  - Diseased case `73` has two lesions, with `LesionID`s 1, and 2 and weights 0.1 and 0.9.
  - Diseased case `74` has one lesion, with `LesionID` = 1 and `Weight` = 1.

[1] `CaseID` should not be so large that it cannot be represented in Excel by an integer; to be safe use unsigned short 8-bit integers. For example, 108057200 or 9971103254 are too large to be a valid `caseID` and may cause errors.

- `ReaderID`: a comma-separated listing of readers, each represented by a unique **text label**, that have interpreted the case. In the example shown below each cell has the value '0, 1, 2'.
- There are 3 readers in the dataset, as each cell in the `ReaderID` column contains '0, 1, 2'.
- `ModalityID`: a comma-separated listing of modalities (or treatments), each represented by a unique **integer**, that apply to each case. In the example each cell has the value `0, 1`. **Each cell has to be text formatted.**
- There are 2 modalities in the dataset, as each cell in the `ModalityID` column contains '0, 1'.
- `Paradigm`: The contents are `FROC` and `FCTRL`: this is an `FROC` dataset and the design is "factorial".

4.4 Reading the FROC dataset

The example shown above corresponds to file `R/quick-start/frocCr.xlsx` in the project directory. The next code chunk reads this file into an `R` object `x`.

```
frocCr <- "R/quick-start/frocCr.xlsx"
x <- DfReadDataFile(frocCr, newExcelFileFormat = TRUE)
str(x)
#> List of 3
#>  $ ratings     :List of 3
#>   ..$ NL   : num [1:2, 1:3, 1:8, 1:2] 1.02 2.89 2.21 3.01 2.14 ...
#>   ..$ LL   : num [1:2, 1:3, 1:5, 1:3] 5.28 5.2 5.14 4.77 4.66 4.87 3.01 3.27 3.31 3
#>   ..$ LL_IL: logi NA
#>  $ lesions     :List of 3
#>   ..$ perCase: int [1:5] 2 1 3 2 1
#>   ..$ IDs    : num [1:5, 1:3] 1 1 1 1 1 ...
#>   ..$ weights: num [1:5, 1:3] 0.3 1 0.333 0.1 1 ...
#>  $ descriptions:List of 7
#>   ..$ fileName     : chr "frocCr"
#>   ..$ type         : chr "FROC"
#>   ..$ name         : logi NA
#>   ..$ truthTableStr: num [1:2, 1:3, 1:8, 1:4] 1 1 1 1 1 1 1 1 1 1 ...
#>   ..$ design       : chr "FCTRL"
#>   ..$ modalityID   : Named chr [1:2] "0" "1"
#>   .. ..- attr(*, "names")= chr [1:2] "0" "1"
#>   ..$ readerID     : Named chr [1:3] "0" "1" "2"
#>   .. ..- attr(*, "names")= chr [1:3] "0" "1" "2"
```

This follows the general description in Chapter 2. The differences are described below.

- The `x$descriptions$type` member indicates that this is an `FROC` dataset.
- The `x$lesions$perCase` member is a vector whose contents reflect the number of lesions in each diseased case, i.e., 2, 1, 3, 2, 1 in the current example.
- The `x$lesions$IDs` member indicates the labeling of the lesions in each diseased case.

```
x$lesions$IDs
#>      [,1] [,2] [,3]
#> [1,]    1    2 -Inf
#> [2,]    1 -Inf -Inf
#> [3,]    1    2    3
#> [4,]    1    2 -Inf
#> [5,]    1 -Inf -Inf
```

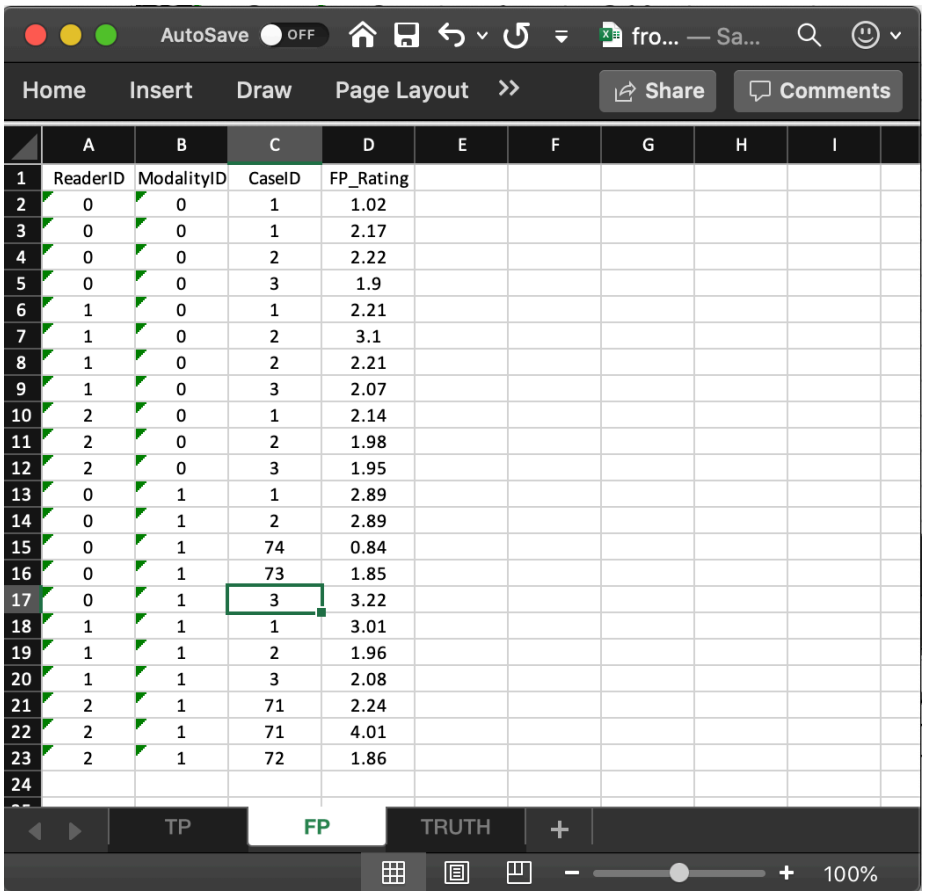
- This shows that the lesions on the first diseased case are labeled '1' and '2'. The `-Inf` is a filler used to denote a missing value. The second diseased case has one lesion labeled '1'. The third diseased case has three lesions labeled '1', '2' and '3', etc.
- The `lesionWeight` member is the clinical importance of each lesion. Lacking specific clinical reasons, the lesions should be equally weighted; this is *not* true for this toy dataset.

```
x$lesions$weights
#>           [,1]      [,2]      [,3]
#> [1,] 0.3000000 0.7000000      -Inf
#> [2,] 1.0000000      -Inf      -Inf
#> [3,] 0.3333333 0.3333333 0.3333333
#> [4,] 0.1000000 0.9000000      -Inf
#> [5,] 1.0000000      -Inf      -Inf
```

- The first diseased case has two lesions, the first has weight 0.3 and the second has weight 0.7.
- The second diseased case has one lesion with weight 1.
- The third diseased case has three equally weighted lesions, each with weight 1/3. Etc.

4.5 The false positive (FP) ratings

These are found in the `FP` or `NL` worksheet.



	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	FP_Rating					
2	0	0	1	1.02					
3	0	0	1	2.17					
4	0	0	2	2.22					
5	0	0	3	1.9					
6	1	0	1	2.21					
7	1	0	2	3.1					
8	1	0	2	2.21					
9	1	0	3	2.07					
10	2	0	1	2.14					
11	2	0	2	1.98					
12	2	0	3	1.95					
13	0	1	1	2.89					
14	0	1	2	2.89					
15	0	1	74	0.84					
16	0	1	73	1.85					
17	0	1	3	3.22					
18	1	1	1	3.01					
19	1	1	2	1.96					
20	1	1	3	2.08					
21	2	1	71	2.24					
22	2	1	71	4.01					
23	2	1	72	1.86					
24									

- It consists of 4 columns, of equal length. The common length is an integer random variable greater than or equal to zero. It could be zero if the dataset has no NL marks (a possibility if the lesions are very easy to find and the observer has perfect performance).

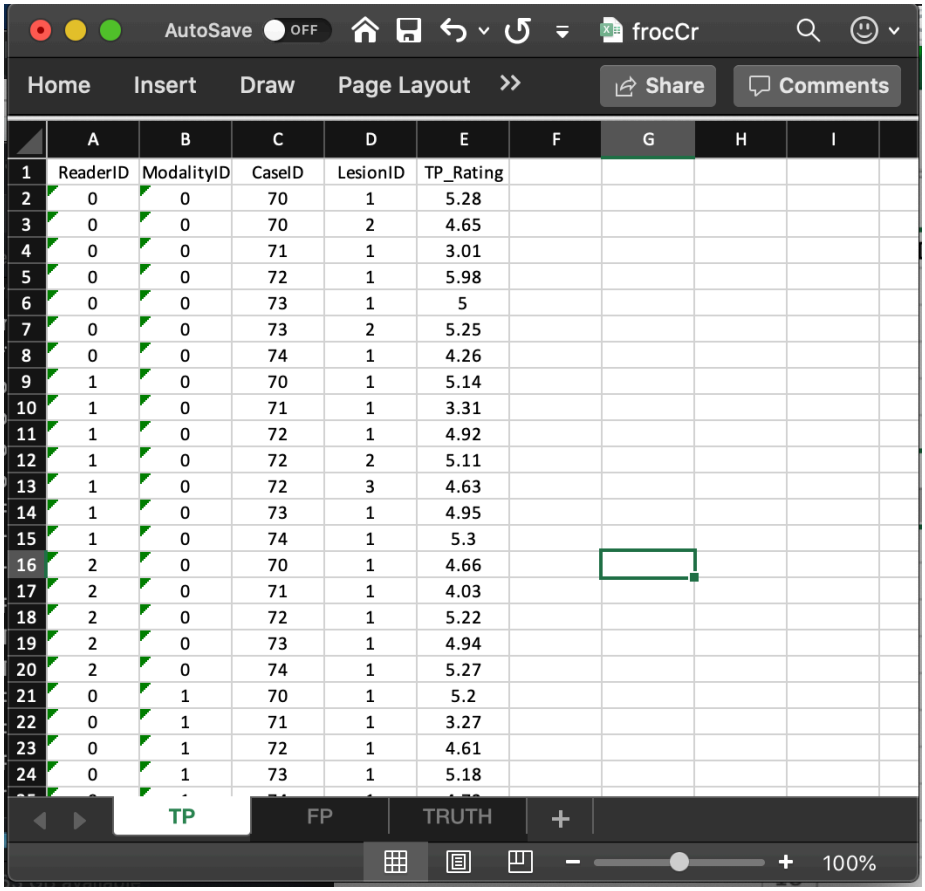
- In the example dataset, the common length is 22.
- `ReaderID`: the reader labels: these must be `0`, `1`, or `2`, as declared in the `Truth` worksheet.
- `ModalityID`: the modality labels: must be `0` or `1`, as declared in the `Truth` worksheet.
- `CaseID`: the labels of cases with `NL` marks. In the FROC paradigm `NL` events can occur on non-diseased **and** diseased cases.
- `FP_Rating`: the floating point ratings of `NL` marks. Each row of this worksheet yields a rating corresponding to the values of `ReaderID`, `ModalityID` and `CaseID` for that row.
- For `ModalityID` 0, `ReaderID` 0 and `CaseID` 1 (the first non-diseased case declared in the `Truth` worksheet), there is a single `NL` mark that was rated

1.02, corresponding to row 2 of the FP worksheet.

- Diseased cases with NL marks are also recorded in the FP worksheet. Some examples are seen at rows 15, 16 and 21, 22, 23.
- Rows 21 and 22 show that caseID = 71 got two NL marks, rated 2.24, 4.01.
- Since this is the *only* case with two NL marks, it determines the length of the fourth dimension of the x$ratings$NL list member, 2. Absent this case, the length would have been one.
- The case with the most NL marks determines the length of the fourth dimension of the x$ratings$NL list member.
- The reader should confirm that the ratings in x$ratings$NL reflect the contents of the FP worksheet.

4.6 The true positive (TP) ratings

These are found in the TP or LL worksheet, see below.



	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	LesionID	TP_Rating				
2	0	0	70	1	5.28				
3	0	0	70	2	4.65				
4	0	0	71	1	3.01				
5	0	0	72	1	5.98				
6	0	0	73	1	5				
7	0	0	73	2	5.25				
8	0	0	74	1	4.26				
9	1	0	70	1	5.14				
10	1	0	71	1	3.31				
11	1	0	72	1	4.92				
12	1	0	72	2	5.11				
13	1	0	72	3	4.63				
14	1	0	73	1	4.95				
15	1	0	74	1	5.3				
16	2	0	70	1	4.66				
17	2	0	71	1	4.03				
18	2	0	72	1	5.22				
19	2	0	73	1	4.94				
20	2	0	74	1	5.27				
21	0	1	70	1	5.2				
22	0	1	71	1	3.27				
23	0	1	72	1	4.61				
24	0	1	73	1	5.18				

- This worksheet can only have diseased cases. The presence of a non-diseased case in this worksheet will generate an error.
- The common vertical length, 31 in this example, is a-priori unpredictable. The maximum possible length, assuming every lesion is marked for each modality, reader and diseased case, is 9 X 2 X 3 = 54. The 9 comes from the total number of non-zero entries in the `LesionID` column of the `Truth` worksheet, the 2 from the number of modalities and 3 from the number of readers.
- The fact that the actual length (31) is smaller than the maximum length (54) means that there are combinations of modality, reader and diseased cases on which some lesions were not marked.
- As examples, line 2 in the worksheet, the first lesion in `CaseID` equal to `70` was marked (and rated 5.28) in `ModalityID 0` and `ReaderID 0`. Line 3 in the worksheet, the second lesion in `CaseID` equal to `70` was also marked (and rated 4.65) in `ModalityID 0` and `ReaderID 0`. However, lesions 2 and 3 in `CaseID` = 72 were not marked (line 5 in the worksheet indicates that for this modality-reader-case combination only the first lesion was marked).
- The length of the fourth dimension of the `x$ratings$LL` list member, 3 in the present example, is determined by the diseased case (72) with the most lesions in the `Truth` worksheet.
- The reader should confirm that the ratings in `x$ratings$LL` reflect the contents of the `TP` worksheet.

4.7 On the distribution of numbers of lesions in diseased cases

- Consider a much larger dataset, `dataset11`, with structure as shown below (for descriptions of all embedded datasets the `RJafroc` documentation):

```
x <- dataset11
str(x)
#> List of 3
#>  $ ratings     :List of 3
#>   ..$ NL   : num [1:4, 1:5, 1:158, 1:4] -Inf -Inf -Inf -Inf -Inf ...
#>   ..$ LL   : num [1:4, 1:5, 1:115, 1:20] -Inf -Inf -Inf -Inf -Inf ...
#>   ..$ LL_IL: logi NA
#>  $ lesions     :List of 3
#>   ..$ perCase: int [1:115] 6 4 7 1 3 3 3 8 11 2 ...
#>   ..$ IDs    : num [1:115, 1:20] 1 1 1 1 1 1 1 1 1 1 ...
#>   ..$ weights: num [1:115, 1:20] 0.167 0.25 0.143 1 0.333 ...
#>  $ descriptions:List of 7
#>   ..$ fileName     : chr "dataset11"
```

```
#>   ..$ type          : chr "FROC"
#>   ..$ name          : chr "DOBBINS-1"
#>   ..$ truthTableStr: num [1:4, 1:5, 1:158, 1:21] 1 1 1 1 1 1 1 1 1 ...
#>   ..$ design        : chr "FCTRL"
#>   ..$ modalityID    : Named chr [1:4] "1" "2" "3" "4"
#>   .. ..- attr(*, "names")= chr [1:4] "1" "2" "3" "4"
#>   ..$ readerID      : Named chr [1:5] "1" "2" "3" "4" ...
#>   .. ..- attr(*, "names")= chr [1:5] "1" "2" "3" "4" ...
```

- Focus for now in the 115 diseased cases.
- The numbers of lesions in these cases is contained in `x$lesions$perCase`.

```
x$lesions$perCase
#>   [1]  6  4  7  1  3  3  3  8 11  2  4  6  2 16  5  2  8  3  4  7 11  1  4  3  4
#>  [26]  4  7  3  2  5  2  2  7  6  6  4 10 20 12  6  4  7 12  5  1  1  5  1  2  8
#>  [51]  3  1  2  2  3  2  8 16 10  1  2  2  6  3  2  2  4  6 10 11  1  2  6  2  4
#>  [76]  5  2  9  6  6  8  3  8  7  1  1  6  3  2  1  9  8  8  2  2 12  1  1  1  1
#> [101]  1  3  1  2  2  1  1  1  1  3  1  1  1  2  1
```

- For example, the first diseased case contains 6 lesions, the second contains 4 lesions, the third contains 7 lesions, etc. and the last diseased case contains 1 lesion.
- To get an idea of the distribution of the numbers of lesions per diseased cases, one could interrogate this vector as shown below using the `which()` function:

```
for (el in 1:max(x$lesions$perCase)) cat(
  "number of diseased cases with", el, "lesions = ",
  length(which(x$lesions$perCase == el)), "\n")
#> number of diseased cases with 1 lesions =  25
#> number of diseased cases with 2 lesions =  23
#> number of diseased cases with 3 lesions =  13
#> number of diseased cases with 4 lesions =  10
#> number of diseased cases with 5 lesions =  5
#> number of diseased cases with 6 lesions =  11
#> number of diseased cases with 7 lesions =  6
#> number of diseased cases with 8 lesions =  8
#> number of diseased cases with 9 lesions =  2
#> number of diseased cases with 10 lesions =  3
#> number of diseased cases with 11 lesions =  3
#> number of diseased cases with 12 lesions =  3
#> number of diseased cases with 13 lesions =  0
#> number of diseased cases with 14 lesions =  0
#> number of diseased cases with 15 lesions =  0
```

```
#> number of diseased cases with 16 lesions =  2
#> number of diseased cases with 17 lesions =  0
#> number of diseased cases with 18 lesions =  0
#> number of diseased cases with 19 lesions =  0
#> number of diseased cases with 20 lesions =  1
```

- This tells us that 25 cases contain 1 lesion
- Likewise, 23 cases contain 2 lesions
- Etc.

4.7.1 Definition of `lesDistr` array

- What is the fraction of (diseased) cases with 1 lesion, 2 lesions etc.

```
for (el in 1:max(x$lesions$perCase)) cat("fraction of diseased cases with", el, "lesion
                                         length(which(x$lesions$perCase == el))/le
#> fraction of diseased cases with 1 lesions =  0.2173913
#> fraction of diseased cases with 2 lesions =  0.2
#> fraction of diseased cases with 3 lesions =  0.1130435
#> fraction of diseased cases with 4 lesions =  0.08695652
#> fraction of diseased cases with 5 lesions =  0.04347826
#> fraction of diseased cases with 6 lesions =  0.09565217
#> fraction of diseased cases with 7 lesions =  0.05217391
#> fraction of diseased cases with 8 lesions =  0.06956522
#> fraction of diseased cases with 9 lesions =  0.0173913
#> fraction of diseased cases with 10 lesions =  0.02608696
#> fraction of diseased cases with 11 lesions =  0.02608696
#> fraction of diseased cases with 12 lesions =  0.02608696
#> fraction of diseased cases with 13 lesions =  0
#> fraction of diseased cases with 14 lesions =  0
#> fraction of diseased cases with 15 lesions =  0
#> fraction of diseased cases with 16 lesions =  0.0173913
#> fraction of diseased cases with 17 lesions =  0
#> fraction of diseased cases with 18 lesions =  0
#> fraction of diseased cases with 19 lesions =  0
#> fraction of diseased cases with 20 lesions =  0.008695652
```

- This tells us that fraction 0.217 of (diseased) cases contain 1 lesion
- And fraction 0.2 of (diseased) cases contain 2 lesions
- Etc.
- This information is obtained using the function `UtilLesionDistr()`

```
lesDistr <- UtilLesionDistr(x)
lesDistr
#>       [,1]        [,2]
#>  [1,]    1 0.217391304
#>  [2,]    2 0.200000000
#>  [3,]    3 0.113043478
#>  [4,]    4 0.086956522
#>  [5,]    5 0.043478261
#>  [6,]    6 0.095652174
#>  [7,]    7 0.052173913
#>  [8,]    8 0.069565217
#>  [9,]    9 0.017391304
#> [10,]   10 0.026086957
#> [11,]   11 0.026086957
#> [12,]   12 0.026086957
#> [13,]   16 0.017391304
#> [14,]   20 0.008695652
```

- The `UtilLesionDistr()` function returns an array with two columns and number of rows equal to the number of *distinct non-zero* values of lesions per case.
- The first column contains the number of distinct non-zero values of lesions per case, 14 in the current example.
- The second column contains the fraction of diseased cases with the number of lesions indicated in the first column.
- The second column must sum to unity

```
sum(UtilLesionDistr(x)[,2])
#> [1] 1
```

- The lesion distribution array will come in handy when it comes to predicting the operating characteristics from using the Radiological Search Model (RSM), as detailed in TBA Chapter 17.

4.8 Definition of `lesWghtDistr` array

- This is returned by `UtilLesionWeightsDistr()`.
- This contains the same number of rows as `lesDistr`.
- The number of columns is one plus the number of rows as `lesDistr`.
- The first column contains the number of distinct non-zero values of lesions per case, 14 in the current example.
- The second through the last columns contain the weights of cases with number of lesions per case corresponding to row 1.

- Missing values are filled with `-Inf`.

```
lesWghtDistr <- UtilLesionWeightsDistr(x)
cat("dim(lesDistr) =", dim(lesDistr),"\n")
#> dim(lesDistr) = 14 2
cat("dim(lesWghtDistr) =", dim(lesWghtDistr),"\n")
#> dim(lesWghtDistr) = 14 21
cat("lesWghtDistr = \n\n")
#> lesWghtDistr =
lesWghtDistr
#>       [,1]       [,2]       [,3]       [,4]       [,5]       [,6]       [,7]
#>  [1,]    1 1.00000000       -Inf       -Inf       -Inf       -Inf       -Inf
#>  [2,]    2 0.50000000 0.50000000       -Inf       -Inf       -Inf       -Inf
#>  [3,]    3 0.33333333 0.33333333 0.33333333       -Inf       -Inf       -Inf
#>  [4,]    4 0.25000000 0.25000000 0.25000000 0.25000000       -Inf       -Inf
#>  [5,]    5 0.20000000 0.20000000 0.20000000 0.20000000 0.20000000       -Inf
#>  [6,]    6 0.16666667 0.16666667 0.16666667 0.16666667 0.16666667 0.16666667
#>  [7,]    7 0.14285714 0.14285714 0.14285714 0.14285714 0.14285714 0.14285714
#>  [8,]    8 0.12500000 0.12500000 0.12500000 0.12500000 0.12500000 0.12500000
#>  [9,]    9 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111 0.11111111
#> [10,]   10 0.10000000 0.10000000 0.10000000 0.10000000 0.10000000 0.10000000
#> [11,]   11 0.09090909 0.09090909 0.09090909 0.09090909 0.09090909 0.09090909
#> [12,]   12 0.08333333 0.08333333 0.08333333 0.08333333 0.08333333 0.08333333
#> [13,]   16 0.06250000 0.06250000 0.06250000 0.06250000 0.06250000 0.06250000
#> [14,]   20 0.05000000 0.05000000 0.05000000 0.05000000 0.05000000 0.05000000
#>            [,8]       [,9]       [,10]      [,11]      [,12]      [,13]  [,14]
#>  [1,]       -Inf       -Inf       -Inf       -Inf       -Inf       -Inf   -Inf
#>  [2,]       -Inf       -Inf       -Inf       -Inf       -Inf       -Inf   -Inf
#>  [3,]       -Inf       -Inf       -Inf       -Inf       -Inf       -Inf   -Inf
#>  [4,]       -Inf       -Inf       -Inf       -Inf       -Inf       -Inf   -Inf
#>  [5,]       -Inf       -Inf       -Inf       -Inf       -Inf       -Inf   -Inf
#>  [6,]       -Inf       -Inf       -Inf       -Inf       -Inf       -Inf   -Inf
#>  [7,] 0.14285714       -Inf       -Inf       -Inf       -Inf       -Inf   -Inf
#>  [8,] 0.12500000 0.12500000       -Inf       -Inf       -Inf       -Inf   -Inf
#>  [9,] 0.11111111 0.11111111 0.11111111       -Inf       -Inf       -Inf   -Inf
#> [10,] 0.10000000 0.10000000 0.10000000 0.10000000       -Inf       -Inf   -Inf
#> [11,] 0.09090909 0.09090909 0.09090909 0.09090909 0.09090909       -Inf   -Inf
#> [12,] 0.08333333 0.08333333 0.08333333 0.08333333 0.08333333 0.08333333   -Inf
#> [13,] 0.06250000 0.06250000 0.06250000 0.06250000 0.06250000 0.06250000 0.0625
#> [14,] 0.05000000 0.05000000 0.05000000 0.05000000 0.05000000 0.05000000 0.0500
#>        [,15]  [,16]  [,17] [,18] [,19] [,20] [,21]
#>  [1,]   -Inf   -Inf   -Inf  -Inf  -Inf  -Inf  -Inf
#>  [2,]   -Inf   -Inf   -Inf  -Inf  -Inf  -Inf  -Inf
#>  [3,]   -Inf   -Inf   -Inf  -Inf  -Inf  -Inf  -Inf
#>  [4,]   -Inf   -Inf   -Inf  -Inf  -Inf  -Inf  -Inf
```

```
#>  [5,]   -Inf   -Inf   -Inf  -Inf  -Inf  -Inf  -Inf
#>  [6,]   -Inf   -Inf   -Inf  -Inf  -Inf  -Inf  -Inf
#>  [7,]   -Inf   -Inf   -Inf  -Inf  -Inf  -Inf  -Inf
#>  [8,]   -Inf   -Inf   -Inf  -Inf  -Inf  -Inf  -Inf
#>  [9,]   -Inf   -Inf   -Inf  -Inf  -Inf  -Inf  -Inf
#> [10,]   -Inf   -Inf   -Inf  -Inf  -Inf  -Inf  -Inf
#> [11,]   -Inf   -Inf   -Inf  -Inf  -Inf  -Inf  -Inf
#> [12,]   -Inf   -Inf   -Inf  -Inf  -Inf  -Inf  -Inf
#> [13,] 0.0625 0.0625 0.0625  -Inf  -Inf  -Inf  -Inf
#> [14,] 0.0500 0.0500 0.0500  0.05  0.05  0.05  0.05
```

- Row 3 corresponds to 3 lesions per case and the weights are 1/3, 1/3 and 1/3.
- Row 13 corresponds to 16 lesions per case and the weights are 0.06250000, 0.06250000, ..., repeated 13 times.
- Note that the number of rows is less than the maximum number of lesions per case (20).
- This is because some configurations of lesions per case (e.g., cases with 13 lesions per case) do not occur in this dataset.

4.9 References

Chapter 5

Data format and reading LROC data

5.1 TBA How much finished

70%

5.2 Introduction

In the Localization Receiver Operating Characteristic (LROC) paradigm (Starr et al., 1977, 1975; Swensson, 1996) the observer is restricted to at most one mark-rating pair per case. Additionally, each diseased case has *exactly* one lesion. On a diseased case and if the mark is close to the real lesion the investigator classifies the mark as a correct-localization (`CL`). Otherwise it is classified as an incorrect-localization (`IL`). On a non-diseased case the mark is always classified as a false-positive (`FP`).

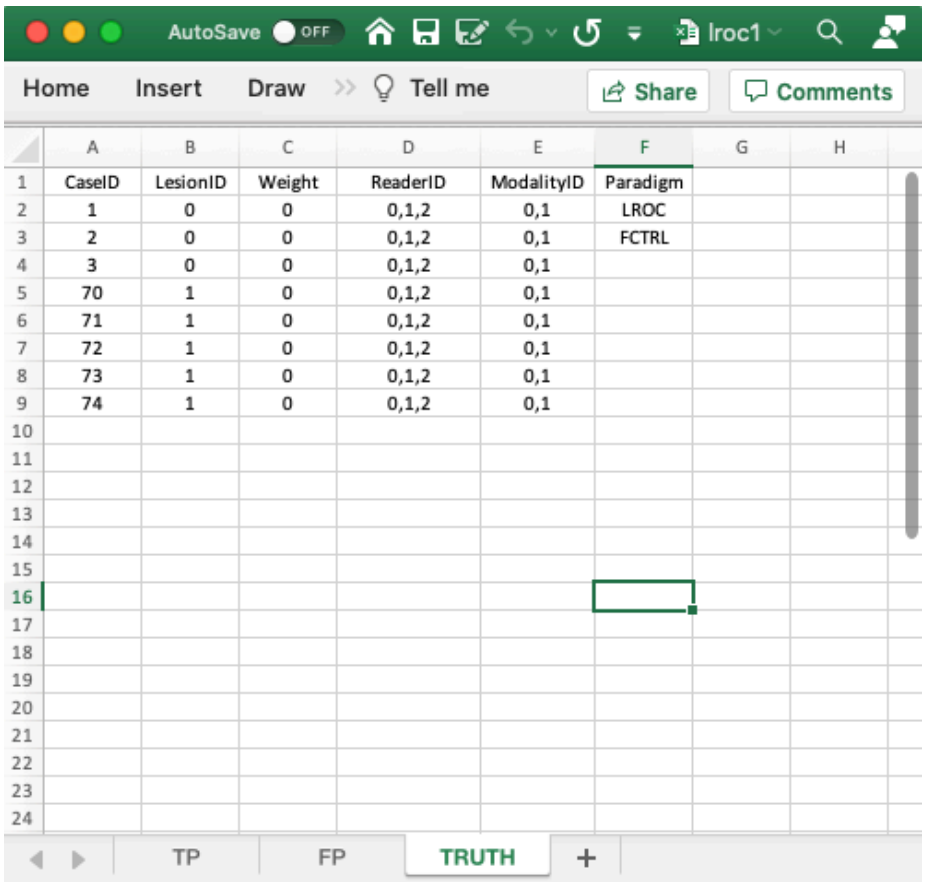
The paradigm is illustrated with a few toy data files, `R/quick-start/lroc?.xlsx`, where `?` is `1` or `2`. These files illustrate two-modality three-reader LROC datasets with 3 non-diseased and 5 diseased cases.

File `lroc1.xlsx` illustrates the classic (i.e., as originally introduced) LROC paradigm where *one mark per case is forced*.

File `lroc2.xlsx` illustrates the paradigm when one mark-rating pair per case is not forced. There is some history behind this: the basic issue was what was the observer supposed to do when there there was nothing to report? Swensson initially thought that even if there was nothing to report, there must be a region, selected from the set of very low confidence regions, which was most likely to be a

lesion. Most radiologists had difficulty with the forced localization requirement - if they saw nothing suspicious, why should they be forced to indicate a location. The paradigm was subsequently altered so that if the confidence level was below a certain value, say 12 percent on a 0 to 100 scale, the radiologist did not have to report a location. LROCFIT software was modified accordingly, and internal to the software the mark was assigned a random location - which ended up being classified as an incorrect-localization in most cases. The fact that there are cases with nothing to report is accounted for in the radiological search model.

5.3 `Truth` worksheet



	A	B	C	D	E	F	G	H
1	CaseID	LesionID	Weight	ReaderID	ModalityID	Paradigm		
2	1	0	0	0,1,2	0,1	LROC		
3	2	0	0	0,1,2	0,1	FCTRL		
4	3	0	0	0,1,2	0,1			
5	70	1	0	0,1,2	0,1			
6	71	1	0	0,1,2	0,1			
7	72	1	0	0,1,2	0,1			
8	73	1	0	0,1,2	0,1			
9	74	1	0	0,1,2	0,1			
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								

- The `Truth` worksheet is similar to that described previously for the ROC and LROC paradigms. The only difference is the first entry in the Paradigm column, which is `LROC`.
- Since a diseased case has one lesion, the first five columns contain as many

rows as there are cases in the dataset. There being 8 cases in the dataset, there are 8 rows of data.

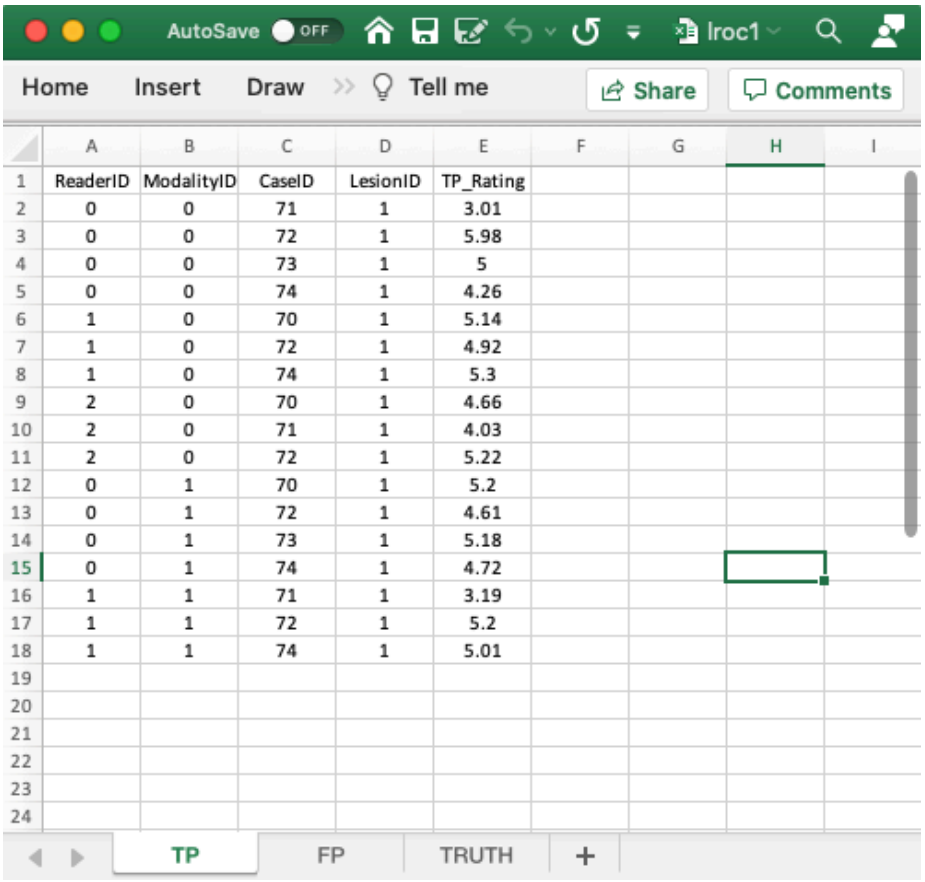
- `CaseID`: unique **integers** representing the cases in the dataset: '1', '2', '3', the 3 non-diseased cases, and '70', '71', '72', '73', '74', the 5 diseased cases.

- `LesionID`: integers 0 or 1.

  - Each 0 represents a non-diseased case,
  - Each 1 represents the solitary lesion in the diseased case.

- There are 3 non-diseased cases in the dataset (the number of 0's in the `LesionID` column).
- There are 5 diseased cases in the dataset (the number of 1's in the `LesionID` column).
- `Weight`: this column is filled with zeroes. With one lesion per case, the weights are irrelevant.
- `ReaderID`: In the example shown each cell has the value '0, 1, 2'. There are 3 readers in the dataset, labeled `0`, `1` and `2`.
- `ModalityID`: In the example each cell has the value `0, 1`. There are 2 modalities in the dataset, labeled `0` and `1`.
- `Paradigm`: The contents are `LROC` and `FCTRL`: this is an `LROC` dataset and the design is "factorial".

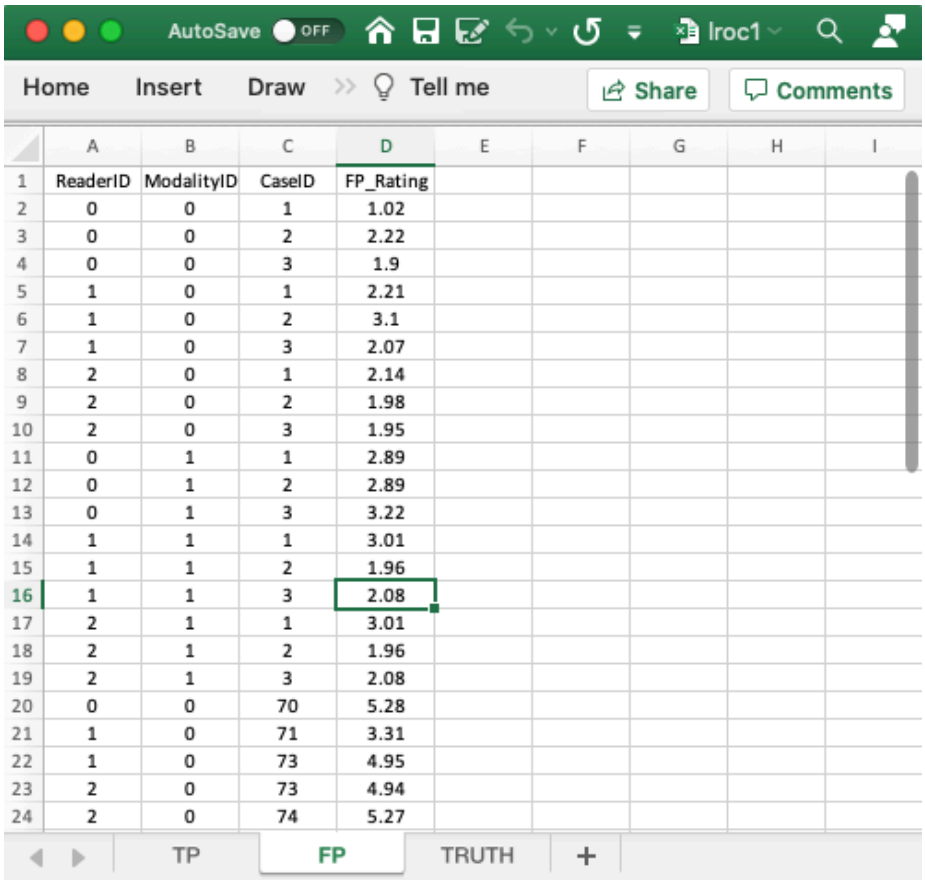
5.4 TP worksheet, forced localization true



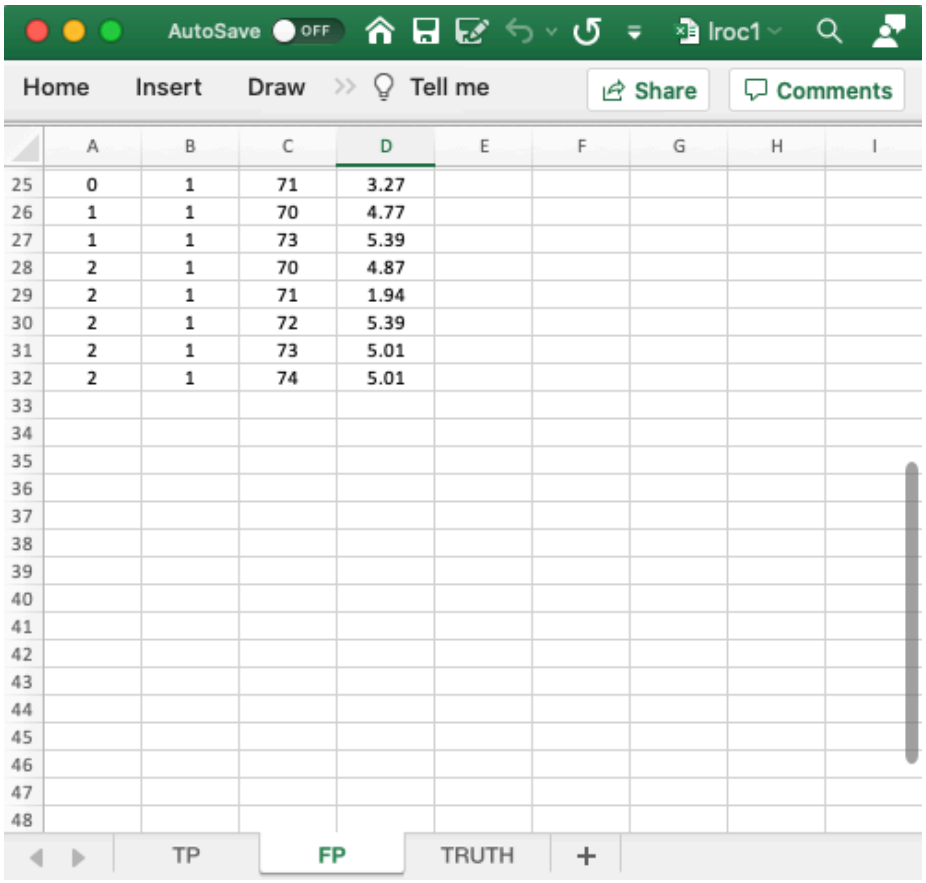
	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	LesionID	TP_Rating				
2	0	0	71	1	3.01				
3	0	0	72	1	5.98				
4	0	0	73	1	5				
5	0	0	74	1	4.26				
6	1	0	70	1	5.14				
7	1	0	72	1	4.92				
8	1	0	74	1	5.3				
9	2	0	70	1	4.66				
10	2	0	71	1	4.03				
11	2	0	72	1	5.22				
12	0	1	70	1	5.2				
13	0	1	72	1	4.61				
14	0	1	73	1	5.18				
15	0	1	74	1	4.72				
16	1	1	71	1	3.19				
17	1	1	72	1	5.2				
18	1	1	74	1	5.01				
19									
20									
21									
22									
23									
24									

- The TP worksheet is similar to that described previously for the ROC and FROC paradigms.
- This worksheet can only have diseased cases. The presence of a non-diseased case in this worksheet will generate an error.
- The key difference is that for each modality-reader and diseased-case combination there can be at most one entry. Also, if a particular combination is missing in the TP worksheet then it must appear in the FP worksheet. This is because this is a forced-mark-per-case dataset.
- There can be at most 30 rows of data in this worksheet: 2 modalities times 3 readers times 5 diseased cases. Since there in fact only 17 rows of data, the missing 13 rows must occur in the FP worksheet.
- Recall that each entry in the TP worksheet represents a correct localization while each missing entry represents an incorrect localization. The incorrect localizations are recorded in the FP worksheet.

5.5 FP worksheet, forced localization true



	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	FP_Rating					
2	0	0	1	1.02					
3	0	0	2	2.22					
4	0	0	3	1.9					
5	1	0	1	2.21					
6	1	0	2	3.1					
7	1	0	3	2.07					
8	2	0	1	2.14					
9	2	0	2	1.98					
10	2	0	3	1.95					
11	0	1	1	2.89					
12	0	1	2	2.89					
13	0	1	3	3.22					
14	1	1	1	3.01					
15	1	1	2	1.96					
16	1	1	3	2.08					
17	2	1	1	3.01					
18	2	1	2	1.96					
19	2	1	3	2.08					
20	0	0	70	5.28					
21	1	0	71	3.31					
22	1	0	73	4.95					
23	2	0	73	4.94					
24	2	0	74	5.27					



	A	B	C	D
25	0	1	71	3.27
26	1	1	70	4.77
27	1	1	73	5.39
28	2	1	70	4.87
29	2	1	71	1.94
30	2	1	72	5.39
31	2	1	73	5.01
32	2	1	74	5.01

- The `FP` worksheet is similar to that described previously for the ROC and FROC paradigms.
- Because of the forced mark requirement, there are 18 rows of data corresponding to non-diseased cases: 2 modalities times 3 readers times 3 non-diseased cases. The missing 13 rows from the TP worksheet are listed next; these correspond to the incorrect localizations on diseased cases. Therefore, the total number of rows in this worksheet is 18 + 13 = 31.
- As an example, it is seen that for `modalityID` = 0 and `readerID` = 0, `caseID` = 70 does not appear in the TP worksheet. The lesion on this case was not localized; therefore it must appear in the FP worksheet as an incorrect localization, which is seen to be true in the FP worksheet.
- As another example, for `modalityID` = 0 and `readerID` = 1, `caseID` = 71 does not appear in the TP worksheet; instead it appears in the FP worksheet.
- As a final example, for `modalityID` = 1 and `readerID` = 2, none of the diseased cases appears in the TP worksheet; instead they all appear in the FP worksheet.

5.6 Reading forced localization true LROC dataset

The images shown above correspond to file `R/quick-start/lroc1.xlsx`. The next code chunk reads this file into an `R` object `x1`. Note the usage of the `lrocForcedMark` flag, which is set to `TRUE`, because this is a forced localization LROC dataset.

```
lroc1 <- "R/quick-start/lroc1.xlsx"
x1 <- DfReadDataFile(lroc1, newExcelFileFormat = TRUE, lrocForcedMark = T)
str(x1)
#> List of 3
#>  $ ratings     :List of 3
#>   ..$ NL   : num [1:2, 1:3, 1:8, 1] 1.02 2.89 2.21 3.01 2.14 3.01 2.22 2.89 3.1 1.96 ...
#>   ..$ LL   : num [1:2, 1:3, 1:5, 1] -Inf 5.2 5.14 -Inf 4.66 ...
#>   ..$ LL_IL: num [1:2, 1:3, 1:5, 1] 5.28 -Inf -Inf 4.77 -Inf ...
#>  $ lesions     :List of 3
#>   ..$ perCase: int [1:5] 1 1 1 1 1
#>   ..$ IDs    : num [1:5, 1] 1 1 1 1 1
#>   ..$ weights: num [1:5, 1] 1 1 1 1 1
#>  $ descriptions:List of 7
#>   ..$ fileName     : chr "lroc1"
#>   ..$ type         : chr "LROC"
#>   ..$ name         : logi NA
#>   ..$ truthTableStr: num [1:2, 1:3, 1:8, 1:2] 1 1 1 1 1 1 1 1 1 1 ...
#>   ..$ design       : chr "FCTRL"
#>   ..$ modalityID   : Named chr [1:2] "0" "1"
#>   .. ..- attr(*, "names")= chr [1:2] "0" "1"
#>   ..$ readerID     : Named chr [1:3] "0" "1" "2"
#>   .. ..- attr(*, "names")= chr [1:3] "0" "1" "2"
```

This follows the general description in Chapter 2. The differences are described below.

- `x1$ratings$NL` is a [2,3,8,1] dimension vector. For each modality and reader, only the first three elements, corresponding to the three non-diseased cases, are finite, the rest are `-Inf`.

For example:

```
x1$ratings$NL[1,1,,1]
#> [1] 1.02 2.22 1.90 -Inf -Inf -Inf -Inf -Inf
```

- x1*ratings*LL`is a [2,3,5,1] dimension vector. For each modality and reader, only the first three elements, corresponding to the three non-diseased cases, are finite, the rest are-Inf`.

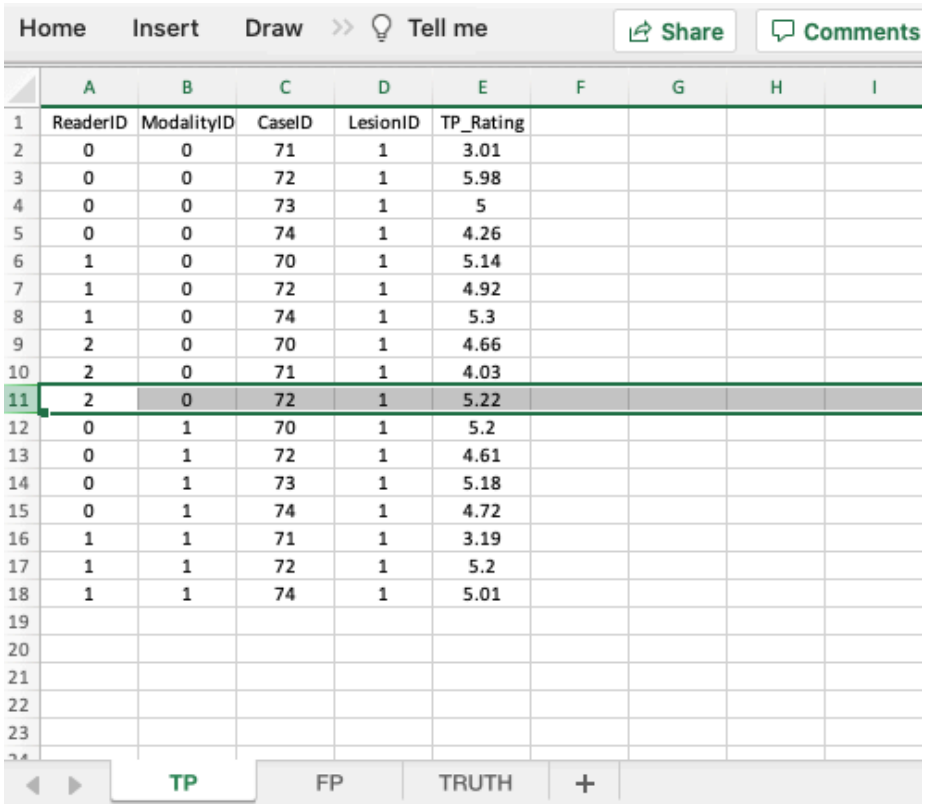
For example, since none of the lesions are localized for `modalityID` = 1 (second modality) and `readerID` = 2 (third reader), the following code yields a vector consisting of five `-Inf` values:

```
x1$ratings$LL[2,3,,1]
#> [1] -Inf -Inf -Inf -Inf -Inf
```

- `x1$ratings$LL_IL` is a [2,3,5,1] dimension vector. These contain the ratings of incorrect localizations on diseased cases. For the just preceding modality-reader combination, this yields a vector with 5 finite values, the ratings of incorrect localizations for `modalityID` = 1 and `readerID` = 2.

```
x1$ratings$LL_IL[2,3,,1]
#> [1] 4.87 1.94 5.39 5.01 5.01
```

5.7 TP worksheet, forced localization false

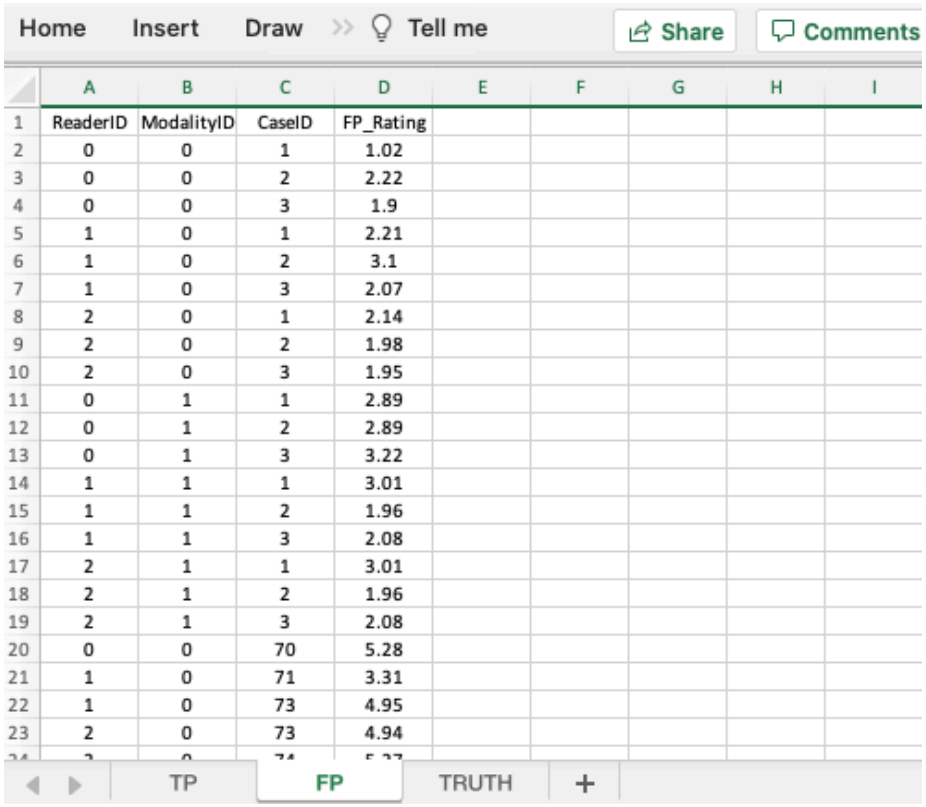


Home Insert Draw Tell me Share Comments

	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	LesionID	TP_Rating				
2	0	0	71	1	3.01				
3	0	0	72	1	5.98				
4	0	0	73	1	5				
5	0	0	74	1	4.26				
6	1	0	70	1	5.14				
7	1	0	72	1	4.92				
8	1	0	74	1	5.3				
9	2	0	70	1	4.66				
10	2	0	71	1	4.03				
11	2	0	72	1	5.22				
12	0	1	70	1	5.2				
13	0	1	72	1	4.61				
14	0	1	73	1	5.18				
15	0	1	74	1	4.72				
16	1	1	71	1	3.19				
17	1	1	72	1	5.2				
18	1	1	74	1	5.01				
19									
20									
21									
22									
23									

TP FP TRUTH +

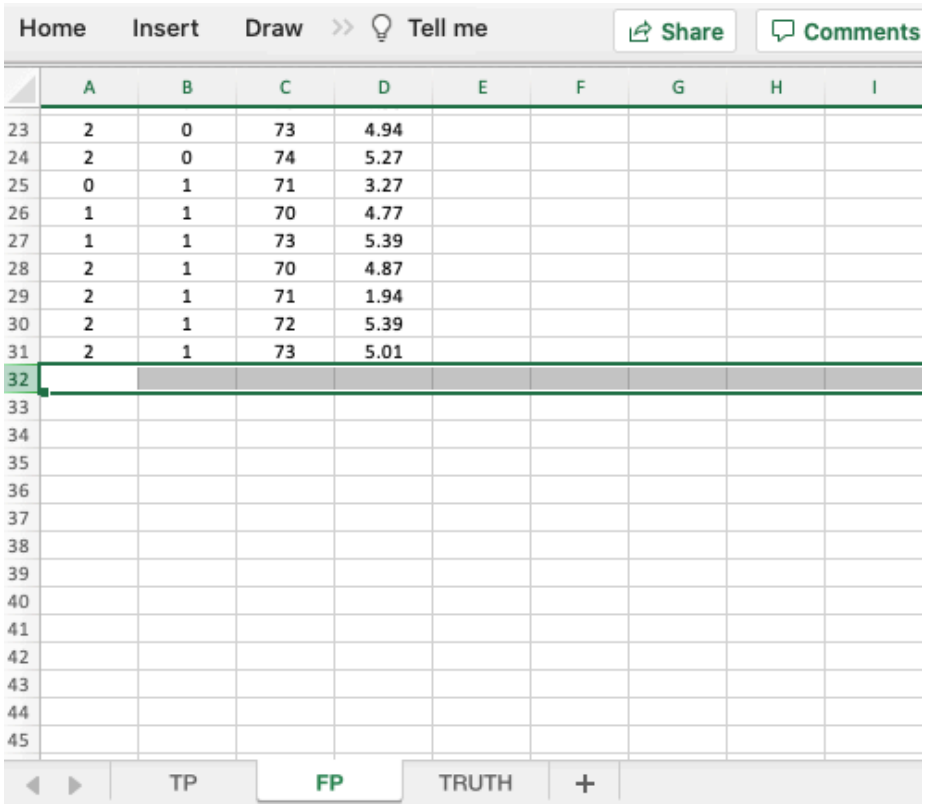
5.8 FP worksheet, forced localization false



Home Insert Draw Tell me Share Comments

	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	FP_Rating					
2	0	0	1	1.02					
3	0	0	2	2.22					
4	0	0	3	1.9					
5	1	0	1	2.21					
6	1	0	2	3.1					
7	1	0	3	2.07					
8	2	0	1	2.14					
9	2	0	2	1.98					
10	2	0	3	1.95					
11	0	1	1	2.89					
12	0	1	2	2.89					
13	0	1	3	3.22					
14	1	1	1	3.01					
15	1	1	2	1.96					
16	1	1	3	2.08					
17	2	1	1	3.01					
18	2	1	2	1.96					
19	2	1	3	2.08					
20	0	0	70	5.28					
21	1	0	71	3.31					
22	1	0	73	4.95					
23	2	0	73	4.94					

TP FP TRUTH +



	A	B	C	D
23	2	0	73	4.94
24	2	0	74	5.27
25	0	1	71	3.27
26	1	1	70	4.77
27	1	1	73	5.39
28	2	1	70	4.87
29	2	1	71	1.94
30	2	1	72	5.39
31	2	1	73	5.01

- If a particular modality-reader-case combination is missing in the TP worksheet then it need not appear in the FP worksheet. This is because this is not a forced-mark-per-case dataset.
- As an example, `modalityID` = 1, `readerID` = 2 and `caseID` = 74 does not appear in either TP or FP worksheets.

5.9 Reading forced localization false LROC dataset

The next example is for file `R/quick-start/lroc2.xlsx`. The following code chunk reads this file into an `R` object `x2`. Note that for this dataset one must set the `lrocForcedMark` flag to `FALSE`, because this is *not* a forced localization LROC dataset. Setting `lrocForcedMark` flag to `TRUE` will generate an error.

```
lroc2 <- "R/quick-start/lroc2.xlsx"
x2 <- DfReadDataFile(lroc2, newExcelFileFormat = TRUE, lrocForcedMark = F)
str(x2)
```

```
#> List of 3
#>  $ ratings     :List of 3
#>   ..$ NL   : num [1:2, 1:3, 1:8, 1] 1.02 2.89 2.21 3.01 2.14 3.01 2.22 2.89 3.1 1.9
#>   ..$ LL   : num [1:2, 1:3, 1:5, 1] -Inf 5.2 5.14 -Inf 4.66 ...
#>   ..$ LL_IL: num [1:2, 1:3, 1:5, 1] 5.28 -Inf -Inf 4.77 -Inf ...
#>  $ lesions     :List of 3
#>   ..$ perCase: int [1:5] 1 1 1 1 1
#>   ..$ IDs    : num [1:5, 1] 1 1 1 1 1
#>   ..$ weights: num [1:5, 1] 1 1 1 1 1
#>  $ descriptions:List of 7
#>   ..$ fileName     : chr "lroc2"
#>   ..$ type         : chr "LROC"
#>   ..$ name         : logi NA
#>   ..$ truthTableStr: num [1:2, 1:3, 1:8, 1:2] 1 1 1 1 1 1 1 1 1 1 ...
#>   ..$ design       : chr "FCTRL"
#>   ..$ modalityID   : Named chr [1:2] "0" "1"
#>   .. ..- attr(*, "names")= chr [1:2] "0" "1"
#>   ..$ readerID     : Named chr [1:3] "0" "1" "2"
#>   .. ..- attr(*, "names")= chr [1:3] "0" "1" "2"
```

- The `x2$ratings$LL` array is a [2,3,5,1] dimension vector. For each modality and reader, only the first three elements, corresponding to the three non-diseased cases, are finite, the rest are `-Inf`.

For example, since none of the lesions are localized for `modalityID` = 1 (second modality) and `readerID` = 2 (third reader), the following code yields a vector consisting of five `-Inf` values:

```
x2$ratings$LL[2,3,,1]
#> [1] -Inf -Inf -Inf -Inf -Inf
```

- The `x2$ratings$LL_IL` is a [2,3,5,1] dimension vector. These contain the ratings of incorrect localizations on diseased cases. For the just preceding modality-reader combination, this yields a vector with 4 finite values, the ratings of incorrect localizations for `modalityID` = 1 and `readerID` = 2.

```
x2$ratings$LL_IL[2,3,,1]
#> [1] 4.87 1.94 5.39 5.01 -Inf
```

For this modality-reader combination case 74 (i.e., the fifth diseased case) was unmarked. It does not appear in either the TP or the FP worksheet.

5.10 Summary

The difference from the previous data structures is the existence of `LL_IL` in the `ratings` list, which contains the ratings of incorrect localizations. Recall that for ROC and FROC paradigms this member was `NA`. When the data obeys forced localization, the corresponding flag should be set to `TRUE`, otherwise it should be set to `FALSE`. The default value of this flag is `NA`, which will work for ROC or FROC datasets. For LROC datasets it should be set to `T/F`.

5.11 References

Chapter 6

DBM analysis text output

6.1 TBA How much finished

50%

6.2 Introduction

This chapter illustrates significance testing using the DBM method.

6.3 Analyzing the ROC dataset

This illustrates the `StSignificanceTesting()` function. The significance testing method is specified as `"DBM"` and the figure of merit `FOM` is specified as "Wilcoxon". The embedded dataset `dataset03` is used.

```
ret <- StSignificanceTesting(dataset03, FOM = "Wilcoxon", method = "DBM")
```

6.4 Explanation of the output

The function returns a list with 5 members:

- `FOMs`: figures of merit.
- `ANOVA`: ANOVA tables.
- `RRRC`: random-reader random-case analyses results.

- FRRC: fixed-reader random-case analyses results.
- RRFC" random-reader fixed-case analyses results.

Let us consider them individually.

```
str(ret$FOMs)
#> List of 3
#>  $ foms        :'data.frame':    2 obs. of  4 variables:
#>   ..$ rdrREADER_1: num [1:2] 0.853 0.85
#>   ..$ rdrREADER_2: num [1:2] 0.865 0.844
#>   ..$ rdrREADER_3: num [1:2] 0.857 0.84
#>   ..$ rdrREADER_4: num [1:2] 0.815 0.814
#>  $ trtMeans    :'data.frame':    2 obs. of  1 variable:
#>   ..$ Estimate: num [1:2] 0.848 0.837
#>  $ trtMeanDiffs:'data.frame':    1 obs. of  1 variable:
#>   ..$ Estimate: num 0.0109
```

- FOMs is a list of 3
  - foms is a [2x4] dataframe: the figure of merit for each of of the four observers in the two treatments.
  - trtMeans is a [2x1] dataframe: the average figure of merit over all readers for each treatment.
  - trtMeanDiffs a [1x1] dataframe: the difference(s) of the reader-averaged figures of merit for all different-treatment pairings. In this example, with only two treatments, there is only one different-treatment pairing.

```
ret$FOMs$foms
#>           rdrREADER_1 rdrREADER_2 rdrREADER_3 rdrREADER_4
#> trtTREAT1  0.85345997  0.86499322  0.85730439  0.81524197
#> trtTREAT2  0.84961556  0.84350972  0.84011759  0.81433740
ret$FOMs$trtMeans
#>             Estimate
#> trtTREAT1 0.84774989
#> trtTREAT2 0.83689507
ret$FOMs$trtMeanDiffs
#>                          Estimate
#> trtTREAT1-trtTREAT2 0.010854817
```

```
str(ret$ANOVA)
#> List of 4
#>  $ TRCanova     :'data.frame':   8 obs. of  3 variables:
#>   ..$ SS: num [1:8] 0.0236 0.2052 52.5284 0.0151 6.41 ...
#>   ..$ DF: num [1:8] 1 3 99 3 99 297 297 799
```

```
#>   ..$ MS: num [1:8] 0.02357 0.06841 0.53059 0.00502 0.06475 ...
#>  $ VarCom       :'data.frame':   6 obs. of  1 variable:
#>   ..$ Estimates: num [1:6] 3.78e-05 5.13e-02 -7.13e-04 -2.89e-03 2.79e-02 ...
#>  $ IndividualTrt:'data.frame':   3 obs. of  3 variables:
#>   ..$ DF       : num [1:3] 3 99 297
#>   ..$ TrtTREAT1: num [1:3] 0.0493 0.294 0.105
#>   ..$ TrtTREAT2: num [1:3] 0.0242 0.3014 0.1034
#>  $ IndividualRdr:'data.frame':   3 obs. of  5 variables:
#>   ..$ DF         : num [1:3] 1 99 99
#>   ..$ rdrREADER_1: num [1:3] 0.000739 0.203875 0.091559
#>   ..$ rdrREADER_2: num [1:3] 0.0231 0.2234 0.0803
#>   ..$ rdrREADER_3: num [1:3] 0.0148 0.2142 0.0612
#>   ..$ rdrREADER_4: num [1:3] 4.09e-05 2.85e-01 6.06e-02
```

- ANOVA is a list of 4
  - `TRCanova` is a [8x3] dataframe: the treatment-reader-case ANOVA table, see below, where SS is the sum of squares, DF is the denominator degrees of freedom and MS is the mean squares, and T = treatment, R = reader, C = case, TR = treatment-reader, TC = treatment-case, RC = reader-case, TRC = treatment-reader-case.
  - `VarCom` is a [6x1] dataframe: the variance components, see below, where `varR` is the reader variance, `varC` is the case variance, `varTR` is the treatment-reader variance, `varTC` is the treatment-case variance, `varRC` is the reader-case variance, and `varTRC` is the treatment-reader-case variance.
  - `IndividualTrt` is a [3x3] dataframe: the individual treatment variance components averaged over all readers, see below, where `msR` is the mean square reader, `msC` is the mean square case and `msRC` is the mean square reader-case.
  - `IndividualRdr` is a [3x5] dataframe: the individual reader variance components averaged over treatments, see below, where `msT` is the mean square treatment, `msC` is the mean square case and `msTC` is the mean square treatment-case.

```
ret$ANOVA$TRCanova
#>                   SS  DF           MS
#> T       0.023565410   1 0.0235654097
#> R       0.205217999   3 0.0684059998
#> C      52.528398680  99 0.5305898857
#> TR      0.015060792   3 0.0050202641
#> TC      6.410048814  99 0.0647479678
#> RC     39.242953812 297 0.1321311576
#> TRC    22.660077641 297 0.0762965577
```

```
#> Total 121.085323149 799           NA
ret$ANOVA$VarCom
#>             Estimates
#> VarR    3.7755679e-05
#> VarC    5.1250915e-02
#> VarTR  -7.1276294e-04
#> VarTC  -2.8871475e-03
#> VarRC   2.7917300e-02
#> VarErr  7.6296558e-02
ret$ANOVA$IndividualTrt
#>       DF   TrtTREAT1   TrtTREAT2
#> msR    3 0.049266349 0.024159915
#> msC   99 0.293967531 0.301370323
#> msRC 297 0.105047872 0.103379843
ret$ANOVA$IndividualRdr
#>      DF   rdrREADER_1 rdrREADER_2 rdrREADER_3   rdrREADER_4
#> msT   1 0.00073897606 0.023077021 0.014769293 0.00004091217
#> msC  99 0.20387477465 0.223441908 0.214246773 0.28541990211
#> msTC 99 0.09155873437 0.080279256 0.061228980 0.06057067104
```

```
str(ret$RRRC)
#> List of 3
#>  $ FTests         :'data.frame': 2 obs. of  4 variables:
#>   ..$ DF   : num [1:2] 1 3
#>   ..$ MS   : num [1:2] 0.02357 0.00502
#>   ..$ FStat: num [1:2] 4.69 NA
#>   ..$ p    : num [1:2] 0.119 NA
#>  $ ciDiffTrt      :'data.frame': 1 obs. of  7 variables:
#>   ..$ Estimate: num 0.0109
#>   ..$ StdErr  : num 0.00501
#>   ..$ DF      : num 3
#>   ..$ t       : num 2.17
#>   ..$ PrGTt   : num 0.119
#>   ..$ CILower : num -0.00509
#>   ..$ CIUpper : num 0.0268
#>  $ ciAvgRdrEachTrt:'data.frame': 2 obs. of  5 variables:
#>   ..$ Estimate: num [1:2] 0.848 0.837
#>   ..$ StdErr  : num [1:2] 0.0244 0.0236
#>   ..$ DF      : num [1:2] 70.1 253.6
#>   ..$ CILower : num [1:2] 0.799 0.79
#>   ..$ CIUpper : num [1:2] 0.896 0.883
```

- RRRC, a list of 3 containing results of random-reader random-case analyses

- `FTtests`: is a [2x4] dataframe: results of the F-tests, see below, where `FStat` is the F-statistic and `p` is the p-value. The first row is the treatment effect and the second is the error term.
- `ciDiffTrt`: is a [1x7] dataframe: the confidence intervals between different-treatments, see below, where `StdErr` is the standard error of the estimate, `t` is the t-statistic and `PrGTt` is the p-value.
- `ciAvgRdrEachTrt`: is a [2x5] dataframe: the confidence intervals for each treatment, averaged over all readers in the treatment, see below, where `CILower` is the lower 95% confidence interval and `CIUpper` is the upper 95% confidence interval.

```
ret$RRRC$FTests
#>           DF           MS     FStat          p
#> Treatment  1 0.0235654097 4.6940577 0.11883786
#> Error      3 0.0050202641        NA         NA
ret$RRRC$ciDiffTrt
#>                          Estimate       StdErr DF         t      PrGTt
#> trtTREAT1-trtTREAT2 0.010854817 0.0050101218  3 2.1665774 0.11883786
#>                            CILower     CIUpper
#> trtTREAT1-trtTREAT2 -0.0050896269 0.026799261
ret$RRRC$ciAvgRdrEachTrt
#>              Estimate      StdErr          DF    CILower    CIUpper
#> trtTREAT1 0.84774989 0.024402152  70.121788 0.79908282 0.89641696
#> trtTREAT2 0.83689507 0.023566416 253.644028 0.79048429 0.88330585
```

```
str(ret$FRRC)
#> List of 4
#>  $ FTests          :'data.frame':    2 obs. of  4 variables:
#>   ..$ DF   : num [1:2] 1 99
#>   ..$ MS   : num [1:2] 0.0236 0.0647
#>   ..$ FStat: num [1:2] 0.364 NA
#>   ..$ p    : num [1:2] 0.548 NA
#>  $ ciDiffTrt       :'data.frame':    1 obs. of  7 variables:
#>   ..$ Estimate: num 0.0109
#>   ..$ StdErr  : num 0.018
#>   ..$ DF      : num 99
#>   ..$ t       : num 0.603
#>   ..$ PrGTt   : num 0.548
#>   ..$ CILower : num -0.0248
#>   ..$ CIUpper : num 0.0466
#>  $ ciAvgRdrEachTrt :'data.frame':    2 obs. of  5 variables:
#>   ..$ Estimate: num [1:2] 0.848 0.837
#>   ..$ StdErr  : num [1:2] 0.0271 0.0274
#>   ..$ DF      : num [1:2] 99 99
#>   ..$ CILower : num [1:2] 0.794 0.782
```

```
#>   ..$ CIUpper : num [1:2] 0.902 0.891
#>  $ ciDiffTrtEachRdr:'data.frame':    4 obs. of  7 variables:
#>   ..$ Estimate: num [1:4] 0.003844 0.021483 0.017187 0.000905
#>   ..$ StdErr  : num [1:4] 0.0428 0.0401 0.035 0.0348
#>   ..$ DF      : num [1:4] 99 99 99 99
#>   ..$ t       : num [1:4] 0.0898 0.5362 0.4911 0.026
#>   ..$ PrGTt   : num [1:4] 0.929 0.593 0.624 0.979
#>   ..$ CILower : num [1:4] -0.0811 -0.058 -0.0522 -0.0682
#>   ..$ CIUpper : num [1:4] 0.0888 0.101 0.0866 0.07
```

- FRRC, a list of 4 containing results of fixed-reader random-case analyses
  - `FTtests`: is a [2x4] dataframe: results of the F-tests, see below.
  - `ciDiffTrt`: is a [1x7] dataframe: the confidence intervals between different-treatments, see below.
  - `ciAvgRdrEachTrt`: is a [2x5] dataframe: the confidence intervals for the average reader over each treatment
  - `ciDiffTrtEachRdr`: is a [4x7] dataframe: the confidence intervals for each different-treatment pairing for each reader.

```
ret$FRRC$FTests
#>           DF          MS      FStat          p
#> Treatment  1 0.023565410 0.36395597 0.54769704
#> Error     99 0.064747968         NA         NA
ret$FRRC$ciDiffTrt
#>                          Estimate      StdErr DF          t      PrGTt
#> trtTREAT1-trtTREAT2 0.010854817 0.017992772 99 0.60328764 0.54769704
#>                           CILower    CIUpper
#> trtTREAT1-trtTREAT2 -0.024846746 0.04655638
ret$FRRC$ciAvgRdrEachTrt
#>             Estimate      StdErr DF    CILower    CIUpper
#> trtTREAT1 0.84774989 0.027109386 99 0.79395898 0.90154079
#> trtTREAT2 0.83689507 0.027448603 99 0.78243109 0.89135905
ret$FRRC$ciDiffTrtEachRdr
#>                                          Estimate      StdErr DF           t
#> rdrREADER_1::trtTREAT1-trtTREAT2 0.00384441429 0.042792227 99 0.089839080
#> rdrREADER_2::trtTREAT1-trtTREAT2 0.02148349163 0.040069753 99 0.536152334
#> rdrREADER_3::trtTREAT1-trtTREAT2 0.01718679331 0.034993994 99 0.491135520
#> rdrREADER_4::trtTREAT1-trtTREAT2 0.00090456807 0.034805365 99 0.025989329
#>                                         PrGTt      CILower     CIUpper
#> rdrREADER_1::trtTREAT1-trtTREAT2 0.92859660 -0.081064648 0.088753476
#> rdrREADER_2::trtTREAT1-trtTREAT2 0.59305592 -0.058023592 0.100990575
#> rdrREADER_3::trtTREAT1-trtTREAT2 0.62441761 -0.052248882 0.086622469
#> rdrREADER_4::trtTREAT1-trtTREAT2 0.97931817 -0.068156827 0.069965963
```

```
str(ret$RRFC)
#> List of 3
#>  $ FTests         :'data.frame': 2 obs. of  4 variables:
#>   ..$ DF   : num [1:2] 1 3
#>   ..$ MS   : num [1:2] 0.02357 0.00502
#>   ..$ FStat: num [1:2] 4.69 NA
#>   ..$ p    : num [1:2] 0.119 NA
#>  $ ciDiffTrt      :'data.frame': 1 obs. of  7 variables:
#>   ..$ Estimate: num 0.0109
#>   ..$ StdErr  : num 0.00501
#>   ..$ DF      : num 3
#>   ..$ t       : num 2.17
#>   ..$ PrGTt   : num 0.119
#>   ..$ CILower : num -0.00509
#>   ..$ CIUpper : num 0.0268
#>  $ ciAvgRdrEachTrt:'data.frame': 2 obs. of  5 variables:
#>   ..$ Estimate: num [1:2] 0.848 0.837
#>   ..$ StdErr  : num [1:2] 0.0111 0.00777
#>   ..$ DF      : num [1:2] 3 3
#>   ..$ CILower : num [1:2] 0.812 0.812
#>   ..$ CIUpper : num [1:2] 0.883 0.862
```

- RRFC, a list of 3 containing results of random-reader fixed-case analyses
  - `FTtests`: is a [2x4] dataframe: results of the F-tests, see below.
  - `ciDiffTrt`: is a [1x7] dataframe: the confidence intervals between different-treatments, see below.
  - `ciAvgRdrEachTrt`: is a [2x5] dataframe: the confidence intervals for the average reader over each over each treatment.

```
ret$RRFC$FTests
#>           DF           MS     FStat          p
#> Treatment  1 0.0235654097 4.6940577 0.11883786
#> Error      3 0.0050202641        NA         NA
ret$RRFC$ciDiffTrt
#>                          Estimate       StdErr DF         t      PrGTt
#> trtTREAT1-trtTREAT2 0.010854817 0.0050101218  3 2.1665774 0.11883786
#>                            CILower     CIUpper
#> trtTREAT1-trtTREAT2 -0.0050896269 0.026799261
ret$RRFC$ciAvgRdrEachTrt
#>             Estimate       StdErr DF    CILower    CIUpper
#> trtTREAT1 0.84774989 0.011098012  3 0.81243106 0.88306871
#> trtTREAT2 0.83689507 0.007771730  3 0.81216196 0.86162818
```

6.5 References

Chapter 7

OR analysis text output

7.1 TBA How much finished

90%

7.2 Introduction

This chapter illustrates significance testing using the DBM and OR methods.

7.3 Analyzing the ROC dataset

The only change is to specify `method = "OR"` in the significance testing function.
The same dataset is used as was used in the previous chapter.

```
ret <- StSignificanceTesting(dataset03, FOM = "Wilcoxon", method = "OR")
```

7.4 Explanation of the output

The function returns a list with 5 members.

- `FOMs`: figures of merit, identical to that in the DBM method.
- `ANOVA`: ANOVA tables.
- `RRRC`: random-reader random-case analyses results.
- `FRRC`: fixed-reader random-case analyses results.

- `RRFC`" random-reader fixed-case analyses results.

Let us consider the ones that are different from the DBM method.

- ANOVA is a list of 4
  - `TRanova` is a [3x3] dataframe: the treatment-reader ANOVA table, see below, where SS is the sum of squares, DF is the denominator degrees of freedom and MS is the mean squares, and T = treatment, R = reader, TR = treatment-reader.
  - `VarCom` is a [6x2] dataframe: the variance components, see below, where `varR` is the reader variance, `varTR` is the treatment-reader variance, `Cov1`, `Cov2`,`Cov3` and `Var` are as defined in the OR model. The second column lists the correlations defined in the OR model.
  - `IndividualTrt` is a [2x4] dataframe: the individual treatment mean-squares, variances and Cov_2, averaged over all readers, see below, where `msREachTrt` is the mean square reader, `varEachTrt` is the variance and `cov2EachTrt` is `Cov2EachTrt` in each treatment.
  - `IndividualRdr` is a [2x4] dataframe: the individual reader variance components averaged over treatments, see below, where `msTEachRdr` is the mean square treatment, `varEachRdr` is the variance and `cov1EachRdr` is Cov_1 for each reader.

```
ret$ANOVA$TRanova
#>                SS DF            MS
#> T  0.00023565410  1 2.3565410e-04
#> R  0.00205217999  3 6.8406000e-04
#> TR 0.00015060792  3 5.0202641e-05
ret$ANOVA$VarCom
#>            Estimates       Rhos
#> VarR   2.3319942e-05         NA
#> VarTR -6.8389146e-04         NA
#> Cov1   7.9168215e-04 0.51887172
#> Cov2   4.8363767e-04 0.31697811
#> Cov3   5.1250915e-04 0.33590059
#> Var    1.5257762e-03         NA
ret$ANOVA$IndividualTrt
#>           DF    msREachTrt   varEachTrt   cov2EachTrt
#> trtTREAT1  3 0.00049266349 0.00152277779 0.00047229915
#> trtTREAT2  3 0.00024159915 0.00152877746 0.00049497620
ret$ANOVA$IndividualRdr
#>             DF    msTEachRdr   varEachRdr   cov1EachRdr
#> rdrREADER_1  1 7.3897606e-06 0.00147771675 0.00056158020
#> rdrREADER_2  1 2.3077021e-04 0.00151860580 0.00071581326
```

```
#> rdrREADER_3  1 1.4769293e-04 0.0013773788 0.00076508897
#> rdrREADER_4  1 4.0912170e-07 0.0017299529 0.00112424616
```

- RRRC, a list of 3 containing results of random-reader random-case analyses
  - `FTtests`: is a [2x4] dataframe: results of the F-tests, see below, where `FStat` is the F-statistic and `p` is the p-value. The first row is the treatment effect and the second is the error term.
  - `ciDiffTrt`: is a [1x7] dataframe: the confidence intervals between different treatments, see below, where `StdErr` is the standard error of the estimate, `t` is the t-statistic and `PrGTt` is the p-value.
  - `ciAvgRdrEachTrt`: is a [2x5] dataframe: the confidence intervals for the average reader over each treatment, see below, where `CILower` is the lower 95% confidence interval and `CIUpper` is the upper 95% confidence interval.

```
ret$RRRC$FTests
#>           DF             MS     FStat          p
#> Treatment  1 2.3565410e-04 4.6940577 0.11883786
#> Error      3 5.0202641e-05        NA         NA
ret$RRRC$ciDiffTrt
#>                          Estimate       StdErr DF         t      PrGTt
#> trtTREAT1-trtTREAT2 0.010854817 0.0050101218  3 2.1665774 0.11883786
#>                            CILower     CIUpper
#> trtTREAT1-trtTREAT2 -0.0050896269 0.026799261
ret$RRRC$ciAvgRdrEachTrt
#>             Estimate      StdErr          DF    CILower    CIUpper          Cov2
#> trtTREAT1 0.84774989 0.024402152  70.121788 0.79908282 0.89641696 0.00047229915
#> trtTREAT2 0.83689507 0.023566416 253.644028 0.79048429 0.88330585 0.00049497620
```

- FRRC, a list of 5 containing results of fixed-reader random-case analyses
  - `FTtests`: is a [2x4] dataframe: results of the chisquare-tests, see below. Here is a difference from DBM: in the OR method for FRRC the denominator degrees of freedom of the F-statistic is infinite, and the test becomes equivalent to a chisquare test with the degrees of freedom equal to $I - 1$, where I is the number of treatments.
  - `ciDiffTrt`: is a [1x6] dataframe: the confidence intervals between different treatments, see below. An additional column lists
  - `ciAvgRdrEachTrt`: is a [2x5] dataframe: the confidence intervals for the average reader over each treatment
  - `ciDiffTrtEachRdr`: is a [4x6] dataframe: the confidence intervals for each different-treatment pairing for each reader.
  - `IndividualRdrVarCov1`: is a [4x2] dataframe: Var and Cov_1 for individual readers.

```
ret$FRRC$FTests
#>                      MS      Chisq DF          p
#> Treatment 0.0002356541 0.32101347  1 0.57099922
#> Error     0.0007340941         NA NA         NA
ret$FRRC$ciDiffTrt
#>                          Estimate      StdErr          z      PrGTz      CILower
#> trtTREAT1-trtTREAT2 0.010854817 0.019158472 0.56658051 0.57099922 -0.026695098
#>                          CIUpper
#> trtTREAT1-trtTREAT2 0.048404732
ret$FRRC$ciAvgRdrEachTrt
#>             Estimate      StdErr DF    CILower    CIUpper
#> trtTREAT1 0.84774989 0.027109386 99 0.79461647 0.90088331
#> trtTREAT2 0.83689507 0.027448603 99 0.78309680 0.89069334
ret$FRRC$ciDiffTrtEachRdr
#>                                              Estimate      StdErr           z
#> rdrREADER_1::trtTREAT1-trtTREAT2 0.00384441429 0.042792227 0.089839080
#> rdrREADER_2::trtTREAT1-trtTREAT2 0.02148349163 0.040069753 0.536152334
#> rdrREADER_3::trtTREAT1-trtTREAT2 0.01718679331 0.034993994 0.491135520
#> rdrREADER_4::trtTREAT1-trtTREAT2 0.00090456807 0.034805365 0.025989329
#>                                         PrGTz      CILower     CIUpper
#> rdrREADER_1::trtTREAT1-trtTREAT2 0.92841509 -0.080026809 0.087715638
#> rdrREADER_2::trtTREAT1-trtTREAT2 0.59185327 -0.057051781 0.100018765
#> rdrREADER_3::trtTREAT1-trtTREAT2 0.62333060 -0.051400174 0.085773761
#> rdrREADER_4::trtTREAT1-trtTREAT2 0.97926585 -0.067312693 0.069121830
ret$FRRC$IndividualRdrVarCov1
#>               varEachRdr   cov1EachRdr
#> rdrREADER_1 0.0014771675 0.00056158020
#> rdrREADER_2 0.0015186058 0.00071581326
#> rdrREADER_3 0.0013773788 0.00076508897
#> rdrREADER_4 0.0017299529 0.00112424616
```

- RRFC, a list of 3 containing results of random-reader fixed-case analyses
  - `FTtests`: is a [2x4] dataframe: results of the F-tests, see below.
  - `ciDiffTrt`: is a [1x7] dataframe: the confidence intervals between different treatments, see below.
  - `ciAvgRdrEachTrt`: is a [2x5] dataframe: the confidence intervals for the average reader over each over each treatment.

```
ret$RRFC$FTests
#>    DF            MS         F          p
#> T   1 2.3565410e-04 4.6940577 0.11883786
#> TR  3 5.0202641e-05        NA         NA
ret$RRFC$ciDiffTrt
#>                          Estimate       StdErr DF         t      PrGTt
```

```
#> trtTREAT1-trtTREAT2 0.010854817 0.0050101218  3 2.1665774 0.11883786
#>                            CILower     CIUpper
#> trtTREAT1-trtTREAT2 -0.0050896269 0.026799261
ret$RRFC$ciAvgRdrEachTrt
#>             Estimate      StdErr DF    CILower    CIUpper
#> TrtTREAT1 0.84774989 0.011098012  3 0.81243106 0.88306871
#> TrtTREAT2 0.83689507 0.007771730  3 0.81216196 0.86162818
```

7.5 References

Chapter 8

OR analysis Excel output

8.1 TBA How much finished

90%

8.2 Introduction

This chapter illustrates significance testing using the OR method. But, instead of the perhaps unwieldy output in Chapter 7, it generates an Excel output file containing the following worksheets:

- `Summary`
- `FOMs`
- `ANOVA`
- `RRRC`
- `FRRC`
- `RRFC`

8.3 Generating the Excel output file

This illustrates the `UtilOutputReport()` function. The arguments are the embedded dataset, `dataset03`, the same dataset as in the previous two chapters, the report file base name `ReportFileBaseName` is set to `R/quick-start/MyResults`, the report file extension `ReportFileExt` is set to `xlsx`, the `FOM` is set to "Wilcoxon", the `method` of analysis is set to "OR", and the flag `overWrite = TRUE` overwrites any existing file with the same name, as otherwise the program will pause for user input.

```
ret <- UtilOutputReport(get("dataset03"),
                        ReportFileBaseName = "R/quick-start/MyResults",
                        ReportFileExt = "xlsx",
                        FOM = "Wilcoxon",
                        method = "OR",
                        overWrite = TRUE)
```

The following screen shots display the contents of the created file `"R/quick-start/MyResults.xlsx"`.

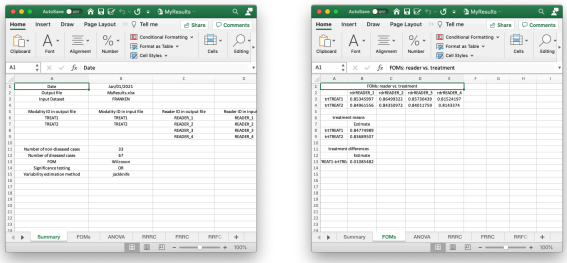


Figure 8.1: 'Summary' and 'FOMs' worksheets of Excel file 'R/quick-start/MyResults.xlsx'

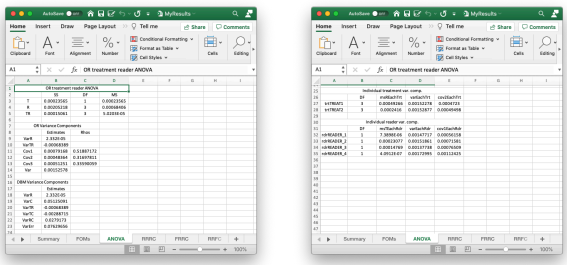


Figure 8.2: 'ANOVA' worksheet of Excel file 'R/quick-start/MyResults.xlsx'

8.4 References

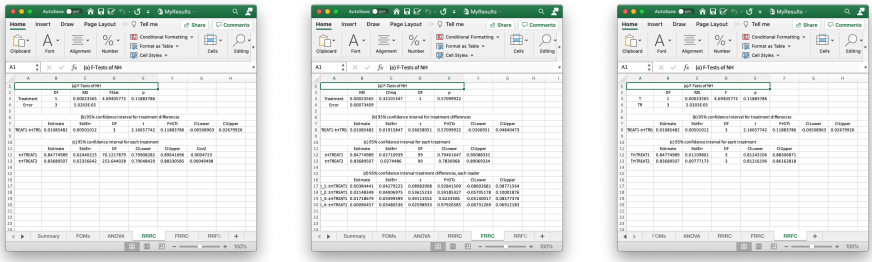


Figure 8.3: ‘RRRC‘, ‘FRRC‘ and ‘RRFC‘ worksheets of Excel file ‘R/quick-start/MyResults.xlsx‘

Chapter 9

DBM method background

9.1 TBA How much finished

80%

9.2 Introduction

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

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

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

9.2.1 Historical background

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

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

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

9.2.2 The Wagner analogy

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

9.2.3 The shortage of numbers to analyze and a pivotal breakthrough

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

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

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

9.2.4 Organization of chapter

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

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

9.3 Random and fixed factors

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

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

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

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

9.4 Reader and case populations

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

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

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

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

9.5 Three types of analyses

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

9.6 General approach

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

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

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

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

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

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

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

9.7 Summary TBA

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

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

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

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

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

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

9.8 References

Chapter 10

Significance Testing using the DBM Method

10.1 TBA How much finished

60%

10.2 The DBM sampling model

DBM = Dorfman Berbaum Metz

The figure-of-merit has three indices:

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

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

$$Y'_{ijk} = K\theta_{ij} - (K-1)\theta_{ij(k)} \tag{10.1}$$

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 (10.1) 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*):

$$Y_{ijk} = Y'_{ijk} + \left(\theta_{ij} - Y'_{ij\bullet}\right) \tag{10.2}$$

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:

$$Y_{ij\bullet} = Y'_{ij\bullet} + \left(\theta_{ij} - Y'_{ij\bullet}\right) = \theta_{ij} \tag{10.3}$$

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

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

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

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

$$Y_{n(ijk)} = \mu + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + (\tau RC)_{ijk} + \epsilon_{n(ijk)} \tag{10.4}$$

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

$$\sum_{i=1}^{I} \tau_i = 0 \Rightarrow \tau_\bullet = 0 \tag{10.5}$$

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:

$$Y_{ijk} \equiv Y_{1(ijk)}\tau_i = Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet} \tag{10.6}$$

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

10.2.1 Explanation of terms in the model

The right hand side of Eqn. (10.1) 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:

$$R_j, C_k, (\tau R)_{ij}, (\tau C)_{ik}, (RC)_{jk}, (\tau RC)_{ijk}, \epsilon_{ijk} \tag{10.7}$$

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

$$\left.\begin{aligned} R_j &\sim N\left(0,\sigma_R^2\right) \\ C_k &\sim N\left(0,\sigma_C^2\right) \\ (\tau R)_{ij} &\sim N\left(0,\sigma_{\tau R}^2\right) \\ (\tau C)_{ik} &\sim N\left(0,\sigma_{\tau C}^2\right) \\ (RC)_{jk} &\sim N\left(0,\sigma_{RC}^2\right) \\ (\tau RC)_{ijk} &\sim N\left(0,\sigma_{\tau RC}^2\right) \\ \epsilon_{ijk} &\sim N\left(0,\sigma_\epsilon^2\right) \end{aligned}\right\} \tag{10.8}$$

Equation (10.8) defines the meanings of the variance components appearing in Equation (10.7). 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 (10.8).

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

The variances defined in (10.8) 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 (10.1). Its meaning will be explored later. For now, *a pseudovalue variance is distinct from a FOM variance*.]
- The term σ_C^2 is the variance of cases that is independent of treatment or reader.
- The term $\sigma_{\tau R}^2$ is the treatment-dependent variance of readers that was excluded in the definition of σ_R^2. If one were to sample readers and treatments for the same case-set, the net variance would be $\sigma_R^2 + \sigma_{\tau R}^2 + \sigma_\epsilon^2$.

- The term $\sigma^2_{\tau C}$ is the treatment-dependent variance of cases that was excluded in the definition of σ^2_C. So, if one were to sample cases and treatments for the same readers, the net variance would be $\sigma^2_C + \sigma^2_{\tau C} + \sigma^2_\epsilon$.
- The term σ^2_{RC} is the treatment-independent variance of readers and cases that were excluded in the definitions of σ^2_R and σ^2_C. So, if one were to sample readers and cases for the same treatment, the net variance would be $\sigma^2_R + \sigma^2_C + \sigma^2_{RC} + \sigma^2_\epsilon$.
- The term $\sigma^2_{\tau RC}$ is the variance of treatments, readers and cases that were excluded in the definitions of all the preceding terms in TBA (10.1). So, if one were to sample treatments, readers and cases the net variance would be $\sigma^2_R + \sigma^2_C + \sigma^2_{\tau C} + \sigma^2_{RC} + \sigma^2_{\tau RC} + \sigma^2_\epsilon$.
- The last term, σ^2_ϵ describes the variance arising from different replications of the study using the same treatments, readers and cases. Measuring this variance requires repeating the study several (N) times with the same treatments, readers and cases, and computing the variance of $Y_{n(ijk)}$, where the additional n-index refers to true replications, $n = 1, 2, \ldots, N$.

$$\sigma^2_\epsilon = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{k}\frac{1}{N-1}\sum_{n=1}^{N}\left(Y_{n(ijk)} - Y_{\bullet(ijk)}\right)^2 \tag{10.9}$$

The right hand side of TBA (10.1) 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 (10.1) 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 (10.1), it being understood that for the random effects, one estimates the variances of the zero-mean normal distributions, TBA (10.1)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 (10.1)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 (10.1) 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.

10.2.3 Definitions of mean-squares

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

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

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

$$\left.\begin{aligned}
\text{MST} &= \frac{JK\sum_{i=1}^{I}(Y_{i\bullet\bullet}-Y_{\bullet\bullet\bullet})^2}{I-1} \\
\text{MSR} &= \frac{IK\sum_{j=1}^{J}(Y_{\bullet j\bullet}-Y_{\bullet\bullet\bullet})^2}{J-1} \\
\text{MS(C)} &= \frac{IJ\sum_{k=1}^{K}(Y_{\bullet\bullet k}-Y_{\bullet\bullet\bullet})^2}{K-1} \\
\text{MSTR} &= \frac{K\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij\bullet}-Y_{i\bullet\bullet}-Y_{\bullet j\bullet}+Y_{\bullet\bullet\bullet})^2}{(I-1)(J-1)} \\
\text{MSTC} &= \frac{J\sum_{i=1}^{I}\sum_{k=1}^{K}(Y_{i\bullet k}-Y_{i\bullet\bullet}-Y_{\bullet\bullet k}+Y_{\bullet\bullet\bullet})^2}{(I-1)(K-1)} \\
\text{MSRC} &= \frac{I\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{\bullet jk}-Y_{\bullet j\bullet}-Y_{\bullet\bullet k}+Y_{\bullet\bullet\bullet})^2}{(J-1)(K-1)} \\
\text{MSTRC} &= \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk}-Y_{ij\bullet}-Y_{i\bullet k}-Y_{\bullet jk}+Y_{i\bullet\bullet}+Y_{\bullet j\bullet}+Y_{\bullet\bullet k}-Y_{\bullet\bullet\bullet})^2}{(I-1)(J-1)K-1)}
\end{aligned}\right\} \quad (10.10)$$

Note the absence of MSE, corresponding to the ϵ term on the right hand side of (10.10). With only one observation per treatment-reader-case combination, MSE cannot be estimated; it effectively gets absorbed into the $MSTRC$ term.

10.3 Expected values of mean squares

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

The *observed* mean squares defined in Equation (10.10) 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 (10.10). Assuming no replications, i.e., $N = 1$, the expected mean squares are as follows, Table Table 10.1; understanding how this table is derived, would lead me outside my expertise and the scope of this book; suffice to say that these are *unconstrained* estimates (as summarized in the quotation above) which are different from the *constrained* estimates appearing in the original DBM publication (Dorfman et al., 1992).

Table 10.1: 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 10.1 the following notation is used as a shorthand:

$$\sigma^2_\tau = \frac{1}{I-1} \sum_{i=1}^{I} \left(Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet}\right)^2 \tag{10.11}$$

Since treatment is a fixed effect, the variance symbol σ_τ^2, which is used for notational consistency in Table 10.1, 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.

10.4 Random-reader random-case (RRRC) analysis

Both readers and cases are regarded as random factors. The expected mean squares in Table Table 10.1 are variance-like quantities; specifically, they are weighted linear combinations of the variances appearing in (10.8). 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.

10.4.1 Calculation of mean squares: an example

- We choose `dataset02` to illustrate calculation of mean squares for pseudovalues. This is referred to in the book as the "VD" dataset (Van Dyke et al., 1993). It consists of 114 cases, 45 of which are diseased, interpreted in two treatments by five radiologists using the ROC paradigm.
- The first line computes the pseudovalues using the `RJafroc` function `UtilPseudoValues()`, and the second line extracts the numbers of treatments, readers and cases. The following lines calculate, using Equation (10.10) 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
```

10.4.2 Significance testing

If the NH of no treatment effect is true, i.e., if $\sigma_\tau^2 = 0$, then according to Table 10.1 the following holds (the last term in the row labeled T in Table 10.1 drops out):

$$E\left(MST \mid NH\right) = \sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2 + J\sigma_{\tau C}^2 \tag{10.12}$$

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

$$\begin{aligned} & E(MSTR) + E(MSTC) - E(MSTRC) \\ & = (\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2) + (\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2) - (\sigma_\epsilon^2 + \sigma_{\tau RC}^2) \\ & = \sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2 + K\sigma_{\tau R}^2 \\ & = E(MST \mid NH) \end{aligned} \tag{10.13}$$

Therefore, under the NH, the ratio:

$$\frac{E(MST \mid NH)}{E(MSTR) + E(MSTC) - E(MSTRC)} = 1 \tag{10.14}$$

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 (10.10):

$$F_{DBM} = \frac{MST}{MSTR + MSTC - MSTRC} \tag{10.15}$$

F_{DBM} is a realization of a random variable. A non-zero treatment effect, i.e., $\sigma_\tau^2 > 0$, will cause the ratio to be larger than one, because $E(MST)$ will be larger, see row labeled T in Table 10.1. 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. (10.15). This is completely analogous to Chapter 08 where knowledge of the distribution of AUC under the NH enabled testing the null hypothesis that the observed value of AUC equals a pre-specified value.

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

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

Summarizing,

$$\left.\begin{array}{r} F_{DBM|NH} \sim F_{\text{ndf,ddf}} \\ \text{ndf} = I - 1 \end{array}\right\} \tag{10.16}$$

The next topic is estimating ddf.

10.4.3 The Satterthwaite approximation

The denominator of the F-ratio is $MSTR + MSTC - MSTRC$. This is not a *simple* mean square (I am using terminology in the Satterthwaite papers - he means any mean square defined by equations such as in Equation (10.10)). 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 (10.15) has an approximate degree of freedom defined by (this is called the *Satterthwaite's approximation*):

$$ddf_{Sat} = \frac{\left(MSTR + MSTC - MSTRC\right)^2}{\left(\frac{MSTR^2}{(I-1)(J-1)} + \frac{MSTC^2}{(I-1)(K-1)} + \frac{MSTRC^2}{(I-1)(J-1)(K-1)}\right)} \tag{10.17}$$

The subscript Sat is for Satterthwaite. From Equation (10.17) it should be fairly obvious that in general ddf_{Sat} is not an integer. To accommodate possible negative estimates of the denominator of Equation (10.17), the original DBM method (Dorfman et al., 1992) proposed, depending on the signs of $\sigma^2_{\tau R}$ and $\sigma^2_{\tau C}$, four expressions for the F-statistic and corresponding expressions for ddf. Rather than repeat them here, since they have been superseded by the method described below, the interested reader is referred to Eqn. 6 and Eqn. 7 in Reference (Hillis et al., 2008).

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

$$F_{DBM} = \frac{MST}{MSTR + \max\left(MSTC - MSTRC, 0\right)} \tag{10.18}$$

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 (10.18) is distributed as an F-statistic with $\text{ndf} = I - 1$ and denominator degrees of freedom ddf_H defined by:

$$ddf_H = \frac{(MSTR + \max(MSTC - MSTRC, 0))^2}{\text{MSTR}^2}(I-1)(J-1) \qquad (10.19)$$

Summarizing,

$$F_{DBM} \sim F_{\text{ndf},\text{ddf}_H} \text{ndf} = I - 1 \qquad (10.20)$$

Instead of 4 rules, as in the original DBM method, the Hillis modification involves just one rule, summarized by Equations (10.19) through (10.20). 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 $\alpha = 0.05$.

10.4.4 Decision rules, p-value and confidence intervals

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

$$\Pr\left(F \le F_{1-\alpha,\text{ndf},\text{ddf}_H} \mid F \sim F_{\text{ndf},\text{ddf}_H}\right) = 1 - \alpha \qquad (10.21)$$

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 (10.18), 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 (10.1) 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}.

10.4.4.1 p-value of the F-test

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

$$p = \Pr\left(F > F_{DBM} \mid F \sim F_{\text{ndf,ddf}_H}\right) \tag{10.22}$$

10.4.4.2 Confidence intervals for inter-treatment FOM differences

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

$$CI_{1-\alpha} = (Y_{i\bullet\bullet} - Y_{i'\bullet\bullet}) \pm t_{\alpha/2;\text{ddf}_H} \sqrt{\frac{2}{JK}\left(MSTR + \max\left(MSTC - MSTRC, 0\right)\right)} \tag{10.23}$$

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

$$\Pr\left(T > t_{\alpha/2;\text{ddf}_H}\right) = \alpha/2 \tag{10.24}$$

Since centered pseudovalues were used:

$$(Y_{i\bullet\bullet} - Y_{i'\bullet\bullet}) = (\theta_{i\bullet} - \theta_{i'\bullet}) \tag{10.25}$$

Therefore, Equation (10.23) can be rewritten:

$$CI_{1-\alpha} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2;\text{ddf}_H} \sqrt{\frac{2}{JK}\left(MSTR + \max\left(MSTC - MSTRC, 0\right)\right)} \tag{10.26}$$

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

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

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

Line 1 defines α. Line 2 forms a data frame from previously calculated mean-squares. Line 3 calculates the denominator appearing in Equation (10.18). Line 4 computes the observed value of F_{DBM}, namely the ratio of the numerator and denominator in Equation (10.18). 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 (10.22). 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
```

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

Line 1 computes the FOM matrix using function `UtilFigureOfMerit`. The next 9 lines compute the treatment FOM differences. The next line `nDiffs` (for "number of differences") evaluates to 1, as with two treatments, there is only one difference. The next line initializes `CI_DIFF_FOM_RRRC`, which stands for "confidence intervals, FOM differences, for RRRC analysis". The next 8 lines evaluate, using Equation (10.26), 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
```

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

10.5.1 The non-centrality parameter

In the significance-testing procedure just described, the relevant distribution was that of the F-statistic when the NH is true, Equation (10.20). *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 (10.15), is distributed as a *non-central* F-distribution $F_{\text{ndf,ddf}_H,\Delta}$ with *non-centrality parameter* Δ:

$$F_{DBM|AH} \sim F_{\text{ndf},ddf_H,\Delta} \tag{10.27}$$

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 (10.4). 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):

$$\Delta = \frac{JK\sigma_\tau^2}{\sigma_\epsilon^2 + \max(K\sigma_{\tau R}^2, 0) + \max(J\sigma_{\tau C}^2, 0)} \tag{10.28}$$

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

$$d = \theta_{1\bullet} - \theta_{2\bullet} \tag{10.29}$$

For two treatments, since the individual treatment effects must be the negatives of each other (because they sum to zero, see (10.5)), it follows that:

$$\sigma_\tau^2 = \frac{d^2}{2} \tag{10.30}$$

Therefore, for two treatments the numerator of the expression for Δ is $JKd^2/2$. Dividing numerator and denominator of Equation (10.28) by K, one gets the final expression for Δ, as coded in `RJafroc`, namely:

$$\Delta = \frac{Jd^2/2}{\max(\sigma_{\tau R}^2, 0) + (\sigma_\epsilon^2 + \max(J\sigma_{\tau C}^2, 0))/K} \tag{10.31}$$

The variances, σ_τ^2, $\sigma_{\tau R}^2$ and $\sigma_{\tau C}^2$, appearing in Equation (10.31), can be calculated from the observed mean squares using the following equations, see (Hillis and Berbaum, 2004) Eqn. 4,

$$\left.\begin{aligned} \sigma_\epsilon^2 &= \text{MSTRC}^* \\ \sigma_{\tau R}^2 &= \frac{\text{MSTR}^* - \text{MSTRC}^*}{K^*} \\ \sigma_{\tau C}^2 &= \frac{\text{MSTC}^* - \text{MSTRC}^*}{J^*} \end{aligned}\right\} \tag{10.32}$$

- 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 (10.31), are the corresponding numbers for the *planned or pivotal study*.
- The three variances, determined from the pilot study via Equation (10.32), are assumed to apply unchanged to the pivotal study (as they are sample-size independent parameters of the DBM model).

10.5.2 The denominator degrees of freedom

- (The numerator degrees of freedom of the non-central F distribution is always unity.) It remains to calculate the appropriate denominator degrees of freedom for the pivotal study. This is denoted df_2, to distinguish it from ddf_H, where the latter applies to the pilot study as in Equation (10.19).
- The starting point is Equation (10.19) 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 (10.19) can be related to the variances by an equation analogous to Equation (10.32), except that, again, all quantities in it apply to the *pivotal* study (note the absence of asterisks):

$$\left.\begin{aligned} \sigma_{\epsilon}^2 &= MSTRC \\ \sigma_{\tau R}^2 &= \frac{MSTR - MSTRC}{K} \\ \sigma_{\tau C}^2 &= \frac{MSTC - MSTRC}{J} \end{aligned}\right\} \tag{10.33}$$

Substituting from Equation (10.33) into Equation (10.19) 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`:

$$df_2 = \frac{(\max(\sigma_{\tau R}^2, 0) + (\max(J\sigma_{\tau C}^2, 0) + \sigma_{\epsilon}^2)/K)^2}{(\max(\sigma_{\tau R}^2, 0) + \sigma_{\epsilon}^2/K)^2}(J-1) \tag{10.34}$$

10.5.3 Example of sample size estimation, RRRC generalization

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

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

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

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

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

10.6 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 (10.4). The F-statistic for testing the NH and its distribution under the NH is:

$$F = \frac{\text{MST}}{\text{MSTC}} \sim F_{I-1,(I-1)(K-1)} \tag{10.35}$$

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 (10.31) upon setting $\sigma^2_{\tau R} = 0$.

$$\Delta = \frac{Jd^2/2}{(\sigma^2_\epsilon + \max(J\sigma^2_{\tau C}, 0))/K} \tag{10.36}$$

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
```

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

The extension to RRFC generalization is as follows. One sets $\sigma_C^2 = 0$ and $\sigma_{\tau C}^2 = 0$ in the DBM model (10.4). The F-statistic for testing the NH and its distribution under the NH is:

$$F = \frac{\text{MST}}{\text{MSTR}} \sim F_{I-1,(I-1)(J-1)} \tag{10.37}$$

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 (10.31) upon setting $\sigma_{\tau C}^2 = 0$.

$$\Delta = \frac{Jd^2/2}{\max(\sigma_{\tau R}^2, 0) + \sigma_\epsilon^2/K} \tag{10.38}$$

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.

10.8 Summary TBA

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

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

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

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

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

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

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

10.9 Things for me to think about

10.9.1 Expected values of mean squares

Assuming no replications the expected mean squares are as follows, Table Table 10.1; understanding how this table is derived, would lead me outside my expertise and the scope of this book; suffice to say that these are *unconstrained* estimates (as summarized in the quotation above) which are different from the *constrained* estimates appearing in the original DBM publication (Dorfman et al., 1992), Table 9.2; the differences between these two types of estimates is summarized in (Dorfman et al., 1995). For reference, Table 9.3 is the table published in the most recent paper that I am aware of (Hillis, 2014). All three tables are different! **In this chapter I will stick to Table Table 10.1 for the subsequent development.**

Table 10.2: Table 9.1 Unconstrained expected values of mean-squares, as in (Dorfman et al., 1995)

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

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

Table 10.3: 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 10.4: 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

10.10 References

Chapter 11

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).

11.1 TBA How much finished

30%

11.2 Fixed-reader random-case (FRRC) analysis

The model is the same as in Eqn. (10.4) except one sets $\sigma_R^2 = \sigma_{\tau R}^2 = 0$ in Table 10.1. The appropriate test statistic is:

$$\frac{E\left(MST\right)}{E\left(MSTC\right)} = \frac{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2 + JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2} \tag{11.1}$$

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

$$\frac{E\left(MST\right)}{E\left(MSTC\right)} = 1 \tag{11.2}$$

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

$$F_{DBM|R} = \frac{MST}{MSTC} \tag{11.3}$$

The observed value $F_{DBM|R}$ (the Roe-Metz notation (Roe and Metz, 1997a) is used which indicates that the factor appearing to the right of the vertical bar is regarded as fixed) is distributed as an F-statistic with ndf = I–1 and $ddf = (I-1)(K-1)$; the degrees of freedom follow from the rows labeled T and TC in TBA Table Table 10.1. Therefore, the distribution of the observed value is (no Satterthwaite approximation needed this time as both numerator and denominator are simple mean-squares):

$$F_{DBM|R} \sim F_{I-1,(I-1)(K-1)} \tag{11.4}$$

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

$$F_{DBM|R} > F_{1-\alpha,I-1,(I-1)(K-1)} \tag{11.5}$$

The p-value of the test is the probability that a random sample from the F-distribution TBA (10.1) Eqn. (9.39), exceeds the observed value:

$$p = \Pr\left(F > F_{DBM|R} \mid F \sim F_{I-1,(I-1)(K-1)}\right) \tag{11.6}$$

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

$$CI_{1-\alpha} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,(I-1)(K-1)}\sqrt{2\frac{MST}{JK}} \tag{11.7}$$

11.2.1 Single-reader multiple-treatment analysis

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

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

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

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

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

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

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

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

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

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

11.3 Random-reader fixed-case (RRFC) analysis

The model is the same as in TBA (10.1) Eqn. (9.4) except one puts $\sigma_C^2 = \sigma_{\tau C}^2 = 0$ in Table Table 10.1. It follows that:

$$\frac{E(MST)}{E(MSTR)} = \frac{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2 + JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2} \tag{11.8}$$

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

$$\frac{E(MST)}{E(MSTR)} = 1 \tag{11.9}$$

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

$$F_{DBM|C} \sim \frac{MST}{MSTR} \tag{11.10}$$

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 10.1.

$$F_{DBM|C} \sim F_{I-1,(I-1)(J-1))} \tag{11.11}$$

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

$$F_{DBM|C} > F_{1-\alpha,I-1,(I-1)(J-1))} \tag{11.12}$$

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

$$p = \Pr\left(F > F_{DBM|C} \mid F \sim F_{I-1,(I-1)(J-1)}\right) \tag{11.13}$$

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

$$CI_{1-\alpha} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,(I-1)(J-1)}\sqrt{2\frac{MSTR}{JK}} \tag{11.14}$$

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

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

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

11.4 References

Chapter 12

Introduction to the Obuchowski-Rockette method

12.1 TBA How much finished

70%

12.2 Locations of helper functions

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

12.3 Introduction

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

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

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

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

12.4 Single-reader multiple-treatment

12.4.1 Overview

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

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

$$\theta_{i\{c\}} = \mu + \tau_i + \epsilon_{i\{c\}} \tag{12.1}$$

Eqn. (12.1) models the observed figure-of-merit $\theta_{i\{c\}}$ as a constant term μ, a treatment dependent term τ_i (the treatment-effect), and a random term $\epsilon_{i\{c\}}$. The term τ_i has the constraint:

$$\sum_{i=1}^{I} \tau_i = 0 \tag{12.2}$$

The left hand side of Eqn. (12.1) is the figure-of-merit $\theta_i\{c\}$ for treatment i and case-set index $\{c\}$, where c = 1, 2, ..., C denotes different independent case-sets sampled from the population, i.e., different *collections* of K_1 non-diseased and K_2 diseased cases.

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

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

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

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

Eqn. (12.1) 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:

$$\epsilon_{i\{c\}} \sim N_I\left(\vec{0}, \Sigma\right) \tag{12.3}$$

Here N_I is the I-variate normal distribution (i.e., each sample yields I random numbers). For the single-reader model Eqn. (12.1), the covariance matrix has the following structure :

$$\Sigma_{ii'} = Cov\left(\epsilon_{i\{c\}}, \epsilon_{i'\{c\}}\right) = \begin{cases} \text{Var} & (i = i') \\ Cov_1 & (i \neq i') \end{cases} \tag{12.4}$$

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

$$\Sigma = \begin{pmatrix} \text{Var} & Cov_1 & ... & Cov_1 & Cov_1 \\ Cov_1 & \text{Var} & ... & Cov_1 & Cov_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Cov_1 & Cov_1 & ... & \text{Var} & Cov_1 \\ Cov_1 & Cov_1 & ... & Cov_1 & \text{Var} \end{pmatrix} \tag{12.5}$$

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

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

[1]

Some elementary statistical results are presented in the Appendix.

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

12.4.2 Significance testing

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

$$MS(T) = \frac{1}{I-1}\sum_{i=1}^{I}(\theta_i - \theta_\bullet)^2 \tag{12.6}$$

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

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

$$\chi_{1\text{R}} \equiv \frac{(I-1)MS(T)}{\text{Var} - \text{Cov1}} \sim \chi^2_{I-1} \tag{12.7}$$

Eqn. (12.7) is from §5.4 in (Hillis, 2007) with two covariance terms "zeroed out" because they are multiplied by $J-1=0$ (since we are restricting to $J=1$).

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

$$F_{1\text{R}} \equiv \frac{MS(T)}{\text{Var} - \text{Cov1}} \sim F_{I-1,\infty} \tag{12.8}$$

12.4.3 p-value and confidence interval

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

$$p \equiv \Pr(f > F_{1R} \mid f \sim F_{I-1,\infty}) \tag{12.9}$$

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

$$CI_{1-\alpha,1R} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,\infty}\sqrt{2(\text{Var} - \text{Cov1})} \tag{12.10}$$

Comparing Eqn. (12.10) to Eqn. (12.27) shows that the term $\sqrt{2(\text{Var} - \text{Cov1})}$ is the standard error of the inter-treatment FOM difference, whose square root is the standard deviation. The term $t_{\alpha/2,\infty}$ is -1.96. Therefore, the confidence interval is constructed by adding and subtracting 1.96 times the standard deviation of the difference from the central value. [One has probably encountered the rule that a 95% confidence interval is plus or minus two standard deviations from the central value. The "2" comes from rounding up 1.96.]

12.4.4 Null hypothesis validation

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

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

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

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

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

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

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

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

The `seed` variable, set to 201 at line 1, is equivalent to the case sample index c in Eqn. (12.1). Different values of `seed` correspond to different case samples.

Line 2 sets the value of μ to 0.8, the average figure of merit, appearing in Eqn. (12.1).

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

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

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

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

Lines 16-18 computes the *estimates* of the means and covariances. The helper function used `VarCov1_FomInput` (the name stands for Var and Cov_1 using FOM input) is included in the distribution. The locations of helper functions are shown in Section 12.2.

Lines 21-33 performs the NH testing. It starts by setting the counter variable `reject` to zero. A for-loop is set up to repeat 2000 times. For each iteration line 24-28 computes the treatment mean-square `MS_T`. Note the use, at line 25, of the two values of $theta_i$ corresponding to the s-th sample from the multivariate norrmal distribution (at line 14). Line 30 computes the F-statistic - compare to Eqn. (12.8). Line 31 computes the p-values and, if the p-value is less than $\alpha = 0.05$, line 32 increments `reject` by one. The observed NH rejection rate, `alphaObs`, is the final value of `reject` divided by 2000, line 34. For a valid test it is expected to be in the range (0.04, 0.06). The actual value, for the chosen value of `seed`, is 0.0515.

12.4.5 Application 1

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

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

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

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

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

```
## pval =  0.2693389
```

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

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

12.4.6 Application 2

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

```
ds <- DfExtractDataset(dataset04, rdrs = 1, trts = c(4,5))
fom <- as.vector(UtilFigureOfMerit(ds, FOM = "wAFROC"))
fom <- t(fom)
vc <- UtilORVarComponentsFactorial(ds, FOM = "wAFROC")
Cov1 <- vc$IndividualRdr$cov1EachRdr
Var <- vc$IndividualRdr$varEachRdr
msT <- vc$IndividualRdr$msTEachRdr
I <- length(ds$ratings$NL[,1,1,1])
chiObs <- (I-1)*msT/(Var-Cov1)
```

```
10 pval <- pchisq(chiObs,I-1,lower.tail = F)
11 ci <- array(dim = 2)
12 ci[1] <- (fom[1] - fom[2]) +
13   qt(0.025, Inf, lower.tail = F) * sqrt(2*(Var - Cov1))
14 ci[2] <- (fom[1] - fom[2]) -
15   qt(0.025, Inf, lower.tail = F) * sqrt(2*(Var - Cov1))
```

```
## fom =  0.8101333 0.7488
```

```
## fom diff =  0.06133333
```

```
## pval =  0.03189534
```

```
## ci =  0.117357 0.005309652
```

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

12.5 Single-treatment multiple-reader

12.5.1 Overview

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

$$\theta_{j\{c\}} = \mu + R_j + \epsilon_{j\{c\}} \tag{12.11}$$

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

$$\epsilon_{j\{c\}} \sim N_J\left(\vec{0}, \Sigma\right) \tag{12.12}$$

The covariance matrix has the following structure:

$$\Sigma_{jj'} = Cov\left(\epsilon_{j\{c\}}, \epsilon_{j'\{c\}}\right) = \begin{cases} \text{Var} & (j = j') \\ Cov_2 & (j \neq j') \end{cases} \tag{12.13}$$

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

$$\Sigma = \begin{pmatrix} \text{Var} & Cov_2 & ... & Cov_2 & Cov_2 \\ Cov_2 & \text{Var} & ... & Cov_2 & Cov_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Cov_2 & Cov_2 & ... & \text{Var} & Cov_2 \\ Cov_2 & Cov_2 & ... & Cov_2 & \text{Var} \end{pmatrix} \tag{12.14}$$

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

12.5.2 Significance testing

Unlike the seemingly analogous single-reader multiple-treatment case addressed in Section 12.4.2, the single-treatment multiple-reader case is fundamentally different. This is because reader is a *random* factor while treatment, in Section 12.4.2, was a *fixed* factor. This makes it impossible to define a null hypothesis analogous to that with the treatment factor, e.g., $R_1 = R_2$, since reader is modeled as a random sample from a distribution, i.e., $R \sim N(0, \sigma_R^2)$.

12.5.3 Special case

If reader is regarded as a *fixed* factor significance testing between readers can be performed. The analysis presented in Section 12.4.2 is applicable, with the treatment factor replaced by the reader factor. This is appropriate, for example, when comparing two AI (artificial intelligence) algorithms. The two algorithms, each of which qualifies as a reader, are not random samples from a population of AI readers: rather they are two fixed algorithms, in the literal sense.

12.6 Multiple-reader multiple-treatment

The previous sections introduced Obuchowski and Rockette method using single reader and single treatment examples. This section extends it to multiple-readers interpreting a common case-set in multiple-treatments (MRMC). The

extension is, in principle, fairly straightforward. Compared to Eqn. (12.1), one needs an additional j index to denote reader dependence of the figure of merit, and additional terms to model reader and treatment-reader variability, and the error term needs to be modified to account for the additional random reader factor.

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

$$\theta_{ij\{c\}} = \mu + \tau_i + R_j + (\tau R)_{ij} + \epsilon_{ij\{c\}} \tag{12.15}$$

- The fixed treatment effect τ_i is subject to the usual constraint, Eqn. (12.2).
- The first two terms on the right hand side of Eqn. (12.15) have their usual meanings: a constant term μ representing performance averaged over treatments and readers, and a treatment effect τ_i (i = 1,2, ..., I).
- The next two terms are, by assumption, mutually independent random samples specified as follows:
  - R_j denotes the random treatment-independent figure-of-merit contribution of reader j (j = 1,2, ..., J), modeled by a zero-mean normal distribution with variance σ_R^2;
  - $(\tau R)_{ij}$ denotes the treatment-dependent random contribution of reader j in treatment i, modeled as a sample from a zero-mean normal distribution with variance $\sigma_{\tau R}^2$.
- Summarizing:

$$\left.\begin{array}{c} R_j \sim N(0, \sigma_R^2) \\ \tau R \sim N(0, \sigma_{\tau R}^2) \end{array}\right\} \tag{12.16}$$

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

$$\mu = \theta_{\bullet\bullet\{1\}} \tag{12.17}$$

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

$$\tau_i = \theta_{i\bullet\{1\}} - \theta_{\bullet\bullet\{1\}} \tag{12.18}$$

The τ_i estimates obey the constraint Eqn. (12.2). 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. (12.15) is more complex than the corresponding DBM model error term. Obuchowski and Rockette model this term with a multivariate normal distribution with a length (IJ) zero-mean vector and a $(IJ \times IJ)$ dimensional covariance matrix Σ. In other words,

$$\epsilon_{ij\{c\}} \sim N_{IJ}(\vec{0}, \Sigma) \tag{12.19}$$

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

$$Cov(\epsilon_{ij\{c\}}, \epsilon_{i'j'\{c\}}) = \left\{ \begin{array}{l} \text{Var } (i = i', j = j') \\ \text{Cov1 } (i \neq i', j = j') \\ \text{Cov2 } (i = i', j \neq j') \\ \text{Cov3 } (i \neq i', j \neq j') \end{array} \right\} \tag{12.20}$$

Apart from fixed effects, the model implied by Eqn. (12.15) and Eqn. (12.20) contains 6 parameters:

$$\sigma_R^2, \sigma_{\tau R}^2, \text{Var}, \text{Cov1}, \text{Cov2}, \text{Cov3}$$

This is the same number of variance component parameters as in the DBM model, which should not be a surprise since one is modeling the data with equivalent models. The Obuchowski and Rockette model Eqn. (12.15) "looks" simpler because four covariance terms are encapsulated in the ϵ term. As with the singe-reader multiple-treatment model, the covariance matrix is assumed to be independent of treatment or reader.

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

12.6.1 Structure of the covariance matrix

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

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

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

The diagonal elements are evidently those cells where the row and column index-pairs are equal. These are (11,11), (12,12), (13,13), (21,21), (22,22) and (23,23). According to Eqn. (12.20) these cells represent Var.

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

According to Eqn. (12.20) cells with different treatment indices but identical reader indices represent Cov1. As an example, cell (21,11) has the same reader indices, namely reader 1, but different treatment indices, namely 2 and 1, so it is Cov1:

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

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

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

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

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

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

$$\text{Var} = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})^2 \right\rangle_{ij} \quad \epsilon_{ij\{c\}} \sim N_{IJ}(\vec{0}, \Sigma) \tag{12.21}$$

Of course, in practice one would use the bootstrap or the jackknife as a stand-in for the c-index (with the understanding that if the jackknife is used, then a variance inflation factor has to be included on the right hand side of Eqn. (12.21). Notice that the left-hand-side of Eqn. (12.21) lacks treatment or reader indices. This is because implicit in the notation is averaging the observed variances over all treatments and readers, as implied by $\langle\rangle_{ij}$. Likewise, the covariance terms are estimated as follows:

$$Cov = \begin{cases} \text{Cov1} = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})(\theta_{i'j\{c\}} - \theta_{i'j\{\bullet\}}) \right\rangle_{ii',jj} \\ Cov_2 = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})(\theta_{ij'\{c\}} - \theta_{ij'\{\bullet\}}) \right\rangle_{ii,jj'} \\ Cov_3 = \left\langle \frac{1}{C-1} \sum_{c=1}^{C} (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})(\theta_{i'j'\{c\}} - \theta_{i'j'\{\bullet\}}) \right\rangle_{ii',jj'} \end{cases} \tag{12.22}$$

In Eqn. (12.22) *the convention is that primed and unprimed variables are always different.*

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

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

12.6.2 Physical meanings of the covariance terms

The meanings of the different terms follow a similar description to that given in Eqn. 12.6.1. The diagonal term Var is the variance of the figures-of-merit when reader j interprets different case-sets $\{c\}$ in treatment i. Each case-set yields a number $\theta_{ij\{c\}}$ and the variance of the C numbers, averaged over the $I \times J$ treatments and readers, is Var. It captures the total variability due to varying difficulty levels of the case-sets, inter-reader and within-reader variability.

It is easier to see the physical meanings of $\text{Cov1}, Cov_2, Cov_3$ if one starts with the correlations.

- $\rho_{1;ii'jj}$ is the correlation of the figures-of-merit when reader j interprets case-sets in different treatments i, i'. Each case-set, starting with $c = 1$, yields two numbers $\theta_{ij\{1\}}$ and $\theta_{i'j\{1\}}$. The correlation of the two pairs of C-length arrays, averaged over all pairings of different treatments and same readers, is ρ_1. The correlation exists due to the common contribution of the shared reader. When the common variation is large, the two arrays become more correlated and ρ_1 approaches unity. If there is no common variation, the two arrays become independent, and ρ_1 equals zero. Converting from correlation to covariance, see Eqn. (12.28), one has $\text{Cov1} < \text{Var}$.

- $\rho_{2;iijj'}$ 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 $Cov_2 < \text{Cov1} < \text{Var}$.

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

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

$$Cov_3 \leq Cov_2 \leq \text{Cov1} \leq \text{Var} \tag{12.23}$$

12.7 Summary

12.8 Discussion

12.9 Appendix: Covariance and correlation

Some elementary statistical results are reviewed here.

12.9.1 Relation: chisquare and F with infinite ddf

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

$$\Pr(d \leq D_{1-\alpha} \mid d \sim D) = 1-\alpha \tag{12.24}$$

With definition Eqn. (12.24), the $(1-\alpha)$ quantile of the χ^2_{I-1} distribution, i.e., $\chi^2_{1-\alpha,I-1}$, is related to the $(1-\alpha)$ quantile of the $F_{I-1,\infty}$ distribution, i.e., $F_{1-\alpha,I-1,\infty}$, as follows (see Hillis et al., 2005, Eq. 22):

$$\frac{\chi^2_{1-\alpha,I-1}}{I-1} = F_{1-\alpha,I-1,\infty} \tag{12.25}$$

Eqn. (12.25) implies that the $(1-\alpha)$ quantile of the F-distribution with $ndf = (I-1)$ and $ddf = \infty$ equals the $(1-\alpha)$ quantile of the χ^2_{I-1} distribution *divided by* $(I-1)$.

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

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

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

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

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

12.9.2 Definitions of covariance and correlation

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

$$Cov(X,Y) = \frac{\sum_{i=1}^{N}(x_i - x_\bullet)(y_i - y_\bullet)}{N-1} = E(XY) - E(X) - E(Y) \qquad (12.26)$$

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

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

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

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

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

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

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

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

It has the property:

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

12.9.3 Special case when variables have equal variances

Assuming X and Y have the same variance:

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

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

$$\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) - 2Cov(X,Y) = 2(\text{Var} - Cov) \qquad (12.27)$$

The right hand side specializes to the OR single-reader multiple-treatment model where the variances (for different treatments) are equal and likewise the covariances in Eqn. (12.5) are equal) The correlation ρ_1 is defined by (the reason for the subscript 1 on ρ is the same as the reason for the subscript 1 on Cov1, which will be explained later):

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

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

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & & \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ & & & \sigma^2 & \rho_1\sigma^2 \\ & & & & \sigma^2 \end{bmatrix} \qquad (12.28)$$

12.9.4 Estimating the variance-covariance matrix

An unbiased estimate of the covariance matrix Eqn. (12.4) follows from:

$$\Sigma_{ii'} \mid_{ps} = \frac{1}{C-1} \sum_{c=1}^{C} \left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right) \left(\theta_{i'\{c\}} - \theta_{i'\{\bullet\}}\right) \qquad (12.29)$$

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

$$\text{Var}_i \mid_{ps} = \frac{1}{C-1} \sum_{c=1}^{C} \left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)^2 \qquad (12.30)$$

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

Sampling different case-sets, as required by Eqn. (12.29), is unrealistic. In reality one has $C = 1$, i.e., a single dataset. Therefore, direct application of this formula is impossible. However, as seen when this situation was encountered before in (book) Chapter 07, one uses resampling methods to realize, for example, different bootstrap samples, which are resampling-based "stand-ins" for

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

$$\Sigma_{ii'} \mid_{bs} = \frac{1}{B-1} \sum_{b=1}^{B} \left(\theta_{i\{b\}} - \theta_{i\{\bullet\}}\right) \left(\theta_{i'\{b\}} - \theta_{i'\{\bullet\}}\right) \tag{12.31}$$

Eqn. (12.31), the bootstrap method of estimating the covariance matrix, is a direct translation of Eqn. (12.29). Alternatively, one could have used the jackknife FOM values $\theta_{i(k)}$, i.e., the figure of merit with a case k removed, repeated for all k, to estimate the covariance matrix:

$$\Sigma_{ii'} \mid_{jk} = \frac{(K-1)^2}{K} \left[\frac{1}{K-1} \sum_{k=1}^{K} \left(\theta_{i(k)} - \theta_{i(\bullet)}\right) \left(\theta_{i'(k)} - \theta_{i'(\bullet)}\right) \right] \tag{12.32}$$

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

Note the subtle difference in usage of ellipses and parentheses between Eqn. (12.29) and Eqn. (12.32). In the former, the subscript $\{c\}$ denotes a set of K cases while in the latter, (k) denotes the original case set with case k removed, leaving $K-1$ cases. There is a similar subtle difference in usage of ellipses and parentheses between Eqn. (12.31) and Eqn. (12.32). The subscript enclosed in parenthesis, i.e., (k), denotes the FOM with case k removed, while in the bootstrap equation one uses the ellipses (curly brackets) $\{b\}$ to denote the b^{th} bootstrap *case-set*, i.e., a whole set of K_1 non-diseased and K_2 diseased cases, sampled with replacement from the original dataset.

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

12.9.5 The variance inflation factor

In Eqn. (12.32), the expression for the jackknife covariance estimate contains a *variance inflation factor*:

$$\frac{(K-1)^2}{K} \tag{12.33}$$

This factor multiplies the traditional expression for the covariance, shown in square brackets in Eqn. (12.32). It is only needed for the jackknife estimate. The bootstrap and the DeLong estimate, see next, do not require this factor.

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

12.9.6 Meaning of the covariance matrix

With reference to Eqn. (12.5), 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:

$$\sigma_i^2 \equiv Var_i = \frac{1}{C-1}\sum_{c=1}^{C}\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right) \tag{12.34}$$

The process is repeated for all treatments and the I-variance values are averaged. This is the final estimate of Var appearing in Eqn. (12.3).

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:

$$\mathrm{Cov}_{1;ii'} = \frac{1}{C-1}\sum_{c=1}^{C}\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i'\{c\}} - \theta_{i'\{\bullet\}}\right) \tag{12.35}$$

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

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

$$\rho_{1;ii'} = \frac{1}{C-1}\sum_{c=1}^{C}\frac{\left(\theta_{i\{c\}} - \theta_{i\{\bullet\}}\right)\left(\theta_{i'\{c\}} - \theta_{i'\{\bullet\}}\right)}{\sigma_i\sigma_{i'}} \tag{12.36}$$

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

To summarize, the covariance matrix can be estimated using the jackknife or the bootstrap, or, in the special case of the empirical AUC figure of merit, the DeLong method can be used. In (book) Chapter 07, these three methods were described in the context of estimating the *variance* of AUC. Eqn. (12.31) and Eqn. (12.32) extend the jackknife and the bootstrap methods, respectively, to estimating the *covariance* of AUC (whose diagonal elements are the variances estimated in the earlier chapter).

12.9.7 Code illustrating the covariance matrix

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

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

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

The following notation is used in the code below:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

```
##   ORBoot.Chisq ORBoot.ddf ORBoot.P.val
## 1    1.3959859          1   0.23739681
```

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

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

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

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

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

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

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

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

The following shows the corresponding output of `RJafroc`.

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

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

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

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

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

12.9.8.1 Jumping ahead

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

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

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

12.10 References

Chapter 13

Obuchowski Rockette (OR) Analysis

13.1 TBA How much finished

80%

13.2 Introduction

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

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

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

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

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

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

13.3 Random-reader random-case

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

$$NH : \tau_i = 0 \;\; (i = 1, 2, ..., I) \tag{13.1}$$

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

$$F_{ORH} = \frac{MS(T)}{MS(TR) + J \max(\text{Cov2} - \text{Cov3}, 0)} \tag{13.2}$$

Eqn. (13.2) incorporates Hillis' modification of the original OR F-statistic. The modification ensures that the constraint Eqn. (12.23) is always obeyed and also avoids a possibly negative (and hence illegal) F-statistic. The relevant mean squares are defined by (note that these are calculated using *FOM* values, not *pseudovalues*):

$$\left.\begin{aligned} MS(T) &= \tfrac{J}{I-1}\sum_{i=1}^{I}(\theta_{i\bullet}-\theta_{\bullet\bullet})^2 \\ MS(R) &= \tfrac{I}{J-1}\sum_{j=1}^{J}(\theta_{\bullet j}-\theta_{\bullet\bullet})^2 \\ MS(TR) &= \tfrac{1}{(I-1)(J-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}(\theta_{ij}-\theta_{i\bullet}-\theta_{\bullet j}+\theta_{\bullet\bullet}) \end{aligned}\right\} \quad (13.3)$$

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

$$F_{OR} = \frac{MS(T)}{MS(TR) + J(\text{Cov2} - \text{Cov3})} \quad (13.4)$$

Note that Eqn. (13.4) lacks the constraint, subsequently proposed by Hillis, which ensures that the denominator cannot be negative. The following distribution was proposed for the test statistic.

$$F_{OR} \sim F_{\text{ndf,ddf}} \quad (13.5)$$

The original degrees of freedom were defined by:

$$\begin{aligned} \text{ndf} &= I - 1 \\ \text{ddf} &= (I-1) \times (J-1) \end{aligned} \quad (13.6)$$

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

13.3.1 Two anecdotes

- The late Dr. Robert F. Wagner once stated to me (ca. 2001) that the sample-size tables published by Obuchowski (Obuchowski, 1998, 2000), using the version of Eqn. (13.2) with the *ddf* as originally suggested by Obuchowski and Rockette, predicted such high number of readers and cases that he was doubtful about the chances of anyone conducting a practical ROC study!

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

13.3.2 Hillis ddf

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

$$\text{ddf}_H = \frac{[MS(TR) + J \max(\text{Cov2} - \text{Cov3}, 0)]^2}{\frac{[MS(TR)]^2}{(I-1)(J-1)}} \tag{13.7}$$

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

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

If Cov2 $<$ Cov3 (which is the *exact opposite* of the expected ordering), ddf_H reduces to $(I-1) \times (J-1)$, the value originally proposed by Obuchowski and Rockette. With Hillis' proposed changes, under the null hypothesis the observed statistic F_{ORH}, defined in Eqn. (13.2), is distributed as an F-statistic with $\text{ndf} = I - 1$ and *ddf* $= \text{ddf}_H$ degrees of freedom (Hillis et al., 2005; Hillis, 2007; Hillis et al., 2008):

$$F_{ORH} \sim F_{\text{ndf},\text{ddf}_H} \tag{13.8}$$

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

13.3.3 Decision rule, p-value and confidence interval

The critical value of the F-statistic for rejection of the null hypothesis is $F_{1-\alpha,\text{ndf},\text{ddf}_H}$, i.e., that value such that fraction $(1-\alpha)$ of the area under the distribution lies to the left of the critical value. From Eqn. (13.2):

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

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

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

$$p = \Pr(F > F_{ORH} \mid F \sim F_{\text{ndf},\text{ddf}_H}) \tag{13.9}$$

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

$$\begin{aligned} CI_{1-\alpha,RRRC,\theta_{i\bullet}-\theta_{i'\bullet}} =& \theta_{i\bullet} - \theta_{i'\bullet} \\ & \pm t_{\alpha/2,\text{ddf}_H} \sqrt{\tfrac{2}{J}(MS(TR) + J\max(\text{Cov2} - \text{Cov3}, 0))} \end{aligned} \tag{13.10}$$

Define df_i, the degrees of freedom for modality i:

$$\text{df}_i = (\text{MS(R)}_i + J\max(\text{Cov2}_i, 0))^2 / \text{MS(R)}_i^2 * (J-1) \tag{13.11}$$

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

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

$$CI_{1-\alpha,RRRC,\theta_{i\bullet}} = \theta_{i\bullet} \pm t_{\alpha/2,\mathrm{df}_i} \sqrt{\tfrac{1}{J}(\mathrm{MS(R)}_i + J\max(\mathrm{Cov2}_i, 0))} \tag{13.12}$$

13.4 Fixed-reader random-case

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

$$F_{ORH|R} = \frac{MS(T)}{\mathrm{Var} - \mathrm{Cov1} + (J-1)\max(\mathrm{Cov2} - \mathrm{Cov3}, 0)} \tag{13.13}$$

[For $J = 1$, Eqn. (13.13) reduces to Eqn. (12.8), i.e., the single-reader analysis described in the previous chapter.]

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

$$F_{ORH|R} \sim F_{I-1,\infty} \tag{13.14}$$

One can get rid of the infinite denominator degrees of freedom by recognizing, as in the previous chapter, that $(I-1)F_{I-1,\infty}$ is distributed as a χ^2 distribution with $I-1$ degrees of freedom, i.e., as χ^2_{I-1}. Therefore, one has, analogous to Eqn. (12.7),

$$\chi^2_{ORH|R} \equiv (I-1)F_{ORH|R} \sim \chi^2_{I-1} \tag{13.15}$$

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

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

The p-value of the test is the probability that a random sample from the chi-square distribution χ^2_{I-1} exceeds the observed value of the test statistic $\chi^2_{ORH|R}$ statistic defined in Eqn. (13.15):

$$p = \Pr(\chi^2 > \chi^2_{ORH|R} \mid \chi^2 \sim \chi^2_{I-1}) \tag{13.16}$$

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

$$CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,\infty}\sqrt{\tfrac{2}{J}(\text{Var} - \text{Cov1} + (J-1)\max(\text{Cov2} - \text{Cov3}, 0))} \tag{13.17}$$

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

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

Additional confidence intervals are stated below:

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

$$CI_{1-\alpha,FRRC,\theta_{i\bullet}} = \theta_{i\bullet} \pm z_{\alpha/2}\sqrt{\tfrac{1}{J}(\text{Var}_i + (J-1)\max(\text{Cov2}_i, 0)} \tag{13.18}$$

$$CI_{1-\alpha,FRRC,\theta_{ij}-\theta_{i'j}} = (\theta_{ij} - \theta_{i'j}) \pm z_{\alpha/2}\sqrt{2(\text{Var}_j - \text{Cov1}_j)} \tag{13.19}$$

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

13.5 Random-reader fixed-case

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

$$F_{ORH|C} = \frac{MS(T)}{MS(TR)} \tag{13.20}$$

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

$$\left.\begin{array}{rcl} \text{ndf} & = & I-1 \\ \text{ddf} & = & (I-1)(J-1) \\ F_{ORH|C} & \sim & F_{\text{ndf,ddf}} \end{array}\right\} \tag{13.21}$$

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

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

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

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

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

$$CI_{1-\alpha,RRFC,\theta_{i\bullet}-\theta_{i'\bullet}} = (\theta_{i\bullet}-\theta_{i'\bullet}) \pm t_{\alpha/2,(I-1)(J-1)}\sqrt{\tfrac{2}{J}MS(TR)} \tag{13.22}$$

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

$$CI_{1-\alpha,RRFC,\theta_{i\bullet}} = \theta_{i\bullet} \pm t_{\alpha/2,J-1}\sqrt{\tfrac{1}{J}\text{MS(R)}_i} \tag{13.23}$$

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

13.6 Single treatment analysis

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

Chapter 14

Obuchowski Rockette Applications

14.1 TBA How much finished

80%

14.2 Introduction

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

Each analysis involves the following steps:

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

14.3 Hand calculation

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

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

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

14.3.1 Random-Reader Random-Case (RRRC) analysis

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

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

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

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

- The next step is the calculate the NH testing statistic.
- The relevant equation is Eqn. (13.2).
- `vc` contains the values needed in this equation, as follows:
  - MS(T) is in `vc$TRanova["T", "MS"]`, whose value is 0.0047962.
  - MS(TR) is in `vc$TRanova["TR", "MS"]`, whose value is 5.5103062×10^{-4}.
  - `Cov2` is in `vc$VarCom["Cov2", "Estimates"]`, whose value is 3.4407483×10^{-4}.
  - `Cov3` is in `vc$VarCom["Cov3", "Estimates"]`, whose value is 2.3902837×10^{-4}.

Applying Eqn. (13.2) one gets (`den` is the denominator on the right hand side of the referenced equation) and F_ORH_RRRC is the value of the F-statistic:

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

- The F-statistic has numerator degrees of freedom ndf $= I - 1$ and denominator degrees of freedom, `ddf`, to be calculated next.
- From the previous chapter, `ddf` is calculated using Eqn. (13.7)). The numerator of `ddf` is identical to `den^2`, where `den` was calculated in the preceding code block. The implementation follows:

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

- The next step is calculation of the p-value for rejecting the NH
- The relevant equation is Eqn. (13.9) whose implementation follows:

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

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

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

- `trtMeans`contains the reader-averaged figures of merit for each treatment.
- `trtMeanDiffs`contains the reader-averaged difference figure of merit.
- From the previous chapter, the $(1-\alpha)$ confidence interval for $\theta_{1\bullet} - \theta_{2\bullet}$ is given by Eqn. (13.10), in which equation the expression inside the square-root symbol is `2/J*den`.
- α, the significance level of the test, is set to 0.05.
- The implementation follows:

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

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

Calculated next is the confidence interval for the reader-averaged FOM for each treatment, i.e. $CI_{1-\alpha,RRRC,\theta_{i\bullet}}$. The relevant equations are Eqn. (13.11) and Eqn. (13.12). The implementation follows:

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

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

14.3.2 Fixed-Reader Random-Case (FRRC) analysis

- The chi-square statistic is calculated using Eqn. (13.13) and Eqn. (13.15).
- The needed quantities are in `vc`.
- For example, MS(T) is in vc$TRanova["T", "MS"], see above. Likewise for `Cov2` and `Cov3`.
- The remaining needed quantities are:
- `Var` is in `vc$VarCom["Var", "Estimates"]`, whose value is 8.0228827×10^{-4}.
- `Cov1` is in `vc$VarCom["Cov1", "Estimates"]`, whose value is 3.4661371×10^{-4}.
- The degree of freedom is $I - 1$.
- The implementation follows:

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

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

- Since p < 0.05, one has a significant finding.
- Freezing reader variability shows a significant difference between the treatments.
- The downside is that the conclusion applies only to the readers used in the study.
- The next step is to calculate the confidence interval for the reader-averaged FOM difference, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- The relevant equation is Eqn. (13.17), whose implementation follows.

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

- Consistent with the chi-square statistic significant finding, one finds that the treatment difference confidence interval does not include zero.
- The next step is to calculate the confidence interval for the reader-averaged figures of merit for each treatment, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}}$.
- The relevant formula is in Eqn. (13.18), whose implementation follows:

```
stdErr <- vector()
df <- vector()
CI <- array(dim = c(I,2))
ciAvgRdrEachTrt <- data.frame()
for (i in 1:I) {
  df[i] <- K - 1
  stdErr[i] <-
    sqrt((vc$IndividualTrt[i,"varEachTrt"] +
            (J-1)*max(vc$IndividualTrt[i,"cov2EachTrt"],0))/J)
  CI[i, ] <- c(trtMeans[i] + qnorm(alpha/2) * stdErr[i],
               trtMeans[i] + qnorm(1-alpha/2) * stdErr[i])
  rowName <- paste0("trt", modalityID[i])
  ciAvgRdrEachTrt <-
    rbind(ciAvgRdrEachTrt,
          data.frame(Estimate = trtMeans[i],
                     StdErr = stdErr[i],
                     DF = df[i],
                     CILower = CI[i,1],
                     CIUpper = CI[i,2],
                     row.names = rowName,
                     stringsAsFactors = FALSE))
}
print(ciAvgRdrEachTrt, digits = 4)
#>      Estimate  StdErr  DF CILower CIUpper
#> trt0   0.8970 0.02429 113  0.8494  0.9446
#> trt1   0.9408 0.01678 113  0.9080  0.9737
```

- Finally, one calculates confidence intervals for the FOM differences for individual readers, i.e., $CI_{1-\alpha,FRRC,\theta_{ij}-\theta_{i'j}}$.
- The relevant formula is in Eqn. (13.19), whose implementation follows:

```
trtMeanDiffs1 <- array(dim = c(J, choose(I, 2)))
Reader <- array(dim = c(J, choose(I, 2)))
stdErr <- array(dim = c(J, choose(I, 2)))
zStat <- array(dim = c(J, choose(I, 2)))
trDiffNames <- array(dim = c(J, choose(I, 2)))
PrGTz <- array(dim = c(J, choose(I, 2)))
CIReader <- array(dim = c(J, choose(I, 2),2))
ciDiffTrtEachRdr <- data.frame()
for (j in 1:J) {
  Reader[j,] <- rep(readerID[j], choose(I, 2))
  stdErr[j,] <-
    sqrt(
      2 *
```

```
        (vc$IndividualRdr[j,"varEachRdr"] -
           vc$IndividualRdr[j,"cov1EachRdr"]))
  pair <- 1
  for (i in 1:I) {
    if (i == I) break
    for (ip in (i + 1):I) {
      trtMeanDiffs1[j, pair] <- foms[i, j] - foms[ip, j]
      trDiffNames[j,pair] <- diffTRName[pair]
      zStat[j,pair] <- trtMeanDiffs1[j,pair]/stdErr[j,pair]
      PrGTz[j,pair] <-
        2 * pnorm(abs(zStat[j,pair]), lower.tail = FALSE)
      CIReader[j, pair,] <-
        c(trtMeanDiffs1[j,pair] +
            qnorm(alpha/2) * stdErr[j,pair],
          trtMeanDiffs1[j,pair] +
            qnorm(1-alpha/2) * stdErr[j,pair])
      rowName <-
        paste0("rdr", Reader[j,pair], "::", trDiffNames[j, pair])
      ciDiffTrtEachRdr <- rbind(
        ciDiffTrtEachRdr,
        data.frame(Estimate = trtMeanDiffs1[j, pair],
                   StdErr = stdErr[j,pair],
                   z = zStat[j, pair],
                   PrGTz = PrGTz[j, pair],
                   CILower = CIReader[j, pair,1],
                   CIUpper = CIReader[j, pair,2],
                   row.names = rowName,
                   stringsAsFactors = FALSE))
      pair <- pair + 1
    }
  }
}
print(ciDiffTrtEachRdr, digits = 3)
#>                 Estimate StdErr      z  PrGTz CILower  CIUpper
#> rdr0::trt0-trt1  -0.0282 0.0255 -1.105 0.2693 -0.0782  0.02182
#> rdr1::trt0-trt1  -0.0465 0.0263 -1.769 0.0768 -0.0981  0.00501
#> rdr2::trt0-trt1  -0.0179 0.0312 -0.573 0.5668 -0.0790  0.04330
#> rdr3::trt0-trt1  -0.0262 0.0173 -1.518 0.1290 -0.0601  0.00764
#> rdr4::trt0-trt1  -0.1002 0.0441 -2.273 0.0230 -0.1865 -0.01381
```

The notation in the first column shows the reader and the treatment pairing. For example, `rdr1::trt0-trt1` means the FOM difference for reader `rdr1`. Only the fifth reader, i.e., `rdr4`, shows a significant difference between the treatments: the p-value is 0.023001 and the confidence interval also does not include zero. The large FOM difference for this reader, -0.100161, was enough to result in a

significant finding for FRRC analysis. The FOM differences for the other readers are about a factor of 2.1522491 or more smaller than that for this reader.

14.3.3 Random-Reader Fixed-Case (RRFC) analysis

The F-statistic is shown in Eqn. (13.20). This time ndf $= I - 1$ and ddf $= (I-1) \times (J-1)$, the values proposed in the Obuchowski-Rockette paper. The implementation follows:

```
den <- vc$TRanova["TR","MS"]
f <- vc$TRanova["T","MS"]/den
ddf <- ((I - 1) * (J - 1))
p <- 1 - pf(f, I - 1, ddf)
FTests_RRFC <-
  data.frame(DF = c(I-1,(I-1)*(J-1)),
             MS = c(vc$TRanova["T","MS"],vc$TRanova["TR","MS"]),
             F = c(f,NA),  p = c(p,NA),
             row.names = c("T","TR"),
             stringsAsFactors = FALSE)
print(FTests_RRFC, digits = 4)
#>    DF       MS     F       p
#> T   1 0.004796 8.704 0.04196
#> TR  4 0.000551    NA      NA
```

Freezing case variability also results in a significant finding, but the conclusion is only applicable to the specific case set used in the study. Next one calculates confidence intervals for the reader-averaged FOM differences, the relevant formula is in Eqn. (13.22), whose implementation follows.

```
stdErr <- sqrt(2 * den/J)
tStat <- vector()
PrGTt <- vector()
CI <- array(dim = c(choose(I,2), 2))
for (i in 1:choose(I,2)) {
  tStat[i] <- trtMeanDiffs[i]/stdErr
  PrGTt[i] <- 2 *
    pt(abs(tStat[i]), ddf, lower.tail = FALSE)
  CI[i, ] <- c(trtMeanDiffs[i] + qt(alpha/2, ddf) * stdErr,
               trtMeanDiffs[i] + qt(1-alpha/2, ddf) * stdErr)
}
ciDiffTrt_RRFC <-
  data.frame(Estimate = trtMeanDiffs,
             StdErr = rep(stdErr, choose(I, 2)),
             DF = rep(ddf, choose(I, 2)),
```

```
              t = tStat,
              PrGTt = PrGTt,
              CILower = CI[,1],
              CIUpper = CI[,2],
              row.names = diffTRName,
              stringsAsFactors = FALSE)

print(ciDiffTrt_RRFC, digits = 4)
#>           Estimate  StdErr DF     t   PrGTt  CILower  CIUpper
#> trt0-trt1  -0.0438 0.01485  4 -2.95 0.04196 -0.08502 -0.00258
```

- As expected because the overall F-test showed significance, the confidence interval does not include zero (the p-value is identical to that found by the F-test).
- This completes the hand calculations.

14.4 RJafroc: dataset02

The second example shows application of the `RJafroc` package function `StSignificanceTesting()` to `dataset02`. This function encapsulates all formulae discussed previously and accomplishes the analyses with a single function call. It returns an object, denoted `st1` below, that contains all results of the analysis. It is a `list` with the following components:

- `FOMs`, this in turn is a `list` containing the following data frames:
  - `foms`, the individual treatment-reader figures of merit, i.e., θ_{ij},
  - `trtMeans`, the treatment figures of merit averaged over readers, i.e., $\theta_{i\bullet}$,
  - `trtMeanDiffs`, the inter-treatment figures of merit differences averaged over readers, i.e., $\theta_{i\bullet} - \theta_{i'\bullet}$.
- `ANOVA`, a `list` containing the following data frames:
  - `TRanova`, the treatment-reader ANOVA table,
  - `VarCom`, Obuchowski-Rockette variance-covariances and correlations,
  - `IndividualTrt`, the mean-squares, `Var` and `Cov2` calculated over individual treatments,
  - `IndividualRdr`, the mean-squares, `Var` and `Cov1` calculated over individual readers.
- `RRRC`, a `list` containing the following data frames:
  - `FTests`, the results of the F-test,

  - `ciDiffTrt`, the confidence intervals for inter-treatment FOM differences, averaged over readers, denoted $CI_{1-\alpha,RRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$ in the previous chapter,
  - `ciAvgRdrEachTrt`, the confidence intervals for individual treatment FOMs, averaged over readers, denoted $CI_{1-\alpha,RRRC,\theta_{i\bullet}}$ in the previous chapter.

- `FRRC`, a `list` containing the following data frames:

  - `FTests`, the results of the F-tests, which in this case specializes to chi-square tests,
  - `ciDiffTrt`, the confidence intervals for inter-treatment FOM differences, averaged over readers, denoted $CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$ in the previous chapter,
  - `ciAvgRdrEachTrt`, the confidence intervals for individual treatment FOMs, averaged over readers, denoted $CI_{1-\alpha,FRRC,\theta_{i\bullet}}$ in the previous chapter,
  - `ciDiffTrtEachRdr`, the confidence intervals for inter-treatment FOM differences for individual readers, denoted $CI_{1-\alpha,FRRC,\theta_{ij}-\theta_{i'j}}$ in the previous chapter,
  - `IndividualRdrVarCov1`, the individual reader variance-covariances and means squares.

- `RRFC`, a `list` containing the following data frames:

  - `FTests`, the results of the F-tests, which in this case specializes to chi-square tests,
  - `ciDiffTrt`, the confidence intervals for inter-treatment FOM differences, averaged over readers, denoted $CI_{1-\alpha,RRFC,\theta_{i\bullet}-\theta_{i'\bullet}}$ in the previous chapter,
  - `ciAvgRdrEachTrt`, the confidence intervals for indvidual treatment FOMs, averaged over readers, denoted $CI_{1-\alpha,RRFC,\theta_{i\bullet}}$ in the previous chapter.

In the interest of clarity, in the first example using the `RJafroc` package the components of the returned object `st1` are listed separately and described explicitly. In the interest of brevity, in subsequent examples the object is listed in its entirety.

Online help on the `StSignificanceTesting()` function is available:

```
?`StSignificanceTesting`
```

The lower right `RStudio` panel contains the online description. Click on the small up-and-right pointing arrow icon to expand this to a new window.

14.4.1 Random-Reader Random-Case (RRRC) analysis

- Since `analysisOption` is not explicitly specified in the following code, the function `StSignificanceTesting` performs all three analyses: `RRRC`, `FRRC` and `RRFC`.
- Likewise, the significance level of the test, also an argument, `alpha`, defaults to 0.05.
- The code below applies `StSignificanceTesting()` and saves the returned object to `st1`.
- The first member of this object, a `list` named `FOMs`, is then displayed.
- `FOMs` contains three data frames:
  - `FOMS$foms`, the figures of merit for each treatment and reader,
  - `FOMS$trtMeans`, the figures of merit for each treatment averaged over readers, and
  - `FOMS$trtMeanDiffs`, the inter-treatment difference figures of merit averaged over readers. The difference is always the first treatment minus the second, etc., in this example, `trt0` minus `trt1`.

```
st1 <- StSignificanceTesting(dataset02, FOM = "Wilcoxon", method = "OR")
print(st1$FOMs, digits = 4)
#> $foms
#>        rdr0   rdr1   rdr2   rdr3   rdr4
#> trt0 0.9196 0.8588 0.9039 0.9731 0.8298
#> trt1 0.9478 0.9053 0.9217 0.9994 0.9300
#>
#> $trtMeans
#>      Estimate
#> trt0   0.8970
#> trt1   0.9408
#>
#> $trtMeanDiffs
#>           Estimate
#> trt0-trt1  -0.0438
```

- Displayed next are the variance components and mean-squares contained in the `ANOVA list`.
  - `ANOVA$TRanova` contains the treatment-reader ANOVA table, i.e. the sum of squares, the degrees of freedom and the mean-squares, for treatment, reader and treatment-reader factors, i.e., `T`, `R` and `TR`.
  - `ANOVA$VarCom` contains the OR variance components and the correlations.
  - `ANOVA$IndividualTrt` contains the quantities necessary for individual treatment analyses.
  - `ANOVA$IndividualRdr` contains the quantities necessary for individual reader analyses.

```
print(st1$ANOVA, digits = 4)
#> $TRanova
#>          SS DF       MS
#> T  0.004796  1 0.004796
#> R  0.015345  4 0.003836
#> TR 0.002204  4 0.000551
#>
#> $VarCom
#>       Estimates   Rhos
#> VarR  0.0015350     NA
#> VarTR 0.0002004     NA
#> Cov1  0.0003466 0.4320
#> Cov2  0.0003441 0.4289
#> Cov3  0.0002390 0.2979
#> Var   0.0008023     NA
#>
#> $IndividualTrt
#>      DF msREachTrt varEachTrt cov2EachTrt
#> trt0  4   0.003083  0.0010141   0.0004840
#> trt1  4   0.001305  0.0005905   0.0002042
#>
#> $IndividualRdr
#>      DF msTEachRdr varEachRdr cov1EachRdr
#> rdr0  1  0.0003971  0.0006989   3.735e-04
#> rdr1  1  0.0010829  0.0011061   7.602e-04
#> rdr2  1  0.0001597  0.0008423   3.553e-04
#> rdr3  1  0.0003445  0.0001506   1.083e-06
#> rdr4  1  0.0050161  0.0012136   2.430e-04
```

- Displayed next are the results of the RRRC significance test, contained in `st1$RRRC`.

```
print(st1$RRRC$FTests, digits = 4)
#>              DF       MS FStat       p
#> Treatment  1.00 0.004796 4.456 0.05167
#> Error     15.26 0.001076    NA      NA
```

- `st1$RRRC$FTests` contains the results of the F-tests: the degrees of freedom, the mean-squares, the observed value of the F-statistic and the p-value for rejecting the NH, listed separately, where applicable, for the treatment and error terms.
- For example, the treatment mean squares is `st1$RRRC$FTests["Treatment", "MS"]` whose value is 0.00479617.

```
print(st1$RRRC$ciDiffTrt, digits = 3)
#>           Estimate StdErr   DF     t  PrGTt CILower  CIUpper
#> trt0-trt1  -0.0438 0.0207 15.3 -2.11 0.0517  -0.088 0.000359
```

- `st1$RRRC$ciDiffTrt` contains the results of the confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.

```
print(st1$RRRC$ciAvgRdrEachTrt, digits = 4)
#>      Estimate  StdErr    DF CILower CIUpper      Cov2
#> trt0   0.8970 0.03317 12.74  0.8252  0.9689 0.0004840
#> trt1   0.9408 0.02157 12.71  0.8941  0.9875 0.0002042
```

- `st1$RRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}}$.

14.4.2 Fixed-Reader Random-Case (FRRC) analysis

- Displayed next are the results of FRRC analysis, contained in `st1$FRRC`.
- `st1$FRRC$FTests` contains the results of the F-tests: the degrees of freedom, the mean-squares, the observed value of the F-statistic and the p-value for rejecting the NH, listed separately, where applicable, for the treatment and error terms.
- For example, the treatment mean squares is `st1$FRRC$FTests["Treatment", "MS"]` whose value is 0.00479617.

```
print(st1$FRRC$FTests, digits = 4)
#>                   MS Chisq DF       p
#> Treatment 0.0047962 5.476  1 0.01928
#> Error     0.0008759    NA NA      NA
```

- Note that this time the output lists a chi-square distribution observed value, 5.47595324, with degree of freedom $df = I - 1 = 1$.
- The listed mean-squares and the p-value agree with the previously performed hand calculations.
- For FRRC analysis the value of the chi-square statistic is significant and the p-value is smaller than α.

```
print(st1$FRRC$ciDiffTrt, digits = 4)
#>           Estimate  StdErr     z   PrGTz  CILower   CIUpper
#> trt0-trt1  -0.0438 0.01872 -2.34 0.01928 -0.08049 -0.007115
```

- `st1$FRRC$ciDiffTrt` contains confidence intervals for inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- The confidence interval excludes zero, and the p-value, listed under `PrGTz` (for probability greater than `z`) is smaller than 0.05.
- One could be using the t-distribution with infinite degrees of freedom, but this is identical to the normal distribution. Hence the listed value is a `z` statistic, i.e., `z = -0.043800322/0.018717483` = -2.34007543.

```
print(st1$FRRC$ciAvgRdrEachTrt, digits = 4)
#>      Estimate  StdErr  DF CILower CIUpper
#> trt0   0.8970 0.02429 113  0.8494  0.9446
#> trt1   0.9408 0.01678 113  0.9080  0.9737
```

- `st1$FRRC$st1$FRRC$ciAvgRdrEachTrt` contains confidence intervals for individual treatment FOMs, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}}$.

```
print(st1$FRRC$ciDiffTrtEachRdr, digits = 3)
#>                  Estimate StdErr      z  PrGTz CILower  CIUpper
#> rdr0::trt0-trt1  -0.0282 0.0255 -1.105 0.2693 -0.0782  0.02182
#> rdr1::trt0-trt1  -0.0465 0.0263 -1.769 0.0768 -0.0981  0.00501
#> rdr2::trt0-trt1  -0.0179 0.0312 -0.573 0.5668 -0.0790  0.04330
#> rdr3::trt0-trt1  -0.0262 0.0173 -1.518 0.1290 -0.0601  0.00764
#> rdr4::trt0-trt1  -0.1002 0.0441 -2.273 0.0230 -0.1865 -0.01381
```

- `st1$FRRC$st1$FRRC$ciDiffTrtEachRdr` contains confidence intervals for inter-treatment difference FOMs, for each reader, i.e., $CI_{1-\alpha,FRRC,\theta_{ij}-\theta_{i'j}}$.

14.4.3 Random-Reader Fixed-Case (RRFC) analysis

```
print(st1$RRFC$FTests, digits = 4)
#>    DF       MS     F       p
#> T   1 0.004796 8.704 0.04196
#> TR  4 0.000551    NA      NA
```

- `st1$RRFC$FTests` contains results of the F-test: the degrees of freedom, the mean-squares, the observed value of the F-statistic and the p-value for rejecting the NH, listed separately, where applicable, for the treatment and treatment-reader terms. The latter is also termed the "error term".
- For example, the treatment-reader mean squares is `st1$RRFC$FTests["TR", "MS"]` whose value is $5.51030622 \times 10^{-4}$.

```
print(st1$RRFC$ciDiffTrt, digits = 4)
#>           Estimate  StdErr DF     t   PrGTt  CILower  CIUpper
#> trt0-trt1  -0.0438 0.01485  4 -2.95 0.04196 -0.08502 -0.00258
```

- `st1$RRFC$ciDiffTrt` contains confidence intervals for the inter-treatment paired difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}-\theta_{i'\bullet}}$.

```
print(st1$RRFC$ciAvgRdrEachTrt, digits = 4)
#>      Estimate  StdErr DF CILower CIUpper
#> Trt0   0.8970 0.02483  4  0.8281  0.9660
#> Trt1   0.9408 0.01615  4  0.8960  0.9857
```

- `st1$RRFC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}}$.

14.5 RJafroc: dataset04

- The third example uses the Federica Zanca dataset (Zanca et al., 2009), i.e., `dataset04`, which has five modalities and four readers.
- It illustrates the situation when multiple treatment pairings are involved. In contrast, the previous example had only one treatment pairing.
- Since this is an FROC dataset, in order to keep it comparable with the previous example, one converts it to an inferred-ROC dataset.
- The function `DfFroc2Roc(dataset04)` converts, using the highest-rating, the FROC dataset to an inferred-ROC dataset.
- The results are contained in `st2`.
- As noted earlier, this time the object is listed in its entirety.

```
ds <- DfFroc2Roc(dataset04) # convert to ROC
I <- length(ds$ratings$NL[,1,1,1])
J <- length(ds$ratings$NL[1,,1,1])
cat("I = ", I, ", J = ", J, "\n")
#> I =  5 , J =  4
st2 <- StSignificanceTesting(ds, FOM = "Wilcoxon", method = "OR")
print(st2, digits = 3)
#> $FOMs
#> $FOMs$foms
#>       rdr1  rdr2  rdr3  rdr4
#> trt1 0.904 0.798 0.812 0.866
#> trt2 0.864 0.845 0.821 0.872
#> trt3 0.813 0.816 0.753 0.857
#> trt4 0.902 0.832 0.789 0.880
```

```
#> trt5 0.841 0.773 0.771 0.848
#>
#> $FOMs$trtMeans
#>      Estimate
#> trt1    0.845
#> trt2    0.850
#> trt3    0.810
#> trt4    0.851
#> trt5    0.808
#>
#> $FOMs$trtMeanDiffs
#>            Estimate
#> trt1-trt2 -0.005100
#> trt1-trt3  0.035325
#> trt1-trt4 -0.005412
#> trt1-trt5  0.036775
#> trt2-trt3  0.040425
#> trt2-trt4 -0.000312
#> trt2-trt5  0.041875
#> trt3-trt4 -0.040737
#> trt3-trt5  0.001450
#> trt4-trt5  0.042187
#>
#>
#> $ANOVA
#> $ANOVA$TRanova
#>         SS DF       MS
#> T  0.00759  4 0.001897
#> R  0.02188  3 0.007294
#> TR 0.00555 12 0.000462
#>
#> $ANOVA$VarCom
#>       Estimates  Rhos
#> VarR   1.28e-03    NA
#> VarTR -1.09e-05    NA
#> Cov1   2.95e-04 0.374
#> Cov2   2.33e-04 0.296
#> Cov3   2.12e-04 0.269
#> Var    7.89e-04    NA
#>
#> $ANOVA$IndividualTrt
#>      DF msREachTrt varEachTrt cov2EachTrt
#> trt1  3   0.002422   0.000711    0.000211
#> trt2  3   0.000523   0.000751    0.000266
#> trt3  3   0.001855   0.000876    0.000246
```

```
#> trt4  3   0.002578   0.000727    0.000220
#> trt5  3   0.001766   0.000882    0.000222
#>
#> $ANOVA$IndividualRdr
#>      DF msTEachRdr varEachRdr cov1EachRdr
#> rdr1  4   0.001551   0.000689    0.000215
#> rdr2  4   0.000794   0.000824    0.000346
#> rdr3  4   0.000786   0.001009    0.000354
#> rdr4  4   0.000153   0.000635    0.000265
#>
#>
#> $RRRC
#> $RRRC$FTests
#>             DF       MS FStat      p
#> Treatment  4.0 0.001897  3.47 0.0305
#> Error     16.8 0.000547    NA     NA
#>
#> $RRRC$ciDiffTrt
#>            Estimate StdErr   DF       t  PrGTt   CILower  CIUpper
#> trt1-trt2 -0.005100 0.0165 16.8 -0.3084 0.7616 -0.040021  0.02982
#> trt1-trt3  0.035325 0.0165 16.8  2.1361 0.0477  0.000404  0.07025
#> trt1-trt4 -0.005412 0.0165 16.8 -0.3273 0.7475 -0.040334  0.02951
#> trt1-trt5  0.036775 0.0165 16.8  2.2238 0.0402  0.001854  0.07170
#> trt2-trt3  0.040425 0.0165 16.8  2.4445 0.0258  0.005504  0.07535
#> trt2-trt4 -0.000312 0.0165 16.8 -0.0189 0.9851 -0.035234  0.03461
#> trt2-trt5  0.041875 0.0165 16.8  2.5322 0.0216  0.006954  0.07680
#> trt3-trt4 -0.040737 0.0165 16.8 -2.4634 0.0249 -0.075659 -0.00582
#> trt3-trt5  0.001450 0.0165 16.8  0.0877 0.9312 -0.033471  0.03637
#> trt4-trt5  0.042187 0.0165 16.8  2.5511 0.0208  0.007266  0.07711
#>
#> $RRRC$ciAvgRdrEachTrt
#>      Estimate StdErr    DF CILower CIUpper     Cov2
#> trt1    0.845 0.0286  5.46   0.774   0.917 0.000211
#> trt2    0.850 0.0199 27.72   0.809   0.891 0.000266
#> trt3    0.810 0.0266  7.04   0.747   0.873 0.000246
#> trt4    0.851 0.0294  5.40   0.777   0.925 0.000220
#> trt5    0.808 0.0258  6.78   0.747   0.870 0.000222
#>
#>
#> $FRRC
#> $FRRC$FTests
#>                  MS Chisq DF       p
#> Treatment 0.001897  13.6  4 0.00868
#> Error     0.000558    NA NA      NA
#>
```

```
#> $FRRC$ciDiffTrt
#>            Estimate StdErr       z  PrGTz  CILower CIUpper
#> trt1-trt2 -0.005100 0.0167 -0.3054 0.7601 -0.03783  0.0276
#> trt1-trt3  0.035325 0.0167  2.1151 0.0344  0.00259  0.0681
#> trt1-trt4 -0.005412 0.0167 -0.3241 0.7459 -0.03815  0.0273
#> trt1-trt5  0.036775 0.0167  2.2019 0.0277  0.00404  0.0695
#> trt2-trt3  0.040425 0.0167  2.4204 0.0155  0.00769  0.0732
#> trt2-trt4 -0.000312 0.0167 -0.0187 0.9851 -0.03305  0.0324
#> trt2-trt5  0.041875 0.0167  2.5073 0.0122  0.00914  0.0746
#> trt3-trt4 -0.040737 0.0167 -2.4392 0.0147 -0.07347 -0.0080
#> trt3-trt5  0.001450 0.0167  0.0868 0.9308 -0.03128  0.0342
#> trt4-trt5  0.042187 0.0167  2.5260 0.0115  0.00945  0.0749
#>
#> $FRRC$ciAvgRdrEachTrt
#>      Estimate StdErr  DF CILower CIUpper
#> trt1    0.845 0.0183 199   0.809   0.881
#> trt2    0.850 0.0197 199   0.812   0.889
#> trt3    0.810 0.0201 199   0.770   0.849
#> trt4    0.851 0.0186 199   0.814   0.887
#> trt5    0.808 0.0197 199   0.770   0.847
#>
#> $FRRC$ciDiffTrtEachRdr
#>                 Estimate StdErr       z   PrGTz  CILower CIUpper
#> rdr1::trt1-trt2  0.04000 0.0308  1.2989 0.19400 -0.02036  0.1004
#> rdr1::trt1-trt3  0.09130 0.0308  2.9646 0.00303  0.03094  0.1517
#> rdr1::trt1-trt4  0.00190 0.0308  0.0617 0.95081 -0.05846  0.0623
#> rdr1::trt1-trt5  0.06285 0.0308  2.0408 0.04127  0.00249  0.1232
#> rdr1::trt2-trt3  0.05130 0.0308  1.6658 0.09576 -0.00906  0.1117
#> rdr1::trt2-trt4 -0.03810 0.0308 -1.2372 0.21603 -0.09846  0.0223
#> rdr1::trt2-trt5  0.02285 0.0308  0.7420 0.45811 -0.03751  0.0832
#> rdr1::trt3-trt4 -0.08940 0.0308 -2.9029 0.00370 -0.14976 -0.0290
#> rdr1::trt3-trt5 -0.02845 0.0308 -0.9238 0.35559 -0.08881  0.0319
#> rdr1::trt4-trt5  0.06095 0.0308  1.9791 0.04780  0.00059  0.1213
#> rdr2::trt1-trt2 -0.04650 0.0309 -1.5039 0.13260 -0.10710  0.0141
#> rdr2::trt1-trt3 -0.01815 0.0309 -0.5870 0.55719 -0.07875  0.0424
#> rdr2::trt1-trt4 -0.03330 0.0309 -1.0770 0.28147 -0.09390  0.0273
#> rdr2::trt1-trt5  0.02520 0.0309  0.8150 0.41505 -0.03540  0.0858
#> rdr2::trt2-trt3  0.02835 0.0309  0.9169 0.35918 -0.03225  0.0889
#> rdr2::trt2-trt4  0.01320 0.0309  0.4269 0.66943 -0.04740  0.0738
#> rdr2::trt2-trt5  0.07170 0.0309  2.3190 0.02040  0.01110  0.1323
#> rdr2::trt3-trt4 -0.01515 0.0309 -0.4900 0.62414 -0.07575  0.0454
#> rdr2::trt3-trt5  0.04335 0.0309  1.4021 0.16090 -0.01725  0.1039
#> rdr2::trt4-trt5  0.05850 0.0309  1.8921 0.05848 -0.00210  0.1191
#> rdr3::trt1-trt2 -0.00875 0.0362 -0.2418 0.80896 -0.07969  0.0622
#> rdr3::trt1-trt3  0.05900 0.0362  1.6302 0.10307 -0.01194  0.1299
```

```
#> rdr3::trt1-trt4  0.02310 0.0362  0.6383 0.52331 -0.04784  0.0940
#> rdr3::trt1-trt5  0.04060 0.0362  1.1218 0.26196 -0.03034  0.1115
#> rdr3::trt2-trt3  0.06775 0.0362  1.8719 0.06122 -0.00319  0.1387
#> rdr3::trt2-trt4  0.03185 0.0362  0.8800 0.37885 -0.03909  0.1028
#> rdr3::trt2-trt5  0.04935 0.0362  1.3635 0.17271 -0.02159  0.1203
#> rdr3::trt3-trt4 -0.03590 0.0362 -0.9919 0.32124 -0.10684  0.0350
#> rdr3::trt3-trt5 -0.01840 0.0362 -0.5084 0.61118 -0.08934  0.0525
#> rdr3::trt4-trt5  0.01750 0.0362  0.4835 0.62872 -0.05344  0.0884
#> rdr4::trt1-trt2 -0.00515 0.0272 -0.1893 0.84987 -0.05848  0.0482
#> rdr4::trt1-trt3  0.00915 0.0272  0.3363 0.73664 -0.04418  0.0625
#> rdr4::trt1-trt4 -0.01335 0.0272 -0.4907 0.62366 -0.06668  0.0400
#> rdr4::trt1-trt5  0.01845 0.0272  0.6781 0.49770 -0.03488  0.0718
#> rdr4::trt2-trt3  0.01430 0.0272  0.5256 0.59918 -0.03903  0.0676
#> rdr4::trt2-trt4 -0.00820 0.0272 -0.3014 0.76312 -0.06153  0.0451
#> rdr4::trt2-trt5  0.02360 0.0272  0.8674 0.38572 -0.02973  0.0769
#> rdr4::trt3-trt4 -0.02250 0.0272 -0.8270 0.40825 -0.07583  0.0308
#> rdr4::trt3-trt5  0.00930 0.0272  0.3418 0.73249 -0.04403  0.0626
#> rdr4::trt4-trt5  0.03180 0.0272  1.1688 0.24249 -0.02153  0.0851
#>
#> $FRRC$IndividualRdrVarCov1
#>      varEachRdr cov1EachRdr
#> rdr1   0.000689    0.000215
#> rdr2   0.000824    0.000346
#> rdr3   0.001009    0.000354
#> rdr4   0.000635    0.000265
#>
#>
#> $RRFC
#> $RRFC$FTests
#>    DF       MS   F      p
#> T   4 0.001897 4.1 0.0253
#> TR 12 0.000462  NA     NA
#>
#> $RRFC$ciDiffTrt
#>            Estimate StdErr DF       t  PrGTt  CILower  CIUpper
#> trt1-trt2 -0.005100 0.0152 12 -0.3355 0.7431 -0.03822  0.02802
#> trt1-trt3  0.035325 0.0152 12  2.3237 0.0385  0.00220  0.06845
#> trt1-trt4 -0.005412 0.0152 12 -0.3560 0.7280 -0.03854  0.02771
#> trt1-trt5  0.036775 0.0152 12  2.4191 0.0324  0.00365  0.06990
#> trt2-trt3  0.040425 0.0152 12  2.6592 0.0208  0.00730  0.07355
#> trt2-trt4 -0.000312 0.0152 12 -0.0206 0.9839 -0.03344  0.03281
#> trt2-trt5  0.041875 0.0152 12  2.7546 0.0175  0.00875  0.07500
#> trt3-trt4 -0.040737 0.0152 12 -2.6797 0.0200 -0.07386 -0.00761
#> trt3-trt5  0.001450 0.0152 12  0.0954 0.9256 -0.03167  0.03457
#> trt4-trt5  0.042187 0.0152 12  2.7751 0.0168  0.00906  0.07531
```

```
#>
#> $RRFC$ciAvgRdrEachTrt
#>      Estimate StdErr DF CILower CIUpper
#> Trt1    0.845 0.0246  3   0.767   0.923
#> Trt2    0.850 0.0114  3   0.814   0.887
#> Trt3    0.810 0.0215  3   0.741   0.878
#> Trt4    0.851 0.0254  3   0.770   0.931
#> Trt5    0.808 0.0210  3   0.742   0.875
```

14.5.1 Random-Reader Random-Case (RRRC) analysis

- `st2$RRRC$FTests` contains the results of the F-test.
- In this example `ndf` = 4 because there are I = 5 treatments. Since the p-value is less than 0.05, at least one treatment-pairing FOM difference is significantly different from zero.
- `st2$RRRC$ciDiffTrt` contains the confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- With I = 5 treatments there are 10 distinct treatment-pairings.
- Looking at the `PrGTt` (for probability greater than `t`) column, one finds six pairings that are significant: `trt1-trt3`, `trt1-trt5`, etc. The smallest p-value is for the `trt4-trt5` pairing.
- `st2$RRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}}$.
- Looking at the `Estimate` column one confirms that `trt5` has the smallest FOM while `trt4` has the highest.

14.5.2 Fixed-Reader Random-Case (FRRC) analysis

- `st2$FRRC$FTests` contains results of the F-tests, which in this situation is actually a chi-square test of the NH.
- Again, `ndf` = 4 because there are I = 5 treatments. Since the p-value is less than 0.05, at least one treatment-pairing FOM difference is significantly different from zero.
- `st2$FRRC$ciDiffTrt` contains confidence intervals for the inter-treatment paired difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- With I = 5 treatments there are 10 distinct treatment-pairings.

- Looking at the `PrGTt` column, one finds six pairings that are significant: `trt1-trt3`, `trt1-trt5`, etc. The smallest p-value is for the `trt4-trt5` pairing.
- `st2$FRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}}$.
- The `Estimate` column confirms that `trt5` has the smallest FOM while `trt4` has the highest.

14.5.3 Random-Reader Fixed-Case (RRFC) analysis

- `st2$RRFC$FTests` contains the results of the F-test of the NH.
- Again, `ndf` = 4 because there are I = 5 treatments. Since the p-value is less than 0.05, at least one treatment-pairing FOM difference is significantly different from zero.
- `st2$RRFC$ciDiffTrt` contains confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- With I = 5 treatments there are 10 distinct treatment-pairings.
- The `PrGTt` column shows that six pairings are significant: `trt1-trt3`, `trt1-trt5`, etc. The smallest p-value is for the `trt4-trt5` pairing.
- `st2$RRFC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}}$.
- The `Estimate` column confirms that `trt5` has the smallest FOM while `trt4` has the highest (the `Estimates` column is identical for RRRC, FRRC and RRFC analyses).

14.6 RJafroc: dataset04, FROC

- The fourth example uses `dataset04`, but this time we use the FROC data, specifically, we do not convert it to inferred-ROC.
- Since this is an FROC dataset, one needs to use an FROC figure of merit.
- In this example the weighted AFROC figure of merit `FOM = "wAFROC"` is specified. This is the recommended figure of merit when both normal and abnormal cases are present in the dataset.
- If the dataset does not contain normal cases, then the weighted AFROC1 figure of merit `FOM = "wAFROC1"` should be specified.
- The results are contained in `st3`.
- As noted earlier, this time the object is listed in its entirety.

```
ds <- dataset04 # do NOT convert to ROC
FOM <- "wAFROC"
st3 <- StSignificanceTesting(ds, FOM = FOM, method = "OR")
print(st3, digits = 3)
#> $FOMs
#> $FOMs$foms
#>       rdr1  rdr3  rdr4  rdr5
#> trt1 0.779 0.725 0.704 0.805
#> trt2 0.787 0.727 0.723 0.804
#> trt3 0.730 0.716 0.672 0.773
#> trt4 0.810 0.743 0.694 0.829
#> trt5 0.749 0.682 0.655 0.771
#>
#> $FOMs$trtMeans
#>      Estimate
#> trt1    0.753
#> trt2    0.760
#> trt3    0.723
#> trt4    0.769
#> trt5    0.714
#>
#> $FOMs$trtMeanDiffs
#>           Estimate
#> trt1-trt2 -0.00686
#> trt1-trt3  0.03061
#> trt1-trt4 -0.01604
#> trt1-trt5  0.03884
#> trt2-trt3  0.03747
#> trt2-trt4 -0.00918
#> trt2-trt5  0.04570
#> trt3-trt4 -0.04665
#> trt3-trt5  0.00823
#> trt4-trt5  0.05488
#>
#>
#> $ANOVA
#> $ANOVA$TRanova
#>         SS DF      MS
#> T  0.00927  4 0.00232
#> R  0.03540  3 0.01180
#> TR 0.00204 12 0.00017
#>
#> $ANOVA$VarCom
#>       Estimates  Rhos
#> VarR   0.002209    NA
```

```
#> VarTR -0.000305    NA
#> Cov1   0.000422 0.455
#> Cov2   0.000336 0.362
#> Cov3   0.000304 0.328
#> Var    0.000928    NA
#>
#> $ANOVA$IndividualTrt
#>      DF msREachTrt varEachTrt cov2EachTrt
#> trt1  3    0.00221   0.000877    0.000333
#> trt2  3    0.00171   0.000939    0.000380
#> trt3  3    0.00171   0.000970    0.000297
#> trt4  3    0.00386   0.000859    0.000311
#> trt5  3    0.00298   0.000995    0.000359
#>
#> $ANOVA$IndividualRdr
#>      DF msTEachRdr varEachRdr cov1EachRdr
#> rdr1  4   0.001014   0.000883    0.000412
#> rdr3  4   0.000509   0.000897    0.000436
#> rdr4  4   0.000698   0.001171    0.000495
#> rdr5  4   0.000604   0.000762    0.000345
#>
#>
#> $RRRC
#> $RRRC$FTests
#>             DF       MS FStat        p
#> Treatment  4.0 0.002317   7.8 0.000117
#> Error     36.8 0.000297    NA       NA
#>
#> $RRRC$ciDiffTrt
#>           Estimate StdErr   DF      t    PrGTt  CILower  CIUpper
#> trt1-trt2 -0.00686 0.0122 36.8 -0.563 5.77e-01 -0.03155  0.01784
#> trt1-trt3  0.03061 0.0122 36.8  2.512 1.65e-02  0.00592  0.05531
#> trt1-trt4 -0.01604 0.0122 36.8 -1.316 1.96e-01 -0.04073  0.00866
#> trt1-trt5  0.03884 0.0122 36.8  3.188 2.92e-03  0.01415  0.06354
#> trt2-trt3  0.03747 0.0122 36.8  3.075 3.96e-03  0.01278  0.06217
#> trt2-trt4 -0.00918 0.0122 36.8 -0.753 4.56e-01 -0.03387  0.01552
#> trt2-trt5  0.04570 0.0122 36.8  3.750 6.07e-04  0.02100  0.07040
#> trt3-trt4 -0.04665 0.0122 36.8 -3.828 4.85e-04 -0.07135 -0.02195
#> trt3-trt5  0.00823 0.0122 36.8  0.675 5.04e-01 -0.01647  0.03292
#> trt4-trt5  0.05488 0.0122 36.8  4.504 6.52e-05  0.03018  0.07957
#>
#> $RRRC$ciAvgRdrEachTrt
#>      Estimate StdErr    DF CILower CIUpper     Cov2
#> trt1    0.753 0.0298  7.71   0.684   0.822 0.000333
#> trt2    0.760 0.0284 10.69   0.697   0.823 0.000380
```

```
#> trt3    0.723 0.0269  8.62   0.661   0.784 0.000297
#> trt4    0.769 0.0357  5.24   0.679   0.860 0.000311
#> trt5    0.714 0.0333  6.59   0.635   0.794 0.000359
#>
#>
#> $FRRC
#> $FRRC$FTests
#>                 MS Chisq DF       p
#> Treatment 0.002317  15.4  4 0.00393
#> Error     0.000602    NA NA      NA
#>
#> $FRRC$ciDiffTrt
#>           Estimate StdErr      z   PrGTz  CILower CIUpper
#> trt1-trt2 -0.00686 0.0173 -0.395 0.69260 -0.04085  0.0271
#> trt1-trt3  0.03061 0.0173  1.765 0.07753 -0.00338  0.0646
#> trt1-trt4 -0.01604 0.0173 -0.925 0.35518 -0.05003  0.0180
#> trt1-trt5  0.03884 0.0173  2.240 0.02511  0.00485  0.0728
#> trt2-trt3  0.03747 0.0173  2.161 0.03073  0.00348  0.0715
#> trt2-trt4 -0.00918 0.0173 -0.529 0.59662 -0.04317  0.0248
#> trt2-trt5  0.04570 0.0173  2.635 0.00841  0.01171  0.0797
#> trt3-trt4 -0.04665 0.0173 -2.690 0.00715 -0.08064 -0.0127
#> trt3-trt5  0.00823 0.0173  0.474 0.63515 -0.02576  0.0422
#> trt4-trt5  0.05488 0.0173  3.164 0.00155  0.02089  0.0889
#>
#> $FRRC$ciAvgRdrEachTrt
#>      Estimate StdErr  DF CILower CIUpper
#> trt1    0.753 0.0217 199   0.711   0.796
#> trt2    0.760 0.0228 199   0.715   0.805
#> trt3    0.723 0.0216 199   0.680   0.765
#> trt4    0.769 0.0212 199   0.728   0.811
#> trt5    0.714 0.0228 199   0.670   0.759
#>
#> $FRRC$ciDiffTrtEachRdr
#>                  Estimate StdErr       z   PrGTz  CILower   CIUpper
#> rdr1::trt1-trt2 -0.00773 0.0307 -0.2520 0.80105 -0.06788  0.052416
#> rdr1::trt1-trt3  0.04957 0.0307  1.6154 0.10622 -0.01057  0.109724
#> rdr1::trt1-trt4 -0.03087 0.0307 -1.0058 0.31451 -0.09102  0.029282
#> rdr1::trt1-trt5  0.03047 0.0307  0.9928 0.32083 -0.02968  0.090616
#> rdr1::trt2-trt3  0.05731 0.0307  1.8674 0.06185 -0.00284  0.117457
#> rdr1::trt2-trt4 -0.02313 0.0307 -0.7538 0.45097 -0.08328  0.037016
#> rdr1::trt2-trt5  0.03820 0.0307  1.2448 0.21322 -0.02195  0.098349
#> rdr1::trt3-trt4 -0.08044 0.0307 -2.6212 0.00876 -0.14059 -0.020293
#> rdr1::trt3-trt5 -0.01911 0.0307 -0.6226 0.53352 -0.07926  0.041041
#> rdr1::trt4-trt5  0.06133 0.0307  1.9986 0.04566  0.00118  0.121482
#> rdr3::trt1-trt2 -0.00201 0.0304 -0.0661 0.94726 -0.06152  0.057504
```

```
#> rdr3::trt1-trt3  0.00913 0.0304  0.3008 0.76357 -0.05038  0.068646
#> rdr3::trt1-trt4 -0.01822 0.0304 -0.6002 0.54836 -0.07774  0.041287
#> rdr3::trt1-trt5  0.04262 0.0304  1.4035 0.16046 -0.01690  0.102129
#> rdr3::trt2-trt3  0.01114 0.0304  0.3669 0.71367 -0.04837  0.070654
#> rdr3::trt2-trt4 -0.01622 0.0304 -0.5341 0.59329 -0.07573  0.043296
#> rdr3::trt2-trt5  0.04462 0.0304  1.4697 0.14165 -0.01489  0.104137
#> rdr3::trt3-trt4 -0.02736 0.0304 -0.9010 0.36758 -0.08687  0.032154
#> rdr3::trt3-trt5  0.03348 0.0304  1.1027 0.27014 -0.02603  0.092996
#> rdr3::trt4-trt5  0.06084 0.0304  2.0037 0.04510  0.00133  0.120354
#> rdr4::trt1-trt2 -0.01899 0.0368 -0.5166 0.60543 -0.09104  0.053061
#> rdr4::trt1-trt3  0.03132 0.0368  0.8519 0.39429 -0.04074  0.103370
#> rdr4::trt1-trt4  0.00927 0.0368  0.2521 0.80099 -0.06279  0.081320
#> rdr4::trt1-trt5  0.04845 0.0368  1.3179 0.18753 -0.02360  0.120503
#> rdr4::trt2-trt3  0.05031 0.0368  1.3685 0.17116 -0.02174  0.122361
#> rdr4::trt2-trt4  0.02826 0.0368  0.7687 0.44209 -0.04379  0.100311
#> rdr4::trt2-trt5  0.06744 0.0368  1.8345 0.06658 -0.00461  0.139495
#> rdr4::trt3-trt4 -0.02205 0.0368 -0.5998 0.54864 -0.09410  0.050003
#> rdr4::trt3-trt5  0.01713 0.0368  0.4661 0.64118 -0.05492  0.089186
#> rdr4::trt4-trt5  0.03918 0.0368  1.0659 0.28649 -0.03287  0.111236
#> rdr5::trt1-trt2  0.00131 0.0289  0.0453 0.96385 -0.05526  0.057881
#> rdr5::trt1-trt3  0.03243 0.0289  1.1237 0.26116 -0.02414  0.089006
#> rdr5::trt1-trt4 -0.02432 0.0289 -0.8425 0.39953 -0.08089  0.032256
#> rdr5::trt1-trt5  0.03384 0.0289  1.1724 0.24102 -0.02273  0.090414
#> rdr5::trt2-trt3  0.03112 0.0289  1.0783 0.28089 -0.02545  0.087698
#> rdr5::trt2-trt4 -0.02563 0.0289 -0.8878 0.37466 -0.08220  0.030948
#> rdr5::trt2-trt5  0.03253 0.0289  1.1271 0.25969 -0.02404  0.089106
#> rdr5::trt3-trt4 -0.05675 0.0289 -1.9661 0.04929 -0.11332 -0.000177
#> rdr5::trt3-trt5  0.00141 0.0289  0.0488 0.96109 -0.05516  0.057981
#> rdr5::trt4-trt5  0.05816 0.0289  2.0149 0.04391  0.00159  0.114731
#>
#> $FRRC$IndividualRdrVarCov1
#>      varEachRdr cov1EachRdr
#> rdr1   0.000883    0.000412
#> rdr3   0.000897    0.000436
#> rdr4   0.001171    0.000495
#> rdr5   0.000762    0.000345
#>
#>
#> $RRFC
#> $RRFC$FTests
#>    DF      MS    F        p
#> T   4 0.00232 13.7 0.000202
#> TR 12 0.00017   NA       NA
#>
#> $RRFC$ciDiffTrt
```

```
#>           Estimate  StdErr DF      t    PrGTt CILower  CIUpper
#> trt1-trt2 -0.00686 0.00921 12 -0.745 4.71e-01 -0.0269  0.01321
#> trt1-trt3  0.03061 0.00921 12  3.324 6.06e-03  0.0106  0.05068
#> trt1-trt4 -0.01604 0.00921 12 -1.741 1.07e-01 -0.0361  0.00403
#> trt1-trt5  0.03884 0.00921 12  4.218 1.19e-03  0.0188  0.05891
#> trt2-trt3  0.03747 0.00921 12  4.069 1.56e-03  0.0174  0.05754
#> trt2-trt4 -0.00918 0.00921 12 -0.997 3.39e-01 -0.0292  0.01089
#> trt2-trt5  0.04570 0.00921 12  4.963 3.29e-04  0.0256  0.06576
#> trt3-trt4 -0.04665 0.00921 12 -5.066 2.77e-04 -0.0667 -0.02659
#> trt3-trt5  0.00823 0.00921 12  0.894 3.89e-01 -0.0118  0.02829
#> trt4-trt5  0.05488 0.00921 12  5.959 6.62e-05  0.0348  0.07494
#>
#> $RRFC$ciAvgRdrEachTrt
#>      Estimate StdErr DF CILower CIUpper
#> Trt1    0.753 0.0235  3   0.678   0.828
#> Trt2    0.760 0.0207  3   0.694   0.826
#> Trt3    0.723 0.0207  3   0.657   0.788
#> Trt4    0.769 0.0311  3   0.670   0.868
#> Trt5    0.714 0.0273  3   0.627   0.801
```

14.6.1 Random-Reader Random-Case (RRRC) analysis

- `st3$RRRC$FTests` contains the results of the F-tests.
- The p-value is much smaller than that obtained after converting to an ROC dataset. Specifically, for FROC analysis, the p-value is $1.17105004 \times 10^{-4}$ while that for ROC analysis is 0.03054456. The F-statistic and the `ddf` are both larger for FROC analysis, both of of which result in increased probability of rejecting the NH, i.e., FROC analysis has greater power than ROC analysis.
- The increased power of FROC analysis has been confirmed in simulation studies (Chakraborty, 2002).
- `st3$RRRC$ciDiffTrt` contains the confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- With I = 5 treatments there are 10 distinct treatment-pairings.
- Looking at the `PrGTt` (for probability greater than `t`) column, one finds six pairings that are significant: `trt1-trt3`, `trt1-trt5`, etc. The smallest p-value is for the `trt4-trt5` pairing. The findings are consistent with the prior ROC analysis, the difference being the smaller p-values.
- `st3$RRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}}$.

- Looking at the `Estimate` column one confirms that `trt5` has the smallest FOM while `trt4` has the highest (the `Estimates` column is identical for RRRC, FRRC and RRFC analyses).
- `st3$RRRC$st1$RRRC$ciDiffTrtEachRdr` contains confidence intervals for inter-treatment difference FOMs, for each reader, i.e., $CI_{1-\alpha,RRRC,\theta_{ij}-\theta_{i'j}}$.

14.6.2 Fixed-Reader Random-Case (FRRC) analysis

- `st3$FRRC$FTests` contains results of the F-test of the NH.
- Again, `ndf` = 4 because there are I = 5 treatments. Since the p-value is less than 0.05, at least one treatment-pairing FOM difference is significantly different from zero.
- `st3$FRRC$ciDiffTrt` contains the confidence intervals for the inter-treatment paired difference FOMs averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- With I = 5 treatments there are 10 distinct treatment-pairings.
- Looking at the `PrGTt` (for probability greater than `t`) column, one finds six pairings that are significant: `trt1-trt3`, `trt1-trt5`, etc. The smallest p-value is for the `trt4-trt5` pairing. The findings are consistent with the prior ROC analysis, the difference being the smaller p-values.
- `st3$FRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}}$.
- Looking at the `Estimate` column one confirms that `trt5` has the smallest FOM while `trt4` has the highest.
- `st3$FRRC$st1$FRRC$ciDiffTrtEachRdr` contains confidence intervals for inter-treatment difference FOMs, for each reader, i.e., $CI_{1-\alpha,FRRC,\theta_{ij}-\theta_{i'j}}$.

14.6.3 Random-Reader Fixed-Case (RRFC) analysis

- `st3$RRFC$FTests` contains results of the F-test of the NH.
- Again, `ndf` = 4 because there are I = 5 treatments. Since the p-value is less than 0.05, at least one treatment-pairing FOM difference is significantly different from zero.
- `st3$RRFC$ciDiffTrt` contains confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- `st3$RRFC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}}$.

- The `Estimate` column confirms that `trt5` has the smallest FOM while `trt4` has the highest (the `Estimates` column is identical for RRRC, FRRC and RRFC analyses).

14.7 RJafroc: dataset04, FROC/DBM

- The fourth example again uses `dataset04`, i.e., FROC data, *but this time using DBM analysis*.
- The key difference below is in the call to `StSignificanceTesting()` function, where we set `method = "DBM"`.
- Since DBM analysis is pseudovalue based, and the figure of merit is not the empirical AUC under the ROC, one expects to see differences from the previously presented OR analysis, contained in `st3`.

```
st4 <- StSignificanceTesting(ds, FOM = FOM, method = "DBM")
# Note: using DBM analysis
print(st4, digits = 3)
#> $FOMs
#> $FOMs$foms
#>       rdr1  rdr3  rdr4  rdr5
#> trt1 0.779 0.725 0.704 0.805
#> trt2 0.787 0.727 0.723 0.804
#> trt3 0.730 0.716 0.672 0.773
#> trt4 0.810 0.743 0.694 0.829
#> trt5 0.749 0.682 0.655 0.771
#>
#> $FOMs$trtMeans
#>      Estimate
#> trt1    0.753
#> trt2    0.760
#> trt3    0.723
#> trt4    0.769
#> trt5    0.714
#>
#> $FOMs$trtMeanDiffs
#>           Estimate
#> trt1-trt2 -0.00686
#> trt1-trt3  0.03061
#> trt1-trt4 -0.01604
#> trt1-trt5  0.03884
#> trt2-trt3  0.03747
#> trt2-trt4 -0.00918
#> trt2-trt5  0.04570
#> trt3-trt4 -0.04665
```

```
#> trt3-trt5  0.00823
#> trt4-trt5  0.05488
#>
#>
#> $ANOVA
#> $ANOVA$TRCanova
#>             SS   DF     MS
#> T       1.853    4 0.4633
#> R       7.081    3 2.3603
#> C     289.602  199 1.4553
#> TR      0.407   12 0.0339
#> TC     95.772  796 0.1203
#> RC    126.902  597 0.2126
#> TRC   226.479 2388 0.0948
#> Total 748.096 3999     NA
#>
#> $ANOVA$VarCom
#>        Estimates
#> VarR    0.002209
#> VarC    0.060862
#> VarTR  -0.000305
#> VarTC   0.006369
#> VarRC   0.023545
#> VarErr  0.094841
#>
#> $ANOVA$IndividualTrt
#>       DF  Trt1  Trt2  Trt3  Trt4  Trt5
#> msR    3 0.442 0.343 0.342 0.772 0.597
#> msC  199 0.375 0.416 0.372 0.358 0.415
#> msRC 597 0.109 0.112 0.134 0.110 0.127
#>
#> $ANOVA$IndividualRdr
#>       DF   rdr1   rdr3  rdr4   rdr5
#> msT    4 0.2027 0.1019 0.140 0.1208
#> msC  199 0.5064 0.5278 0.630 0.4285
#> msTC 796 0.0942 0.0922 0.135 0.0833
#>
#>
#> $RRRC
#> $RRRC$FTests
#>             DF     MS FStat        p
#> Treatment  4.0 0.4633   7.8 0.000117
#> Error     36.8 0.0594    NA       NA
#>
#> $RRRC$ciDiffTrt
```

```
#>           Estimate StdErr   DF      t    PrGTt  CILower  CIUpper
#> trt1-trt2 -0.00686 0.0122 36.8 -0.563 5.77e-01 -0.03155  0.01784
#> trt1-trt3  0.03061 0.0122 36.8  2.512 1.65e-02  0.00592  0.05531
#> trt1-trt4 -0.01604 0.0122 36.8 -1.316 1.96e-01 -0.04073  0.00866
#> trt1-trt5  0.03884 0.0122 36.8  3.188 2.92e-03  0.01415  0.06354
#> trt2-trt3  0.03747 0.0122 36.8  3.075 3.96e-03  0.01278  0.06217
#> trt2-trt4 -0.00918 0.0122 36.8 -0.753 4.56e-01 -0.03387  0.01552
#> trt2-trt5  0.04570 0.0122 36.8  3.750 6.07e-04  0.02100  0.07040
#> trt3-trt4 -0.04665 0.0122 36.8 -3.828 4.85e-04 -0.07135 -0.02195
#> trt3-trt5  0.00823 0.0122 36.8  0.675 5.04e-01 -0.01647  0.03292
#> trt4-trt5  0.05488 0.0122 36.8  4.504 6.52e-05  0.03018  0.07957
#>
#> $RRRC$ciAvgRdrEachTrt
#>      Estimate StdErr    DF CILower CIUpper
#> trt1    0.753 0.0298  7.71   0.684   0.822
#> trt2    0.760 0.0284 10.69   0.697   0.823
#> trt3    0.723 0.0269  8.62   0.661   0.784
#> trt4    0.769 0.0357  5.24   0.679   0.860
#> trt5    0.714 0.0333  6.59   0.635   0.794
#>
#>
#> $FRRC
#> $FRRC$FTests
#>            DF    MS FStat       p
#> Treatment   4 0.463  3.85 0.00416
#> Error     796 0.120    NA      NA
#>
#> $FRRC$ciDiffTrt
#>           Estimate StdErr  DF      t   PrGTt  CILower CIUpper
#> trt1-trt2 -0.00686 0.0173 796 -0.395 0.69271 -0.04090  0.0272
#> trt1-trt3  0.03061 0.0173 796  1.765 0.07791 -0.00343  0.0647
#> trt1-trt4 -0.01604 0.0173 796 -0.925 0.35546 -0.05008  0.0180
#> trt1-trt5  0.03884 0.0173 796  2.240 0.02539  0.00480  0.0729
#> trt2-trt3  0.03747 0.0173 796  2.161 0.03103  0.00343  0.0715
#> trt2-trt4 -0.00918 0.0173 796 -0.529 0.59677 -0.04322  0.0249
#> trt2-trt5  0.04570 0.0173 796  2.635 0.00858  0.01166  0.0797
#> trt3-trt4 -0.04665 0.0173 796 -2.690 0.00730 -0.08069 -0.0126
#> trt3-trt5  0.00823 0.0173 796  0.474 0.63528 -0.02581  0.0423
#> trt4-trt5  0.05488 0.0173 796  3.164 0.00161  0.02084  0.0889
#>
#> $FRRC$ciAvgRdrEachTrt
#>      Estimate StdErr  DF CILower CIUpper
#> trt1    0.753 0.0217 199   0.711   0.796
#> trt2    0.760 0.0228 199   0.715   0.805
#> trt3    0.723 0.0216 199   0.680   0.765
```

```
#> trt4    0.769 0.0212 199   0.728   0.811
#> trt5    0.714 0.0228 199   0.669   0.759
#>
#> $FRRC$ciDiffTrtEachRdr
#>                 Estimate StdErr  DF       t   PrGTt   CILower   CIUpper
#> rdr1::trt1-trt2 -0.00773 0.0307 199 -0.2520 0.80131 -0.068250  0.052784
#> rdr1::trt1-trt3  0.04957 0.0307 199  1.6154 0.10781 -0.010942  0.110092
#> rdr1::trt1-trt4 -0.03087 0.0307 199 -1.0058 0.31573 -0.091384  0.029650
#> rdr1::trt1-trt5  0.03047 0.0307 199  0.9928 0.32203 -0.030050  0.090984
#> rdr1::trt2-trt3  0.05731 0.0307 199  1.8674 0.06332 -0.003209  0.117825
#> rdr1::trt2-trt4 -0.02313 0.0307 199 -0.7538 0.45186 -0.083650  0.037384
#> rdr1::trt2-trt5  0.03820 0.0307 199  1.2448 0.21469 -0.022317  0.098717
#> rdr1::trt3-trt4 -0.08044 0.0307 199 -2.6212 0.00944 -0.140959 -0.019925
#> rdr1::trt3-trt5 -0.01911 0.0307 199 -0.6226 0.53423 -0.079625  0.041409
#> rdr1::trt4-trt5  0.06133 0.0307 199  1.9986 0.04702  0.000816  0.121850
#> rdr3::trt1-trt2 -0.00201 0.0304 199 -0.0661 0.94733 -0.061885  0.057868
#> rdr3::trt1-trt3  0.00913 0.0304 199  0.3008 0.76389 -0.050743  0.069010
#> rdr3::trt1-trt4 -0.01822 0.0304 199 -0.6002 0.54904 -0.078102  0.041652
#> rdr3::trt1-trt5  0.04262 0.0304 199  1.4035 0.16202 -0.017260  0.102493
#> rdr3::trt2-trt3  0.01114 0.0304 199  0.3669 0.71406 -0.048735  0.071018
#> rdr3::trt2-trt4 -0.01622 0.0304 199 -0.5341 0.59389 -0.076093  0.043660
#> rdr3::trt2-trt5  0.04462 0.0304 199  1.4697 0.14323 -0.015252  0.104502
#> rdr3::trt3-trt4 -0.02736 0.0304 199 -0.9010 0.36867 -0.087235  0.032518
#> rdr3::trt3-trt5  0.03348 0.0304 199  1.1027 0.27148 -0.026393  0.093360
#> rdr3::trt4-trt5  0.06084 0.0304 199  2.0037 0.04645  0.000965  0.120718
#> rdr4::trt1-trt2 -0.01899 0.0368 199 -0.5166 0.60600 -0.091485  0.053502
#> rdr4::trt1-trt3  0.03132 0.0368 199  0.8519 0.39531 -0.041177  0.103810
#> rdr4::trt1-trt4  0.00927 0.0368 199  0.2521 0.80125 -0.063227  0.081760
#> rdr4::trt1-trt5  0.04845 0.0368 199  1.3179 0.18904 -0.024044  0.120944
#> rdr4::trt2-trt3  0.05031 0.0368 199  1.3685 0.17271 -0.022185  0.122802
#> rdr4::trt2-trt4  0.02826 0.0368 199  0.7687 0.44300 -0.044235  0.100752
#> rdr4::trt2-trt5  0.06744 0.0368 199  1.8345 0.06807 -0.005052  0.139935
#> rdr4::trt3-trt4 -0.02205 0.0368 199 -0.5998 0.54932 -0.094544  0.050444
#> rdr4::trt3-trt5  0.01713 0.0368 199  0.4661 0.64168 -0.055360  0.089627
#> rdr4::trt4-trt5  0.03918 0.0368 199  1.0659 0.28778 -0.033310  0.111677
#> rdr5::trt1-trt2  0.00131 0.0289 199  0.0453 0.96389 -0.055610  0.058227
#> rdr5::trt1-trt3  0.03243 0.0289 199  1.1237 0.26251 -0.024485  0.089352
#> rdr5::trt1-trt4 -0.02432 0.0289 199 -0.8425 0.40055 -0.081235  0.032602
#> rdr5::trt1-trt5  0.03384 0.0289 199  1.1724 0.24242 -0.023077  0.090760
#> rdr5::trt2-trt3  0.03112 0.0289 199  1.0783 0.28219 -0.025794  0.088044
#> rdr5::trt2-trt4 -0.02563 0.0289 199 -0.8878 0.37573 -0.082544  0.031294
#> rdr5::trt2-trt5  0.03253 0.0289 199  1.1271 0.26105 -0.024385  0.089452
#> rdr5::trt3-trt4 -0.05675 0.0289 199 -1.9661 0.05068 -0.113669  0.000169
#> rdr5::trt3-trt5  0.00141 0.0289 199  0.0488 0.96113 -0.055510  0.058327
#> rdr5::trt4-trt5  0.05816 0.0289 199  2.0149 0.04526  0.001240  0.115077
```

```
#>
#>
#> $RRFC
#> $RRFC$FTests
#>           DF     MS FStat        p
#> Treatment  4 0.4633  13.7 0.000202
#> Error     12 0.0339    NA       NA
#>
#> $RRFC$ciDiffTrt
#>           Estimate  StdErr DF      t    PrGTt CILower  CIUpper
#> trt1-trt2 -0.00686 0.00921 12 -0.745 4.71e-01 -0.0269  0.01321
#> trt1-trt3  0.03061 0.00921 12  3.324 6.06e-03  0.0106  0.05068
#> trt1-trt4 -0.01604 0.00921 12 -1.741 1.07e-01 -0.0361  0.00403
#> trt1-trt5  0.03884 0.00921 12  4.218 1.19e-03  0.0188  0.05891
#> trt2-trt3  0.03747 0.00921 12  4.069 1.56e-03  0.0174  0.05754
#> trt2-trt4 -0.00918 0.00921 12 -0.997 3.39e-01 -0.0292  0.01089
#> trt2-trt5  0.04570 0.00921 12  4.963 3.29e-04  0.0256  0.06576
#> trt3-trt4 -0.04665 0.00921 12 -5.066 2.77e-04 -0.0667 -0.02659
#> trt3-trt5  0.00823 0.00921 12  0.894 3.89e-01 -0.0118  0.02829
#> trt4-trt5  0.05488 0.00921 12  5.959 6.62e-05  0.0348  0.07494
#>
#> $RRFC$ciAvgRdrEachTrt
#>      Estimate StdErr DF CILower CIUpper
#> trt1    0.753 0.0235  3   0.678   0.828
#> trt2    0.760 0.0207  3   0.694   0.826
#> trt3    0.723 0.0207  3   0.657   0.788
#> trt4    0.769 0.0311  3   0.670   0.868
#> trt5    0.714 0.0273  3   0.627   0.801
```

14.7.1 Random-Reader Random-Case (RRRC) analysis

- `st4$RRRC$FTests` contains the results of the F-test of the NH.
- `st4$RRRC$ciDiffTrt` contains the confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- `st4$RRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRRC,\theta_{i\bullet}}$.

14.7.2 Fixed-Reader Random-Case (FRRC) analysis

- `st4$FRRC$FTests` contains results of the F-test of the NH, which is actually a chi-square statistic.

- `st4$FRRC$ciDiffTrt` contains confidence intervals for the inter-treatment difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- With I = 5 treatments there are 10 distinct treatment-pairings.
- Looking at the `PrGTt` (for probability greater than `t`) column, one finds six pairings that are significant: `trt1-trt3`, `trt1-trt5`, etc. The smallest p-value is for the `trt4-trt5` pairing. The findings are consistent with the prior ROC analysis, the difference being the smaller p-values.
- `st4$FRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,FRRC,\theta_{i\bullet}}$.
- `st4$FRRC$ciDiffTrtEachRdr` contains confidence intervals for inter-treatment difference FOMs, for each reader, i.e., $CI_{1-\alpha,FRRC,\theta_{ij}-\theta_{i'j}}$.

14.7.3 Random-Reader Fixed-Case (RRFC) analysis

- `st4$RRFC$FTests` contains the results of the F-test of the NH.
- `st4$RRFC$ciDiffTrt` contains confidence intervals for the inter-treatment paired difference FOMs, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}-\theta_{i'\bullet}}$.
- `st4$RRFC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e., $CI_{1-\alpha,RRFC,\theta_{i\bullet}}$.
- The `Estimate` column confirms that `trt5` has the smallest FOM while `trt4` has the highest (the `Estimates` column is identical for RRRC, FRRC and RRFC analyses).

14.8 Summary

14.9 Discussion

14.10 Tentative

```
ds1 <- dataset04 # do NOT convert to ROC
# comment/uncomment following code to disable/enable unequal weights
# K2 <- length(ds1$ratings$LL[1,1,,1])
# weights <- array(dim = c(K2, max(ds1$lesions$perCase)))
# perCase <- ds1$lesions$perCase
# for (k2 in 1:K2) {
#   sum <- 0
```

```
#   for (el in 1:perCase[k2]) {
#     weights[k2,el] <- 1/el
#     sum <- sum + 1/el
#   }
#   weights[k2,1:perCase[k2]] <- weights[k2,1:perCase[k2]] / sum
# }
# ds1$lesions$weights <- weights
ds <- ds1
FOM <- "wAFROC" # also try wAFROC1, MaxLLF and MaxNLF
st5 <- StSignificanceTesting(ds, FOM = FOM, method = "OR")
print(st5, digits = 4)
```

A comparison was run between results of OR and DBM for the FROC dataset. Except for `FRRC`, where differences are expected (because `ddf` in the former is ∞, while that in the later is $(I-1)\times(J-1)$), the results for the p-values were identical. This was true for the following FOMs: `wAFROC`, with equal and unequal weights, and `MaxLLF`. The confidence intervals (again, excluding `FRRC`) were identical for `FOM` = `wAFROC`. Slight differences were observed for `FOM` = `MaxLLF`.

14.11 References

Chapter 15

Sample size estimation for ROC studies DBM method

15.1 TBA How much finished

80%

15.2 Introduction

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

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

For simplicity and practicality, this chapter, and the next, is restricted to analysis of two-treatment data ($I = 2$). The purpose of most imaging system assessment studies is to determine, for a given diagnostic task, whether radiologists perform better using a new treatment over the conventional treatment, and

whether the difference is statistically significant. Therefore, the two-treatment case is the most common one encountered. While it is possible to extend the methods to more than two treatments, the extensions are not, in my opinion, clinically interesting.

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

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

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

What determines clinical significance? A small effect-size, e.g., 0.01 AUC units, could be clinically significant if it applies to a large population, where the small benefit in detection rate is amplified by the number of patients benefiting from

the new treatment. In contrast, for an "orphan" disease, i.e., one with very low prevalence, an effect-size of 0.05 might not be enough to justify the additional cost of the new treatment. The improvement might have to be 0.1 before it is worth it for a new treatment to be brought to market. One hates to monetize life and death issues, but there is no getting away from it, as cost/benefit issues determine clinical significance. The arbiters of clinical significance are engineers, imaging scientists, clinicians, epidemiologists, insurance companies and those who set government health care policies. The engineers and imaging scientists determine whether the effect-size the clinicians would like is feasible from technical and scientific viewpoints. The clinician determines, based on incidence of disease and other considerations, e.g., altruistic, malpractice, cost of the new device and insurance reimbursement, what effect-size is justifiable. Cohen has suggested that d values of 0.2, 0.5, and 0.8 be considered small, medium, and large, respectively, but he has also argued against their indiscriminate usage. However, after a study is completed, clinicians often find that an effect-size that biostatisticians label as small may, in certain circumstances, be clinically significant and an effect-size that they label as large may in other circumstances be clinically insignificant. Clearly, this is a complex issue. Some suggestions on choosing a clinically significant effect size are made in (TBA §11.12).

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

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

This chapter will focus on the DBM approach. Since it is based on a standard ANOVA model, it is easier to extend the NH testing procedure described in Chapter 09 to the alternative hypothesis, which is relevant for sample size estimation. [TBA Online Appendix 11.A shows how to translate the DBM formulae to the OR method (Hillis et al., 2011).]

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

15.3 Statistical Power

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

$$Power = 1 - \beta \tag{15.1}$$

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.

15.3.1 Observed vs. anticipated effect-size

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

15.3.2 Dependence of statistical power on estimates of model parameters

Examination of the expression for , Eqn. (11.5), shows that statistical power increases if:

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

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

15.3.4 Significance testing

15.3.5 p-value and confidence interval

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

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

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

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

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

$$\Delta = \frac{JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2} \tag{15.2}$$

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

$$F_{AH|R} \equiv \frac{MST}{MSTC} \sim F_{I-1,(I-1)(K-1),\Delta} \tag{15.3}$$

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

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

$$\Delta = \frac{JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2} \tag{15.4}$$

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

$$F_{AH|C} \equiv \frac{MST}{MSTR} \sim F_{I-1,(I-1)(J-1),\Delta} \tag{15.5}$$

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

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

15.4.3 Random-reader fixed-case (RRFC) analysis

15.4.4 Single-treatment multiple-reader analysis

15.5 Discussion/Summary/2

15.6 References

Chapter 16

Sample size estimation for ROC studies OR method

16.1 TBA How much finished

70%

16.2 Introduction

16.3 Statistical Power

$$Power = 1 - \beta \tag{16.1}$$

16.3.1 Sample size estimation for random-reader random-cases

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

$$Y_{n(ijk)} = \mu + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + (\tau RC)_{ijk} + \epsilon_{n(ijk)} \tag{16.2}$$

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:

$$d = \theta_{1\bullet} - \theta_{2\bullet} \tag{16.3}$$

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:

$$F_{DBM|AH} \sim F_{ndf,ddf,\Delta} \tag{16.4}$$

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

$$\Delta = \frac{JK\sigma^2_{Y;\tau}}{\left(\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC}\right) + K\sigma^2_{Y;\tau R} + J\sigma^2_{Y;\tau C}} \tag{16.5}$$

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:

$$\Delta = \frac{JK\sigma^2_{Y;\tau}}{\left(\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC}\right) + K\sigma^2_{Y;\tau R} + \max\left(J\sigma^2_{Y;\tau C}, 0\right)} \tag{16.6}$$

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:

$$ddf_H = \frac{\left[MSTR + \max(MSTR - MSTRC, 0)\right]^2}{\frac{[MSTR]^2}{(I-1)(J-1)}} \tag{16.7}$$

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:

$$\sigma^2_{Y;\tau} = \frac{d^2}{2} \tag{16.8}$$

16.3.2 Dependence of statistical power on estimates of model parameters

Examination of the expression for , Eqn. (11.5), shows that statistical power increases if:

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

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

16.3.4 Significance testing

16.3.5 p-value and confidence interval

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

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

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

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

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

$$\Delta = \frac{JK\sigma^2_{Y;\tau}}{\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC} + J\sigma^2_{Y;\tau C}} \tag{16.9}$$

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

$$F_{AH|R} \equiv \frac{MST}{MSTC} \sim F_{I-1,(I-1)(K-1),\Delta} \tag{16.10}$$

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

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

$$\Delta = \frac{JK\sigma^2_{Y;\tau}}{\sigma^2_{Y;\epsilon} + \sigma^2_{Y;\tau RC} + K\sigma^2_{Y;\tau R}} \tag{16.11}$$

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

$$F_{AH|C} \equiv \frac{MST}{MSTR} \sim F_{I-1,(I-1)(J-1),\Delta} \tag{16.12}$$

16.4.2 Example 1

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

Shown first is the "open" implementation.

```
alpha <- 0.05;cat("alpha = ", alpha, "\n")
#> alpha =  0.05
rocData <- dataset02 # select Van Dyke dataset
retDbm <- StSignificanceTesting(dataset = rocData, FOM = "Wilcoxon", method = "DBM")
```

```
varYTR <- retDbm$ANOVA$VarCom["VarTR","Estimates"]
varYTC <- retDbm$ANOVA$VarCom["VarTC","Estimates"]
varYEps <- retDbm$ANOVA$VarCom["VarErr","Estimates"]
effectSize <- retDbm$FOMs$trtMeanDiffs["trt0-trt1","Estimate"]
cat("effect size = ", effectSize, "\n")
#> effect size =  -0.043800322

#RRRC
J <- 10; K <- 163
ncp <- (0.5*J*K*(effectSize)^2)/(K*varYTR+max(J*varYTC,0)+varYEps)
MS <- UtilMeanSquares(rocData, FOM = "Wilcoxon", method = "DBM")
ddf <- (MS$msTR+max(MS$msTC-MS$msTRC,0))^2/(MS$msTR^2)*(J-1)
FCrit <- qf(1 - alpha, 1, ddf)
Power <- 1-pf(FCrit, 1, ddf, ncp = ncp)
data.frame("J"= J,  "K" = K, "FCrit" = FCrit, "ddf" = ddf, "ncp" = ncp, "RRRCPower" = P
#>    J   K     FCrit       ddf       ncp  RRRCPower
#> 1 10 163 4.1270572 34.334268 8.1269825 0.79111255

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

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

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

16.4.3 Fixed-reader random-case (FRRC) analysis

16.4.4 Random-reader fixed-case (RRFC) analysis

16.4.5 Single-treatment multiple-reader analysis

16.5 Discussion/Summary/3

16.6 References

Chapter 17

Analyzing FROC data

17.1 TBA How much finished

10%

17.2 Introduction

Analyzing FROC data is, apart from a single difference, very similar to analyzing ROC data. *The crucial difference is the selection of an appropriate location-sensitive figure of merit.* The reason is that the DBMH and ORH methods are applicable to any scalar figure of merit. Any appropriate FROC figure of merit reduces the mark rating data for a single dataset (i.e., a single treatment, a single reader and a number of cases) to a single scalar figure of merit.

The author recommends usage of the weighted AFROC figure of merit, where the lesions should be equally weighted, the default, unless there are strong clinical reasons for assigning unequal weights.

The chapter starts with analysis of a sample FROC dataset, #4 in Online Chapter 24. Any analysis should start with visualization of the relevant operating characteristic. Extensive examples are given using RJafroc implemented functions. Suggestions are made on how to report the results of a study (the suggestions apply equally to ROC studies). A method called *crossed-treatment analysis*, applicable when one has two treatment factors and their levels are crossed and one wishes to draw conclusions regarding the effect of treatments after averaging over all levels of the treatments.

17.3 Example 1

The following is a listing of file "mainAnalyzewAFROC.R". It performs both wAFROC and inferred ROC analyses of the same dataset and the results are saved to tables similar in structure to the Excel output tables shown for DBMH analysis of ROC data in §9.10.2. Empirical wAFROC-AUC and ROC-AUC for all combinations of treatments and readers, and reader-averaged AUCs for each treatment (Rdr. Avg.). The weighted AFROC results were obtained from worksheet FOMs in file FedwAfroc.xlsx. The highest rating AUC results were obtained from worksheet FOMs in file FedHrAuc.xlsx. The wAFROC-AUCs are smaller than the corresponding ROC-AUCs.

The datasets that come with this book are described in Online Chapter 24. Four of these are ROC datasets, one an LROC dataset and the rest (nine) are FROC datasets. For non-ROC datasets, the highest rating method was used to infer the corresponding ROC data. The datasets are identified in the code by strings contained in the string-array variable fileNames (line 7 - 8). Line 9 selects the dataset to be analyzed. In the example shown the "FED" dataset has been selected. It is a 5 treatment 4 radiologist FROC dataset1 acquired by Dr. Federica Zanca. Line 13 loads the dataset; this is done internal to the function loadDataFile(). Line 11 constructs the name of the wAFROC file and line 12 does the same for the ROC datafile. Line 15 which "spills over" to line 16 without the need for a special continuation character, generates an output file by performing DBMH significance testing (method = "DBMH") using fom = "wAFROC", i.e., the wAFROC figure of merit – this is the critical change. If one changes this to fom = "HrAuc", lines 19 – 20, then inferred ROC analysis occurs. In either case the default analysis, i.e., option = "ALL" is used, i.e., random-reader random-case (RRRC), fixed-reader random-case (FRRC) and random-reader fixed-case (RRFC). Results are shown below for random-reader random-case only.

The results of wAFROC analysis are saved to FedwAfroc.xlsx and that of inferred ROC analysis are saved to FedHrAuc.xlsx. The output file names need to be explicitly stated as otherwise they would overwrite each other (as a time-saver, checks are made at lines 14 and 18 to determine if the analysis has already been performed, in which case it is skipped).

In the Excel data file the readers are named 1, 3, 4 and 5 – the software treats the reader names as labels. The author's guess is that for some reason complete data for reader 2 could not be obtained. The renumber = TRUE option has the effect of renumbering the readers 1 through 4. Without renumbering, the output would be aesthetically displeasing, but have no effect on the conclusions.

Figures of merit, empirical wAFROC-AUC and empirical ROC-AUC, and the corresponding reader averages for both analyses are summarized in Table 19.1. The weighted AFROC results were obtained by copy and paste operations from worksheet FOMs in file FedwAfroc.xlsx. The highest rating AUC results were obtained by similar operations from worksheet FOMs in Excel file

FedHrAuc.xlsx. As expected, each wAFROC-AUC is smaller than the corresponding ROC-AUC.

Table 19.1: Empirical wAFROC-AUC and ROC-AUC for all combinations of treatments and readers, and reader-averaged AUCs for each treatment (Rdr. Avg.). The weighted AFROC results were obtained from worksheet FOMs in file FedwAfroc.xlsx. The highest rating AUC results were obtained from worksheet FOMs in file FedHrAuc.xlsx. The wAFROC-AUCs are smaller than the corresponding ROC-AUCs.

Table 19.2 shows results for RRRC analysis using the wAFROC-AUC FOM. The overall F-test of the null hypothesis that all treatments have the same reader-averaged FOM, rejected the NH: F(4, 36.8) = 7.8, p = 0.00012 . The numerator degree of freedom ndf is I - 1 = 4. Since the null hypothesis is that all treatments have the same FOM, this implies that at least one pairing of treatments yielded a significant FOM difference. The control for multiple testing is in the formulation of the null hypothesis and no further Bonferroni-like2 correction is needed. To determine which specific pairings are significantly different one examines the p-values (listed under Pr>t) in the "95% CI's FOMs, treatment difference" portion of the table. It shows that the following differences are significant at alpha = 0.05, namely "1 – 3", "1 – 5", "2 – 3", "2 – 5", "3 – 4" and "4 – 5"; these are indicated by asterisks. The values listed under the "95% CI's FOMs, each treatment" portion of the table show that treatment 4 yielded the highest FOM (0.769) followed closely by treatments 2 and 1, while treatment 5 had the least FOM (0.714), slightly worse than treatment 3. This explains why the p-value for the difference 4 - 5 is the smallest (0.00007) of all the listed p-values in the "95% CI's FOMs, each treatment" portion of the table. Each instance where the p-value for the individual treatment comparisons yields a significant p-value is accompanied by a 95% confidence interval that does not include zero. The two statements of significance, one in terms of a p-value and one in terms of a CI, are equivalent. When it comes to presenting results for treatment FOM differences, I prefer the 95% CI but some journals insist on a p-value, even when it is not significant. Note that two sequential tests are involved, an overall F-test of the NH that all treatments have the same performance and only if this yields a significant results is one justified in looking at the p-values of individual treatment pairings.

Table 19.2: wAFROC-AUC analysis: results of random-reader random-case (RRRC) analysis, in worksheet "RRRC"". [ddf = denominator degrees of freedom of F-distribution. df = degrees of freedom of t-distribution. Stderr = standard error. CI = confidence interval. * = Significantly different at alpha = 0.05.]

Table 19.3 shows corresponding results for the inferred ROC-AUC FOM. Again the null hypothesis was rejected: F(4, 16.8) = 3.46, p = 0.032. This means at least two treatments have significantly different FOMs. Looking down the table, one sees that the same 6 pairs (as compared to wAFROC analysis) are significantly different, 1 – 3, 1- 5, etc., as indicated by the asterisks. The

last five rows of the table show that treatment 4 had the highest performance while treatment 5 had the lowest performance. At the 5% significance level, both methods yielded the same significant differences, but this is not always true. While it is incorrect to conclude from a single dataset that a smaller p-value is indicative of higher statistical power, simulation testing under controlled conditions has consistently shown higher statistical power for the wAFROC-AUC FOM3,4 as compared to the inferred ROC-AUC FOM.

Table 19.3: Inferred ROC-AUC analysis: results of random-reader random-case (RRRC) analysis, in worksheet "RRRC"". ddf = denominator degrees of freedom of F-distribution. df = degrees of freedom of t-distribution. Stderr = standard error. CI = confidence interval; * = Significantly different at alpha = 0.05.].

17.4 Plotting wAFROC and ROC curves

It is important to display empirical wAFROC/ROC curves, not just for publication purposes, but to get a better feel for the data. Since treatments 4 and 5 showed the largest difference, the corresponding /ROC plots for them are displayed. The code is in file mainwAfrocRocPlots.R.

Sourcing this code yields Fig. 19.1. Plot (A), originating from lines 16 – 19, shows individual reader wAFROC plots for treatment 4 (solid lines) and treatment 5 (dashed lines). Running the software on one's computer best shows the color-coding. While difficult to see, examination of this plot shows that all readers performed better in treatment 4 than in treatment 5 (i.e., for each color the solid line is above the dashed line). Plot (B), originating from lines 21 – 25, shows reader-averaged wAFROC plots for treatments 4 (red line, upper curve) and 5 (blue line, lower curve). If one changes, for example, line 19 from print(plot1$wAFROCPlot$)$toprint(plot$1$wAFROCPoints) the code will output the coordinates of the points describing the curve, which gives the user the option to copy and paste the operating points into alternative plotting software.

Lines 16 – 19 create plots for all specified treatment-reader combinations. The "trick" to creating reader-averaged curves, such as in (B) is defining two list variables, plotT and plotR, at lines 21 – 22, the first containing the treatments to be plotted, list(4,5), and the second, a list of equal length, containing the arrays of readers to be averaged over, list(c(1:4), c(1:4)). More examples can be found in the help page for PlotEmpiricaOperatingCharacteristics().

Meaningful operating points on the reader average curves cannot be defined. This is because ratings are treatment and reader specific labels, so one cannot for example, average bin counts over all readers to construct a table like ROC Table 4.1 or its AFROC counterpart, Table 13.3.

Instead, the following procedure is used internal to PlotEmpiricaOperatingCharacteristics(). The reader-averaged plot for a specified treatment is obtained by

dividing the FPF range from 0 to 1 into finely spaced steps of 0.005. For each FPF value the wLLF values for that treatment are averaged over all readers, yielding the reader-averaged ordinate. Calculating confidence intervals on the reader-averaged curve is possible but cumbersome and unnecessary in my opinion. The relevant information, namely the 95% confidence interval on the difference in reader-averaged AUCs, is already contained in the program output, see Table 19.2, row labeled "4 – 5*". The difference is 0.05488 with a 95% confidence interval (0.03018, 0.07957).

Fig. 19.1: FED dataset; (A): individual reader wAFROC plots for treatments 4 and 5. While difficult to see, all readers performed better in treatment 4 as indicated by each colored solid line being above the corresponding dashed lines. (B): reader-averaged wAFROC plots for treatments 4 and 5. The performance superiority of treatment 4 is fairly obvious in this curve. The difference is significant, p = 0.00012.

Inferred ROC plots corresponding to Fig. 19.1 were generated by lines 20-24, i.e., by changing opChType = "wAFROC" to opChType = "ROC", and print(plot2$wAFROCPlot$)$toprint(plot$2$ROCPlot), resulting in Fig. 19.2. From Table 19.3 it is seen that the difference in reader-averaged AUCs is 0.04219 with a 95% confidence interval (0.00727, 0.07711). The observed wAFROC effect-size, 0.05488, is larger than the corresponding inferred ROC effect-size, 0.04219. This is a common observation, but sampling variability compounded with small differences, could give different results.

Fig. 19.2: FED dataset; (A): individual reader ROC plots for treatments 4 and 5. While difficult to see, all readers performed better in treatment 4. (B): reader-averaged ROC plots for treatments 4 and 5. The performance superiority of treatment 4 is fairly obvious in this curve. The difference is significant, p = 0.03054.

17.5 Reporting an FROC study

The methods section should make it clear exactly how the study was conducted. The information should be enough to allow some one else to replicate the study. How many readers, how many cases, how many treatments were used. How was ground truth determined and if the FROC paradigm was used, how were true lesion locations determined? The instructions to the readers should be clearly stated in writing. Precautions to minimize reading order effects should be stated – usually this is accomplished by interleaving cases from different treatments so that the chances that cases from a particular treatment is always seen first by every reader are minimized. Additionally, images from the same case, but in different treatments, should not be viewed in the same reading session. Reading sessions are usually an hour, and the different sessions should ideally be separated by at least one day. Users generally pay minimal attention to training sessions. It is recommended that at least 25% of the total number

of interpretations be training cases and cases used for training should not be used in the main study. Feedback should be provided during training session to allow the reader to become familiar with the range of difficulty levels regarding diseased and non-diseased cases in the dataset. Deception, e.g., stating a higher prevalence than is actually used, is usually not a good idea. The user-interface should be explained carefully. The best user interface is intuitive, minimizes keystrokes and requires the least explanation.

In publications, the paradigm used to collect the data (ROC, FROC, etc.) and the figure of merit used for analysis should be stated. If FROC, the proximity criterion should be stated. The analysis should state the NH and the alpha of the test, and the desired generalization. The software used and appropriate references should be cited. The results of the overall F-test, the p-value, the observed F-statistic and its degrees of freedom should be stated. If the NH is not rejected, one should cite the observed inter-treatment FOM differences, confidence intervals and p-values and ideally provide preliminary sample size estimates. This information could be useful to other researchers attempting to conduct a larger study. If the NH is rejected, a table of inter-treatment FOM differences such as Table 19.3 should be summarized. Reader averaged plots of the relevant operating characteristics for each treatment should be provided. In FROC studies it is recommended to vary the proximity criterion, perhaps increasing it by a factor of 2, to test if the final conclusions (is NH rejected and if so which treatment is highest) are unaffected.

Assuming the study has been done properly and with sufficiently large number of cases, the results should be published in some form, even if the NH is not rejected. The dearth of datasets to allow reasonable sample size calculations is a real problem in this field. The dataset set should be made available, perhaps on Research Gate, or if communicated to me, they will be included in the Online Appendix material. Datasets acquired via NIH or other government funding must be made available upon request, in an easily decipherable format. Subsequent users of these datasets must cite the original source of the data. Given the high cost of publishing excess pages in some journals, an expanded version, if appropriate for clarity, should be made available using online posting avenues.

17.6 Crossed-treatment analysis

This analysis was developed for a particular application6 in which nodule detection in an anthropomorphic chest phantom in computed tomography (CT) was evaluated as a function of tube charge and reconstruction method. The phantom was scanned at 4 values of mAs and images were reconstructed with adaptive iterative dose reduction 3D (AIDR3D) and filtered back projection (FBP). Thus there are two treatment factors and the factors are crossed since for each value of the mAs factor there were two values of the reconstruction

algorithm factor. Interest was in determining if whether performance depends on mAs and/or reconstruction method.

In a typical analysis of MRMC ROC or FROC study, treatment is considered as a single factor with I levels, where I is usually small. The figure of merit for treatment i (i =1,2,...,I) and reader j (j =1,2,...,J) is denoted ; the case set index is suppressed. MRMC analysis compares the observed magnitude of the difference in reader-averaged figures of merit between treatments i and i', , to the estimated standard deviation of the difference. For example, the reader-averaged difference in figures of merit is , where the dot symbol represents the average over the corresponding (reader) index. The standard deviation of the difference is estimated using the DBMH or the ORH method, using for example jackknifing to determine the variance components and/or covariances. With I levels, the number of distinct i vs. i' comparisons is I (I −1)/2. If the current study were analyzed in this manner, where I =8 (4 levels of mAs and two image reconstruction methods), then this would imply 28 comparisons. The large number of comparisons leads to loss of statistical power in detecting the effect of a specific pair of treatments, and, more importantly, does not inform one of the main points of interest: whether performance depends on mAs and/or reconstruction method. For example, in standard analysis the two reconstruction algorithms might be compared at different mAs levels, and one is in the dark as to which factor (algorithm or mAs) caused the observed significant difference.

Unlike conventional ROC type studies, the images in this study are defined by two factors. The first factor, tube charge, had four levels: 20, 40, 60 and 80 mAs. The second factor, reconstruction method, had two levels: FBP and AIDR3D. The figure of merit is represented by , where represents the levels of the first factor (mAs), and represents the levels of the second factor (reconstruction method), . Two sequential analyses were performed: (i) mAs analysis, where the figure of merit was averaged over (the reconstruction index); and (ii) reconstruction analysis, where the figure of merit was averaged over (the mAs index). For example, the mAs analysis figure of merit is , where the dot represents the average over the reconstruction index, and the corresponding reconstruction analysis figure of merit is , where the dot represents the average over the mAs index. Thus in either analysis, the figure of merit is dependent on a single treatment factor, and therefore standard DBMH or ORH methods apply.

The mAs analysis determines whether tube charge is a significant factor and in this analysis the number of possible comparisons is only six. The reconstruction analysis determines whether AIDR3D offers any advantage over FBP and in this analysis the number of possible comparisons is only one. Multiple testing on the same dataset increases the probability of Type I error, therefore a Bonferroni correction is applied by setting the threshold for declaring significance at 0.025; this is expected to conservatively maintain the overall probability of a Type I error at = 0.05. Crossed-treatment analysis is used to describe this type of

analysis of ROC/FROC data, which yields clearer answers on which of the two factors effects performance. The averaging over the other treatment has the effect of increasing the power of the study in detecting differences in each of the two factors.

Since the phantom is unique, and conclusions are only possible that are specific to this one phantom, the case (or image) factor was regarded as fixed. For this reason only results of random-reader fixed-case analyses are reported.

17.7 Discussion / Summary

An IDL (Interactive Data Language, currently marketed by Exelis Visual Information Solutions, www.exelisvis.com) version of JAFROC was first posted to a now obsolete website on 4/16/2004. This software required a license for IDL, which most users did not have. Subsequently, (9/27/2005) a version was posted which allowed analysis using the freely downloadable IDL Virtual Machine software (a method for freely distributing compiled IDL code). On 1/11/2011 the standalone Windows-compatible version was posted (4.0) and the current version is 4.2. JAFROC is windows compatible (XP, Vista and Windows 7, 8 and10).

To our knowledge JAFROC is the only easily accessible software currently available that can analyze FROC data. Workstation software for acquiring ROC and FROC data is available from several sources7-9. The Windows version is no longer actively supported (bugs, if pointed out, will be corrected). Current effort to conduct research and distribute software uses the R platform10. There are several advantages to this. R is an open-source platform - we have already benefited from a bug pointed out by a user . R runs on practically any platform (Windows, OSX, Linux, etc.). Also, developing an R package benefits from other contributed R-packages, which allow easy computation of probability integrals, random number generation, and parallel computing to speed up computations, to name just a few. The drawback with R, and this has to with its open source philosophy, is that one cannot readily integrate existing ROC code, developed on other platforms and other programming languages (specifically, DLLs are not allowed in R). So useful programs like CORROC2 and CBM were coded in C++, since R allows C++ programs to be compiled and included in a package.

Due to the random number of marks per image, data entry in the FROC paradigm is inherently more complicated and error-prone than in ROC analysis, and consequently, and in response to feedback from users, much effort has gone into error checking. The users have especially liked the feature where the program indicates the Excel sheet name and line-number where an error is detected. User-feedback has also been very important in detecting program bugs and inconsistencies in the documentation and developing additional features (e.g., ROI analysis).

Interest in the FROC paradigm is evidenced by the fact that Ref. 3 describing the JAFROC method has been cited over 273 times. Over 25,000 unique visitors have viewed my website, at least 73 have downloaded the software and over 107 publications using JAFROC have appeared. The list is available on my website. JAFROC has been applied to magnetic resonance imaging, virtual computerized tomography colonoscopy, digital tomosynthesis (chest and breast), mammography dose and image processing optimization, computer aided detection (CAD), computerized tomography, and other applications.

Since confusion still appears to exist, especially among statisticians, regarding perceived neglect of intra-image correlations of ratings and how true negatives are handled in FROC analysis11, we close with a quote from respected sources 12 "(Chakraborty and Berbaum) have presented a solution to the FROC problem using a jackknife resampling approach that respects the correlation structure in the images … their paradigm successfully passes a rigorous statistical validation test". Since 2005 the National Institutes for Health (NIH) has been generous with supporting the research and users of JAFROC have been equally generous with providing their datasets, which have resulted in several collaborations.

17.8 References

Chapter 18

FROC sample size

18.1 TBA How much finished

10% TBA Merge the vignette into this …

18.2 Introduction

FROC sample size estimation is not fundamentally different from the procedure outlined in Chapter 11 for the ROC paradigm. To recapitulate, based on analysis of a pilot ROC dataset and using a specified FOM, e.g., the ROC-AUC, and either the DBMH or the ORH method for significance testing, one estimates the intrinsic variability of the data expressed in terms of variance components. For DBMH analysis, these are the pseudovalue variance components, while for ORH analysis these are the FOM treatment-reader variance component and the FOM covariances. The second step is to specify a clinically realistic effect-size, e.g., the AUC difference between the two modalities. Given these values, the sample size functions implemented in `RJafroc` allow one to estimate the number of readers and cases necessary to detect (i.e., reject the null hypothesis) the modality AUC difference at specified Type II error rate β, typically chosen to be 20% - corresponding to 80% statistical power - and specified Type I error rate α, typically chosen to be 5%.

In FROC analysis the only difference, indeed the critical difference, is the choice of FOM; e.g., the wAFROC-AUC instead of the inferred ROC-AUC. The FROC dataset is analyzed using either the DBMH or the ORH method. This yields the necessary variance components or the covariance matrix corresponding to the wAFROC-AUC. The next step is to specify an effect-size in wAFROC-AUC units, and therein lies the problem. What value does one use? The ROC-AUC has a historically well-known interpretation: the classification ability at

separating diseased patients from non-diseased patients. Needed is a way of relating the effect-size in ROC-AUC units to one in wAFROC-AUC units.

1. Choose an ROC-AUC effect-size that is realistic, one that clinicians understand and can therefore participate in, in the effect-size postulation process.
2. Convert the ROC effect-size to a wAFROC-AUC effect-size: the method for this is described in the next section.
3. Use the sample size tools in in `RJafroc`, i.e., functions with names beginning with `Ss`, to determine the necessary sample size.

It is important to recognize is that all quantities have to be in the same units. When performing ROC analysis, everything (variance components and effect-size) has to be in units of the selected FOM, e.g., Wilcoxon statistic. When performing wAFROC analysis, everything has to be in units of the wAFROC-AUC. The variance components and effect-size in wAFROC-AUC units will be different from their corresponding ROC counterparts. In particular, as shown next, an ROC-AUC effect-size of 0.05 generally correspond to a larger effect-size in wAFROC-AUC units. The reason for this is that the range over which wAFROC-AUC can vary, namely 0 to 1, is twice the corresponding ROC-AUC range.

The next section explains the steps used to implement #2 above.

For each modality-reader (ij) dataset, the inferred ROC data is fitted by the procedure described above, yielding estimates of the parameters (notice the usage of intrinsic RSM parameters, not the primed values; the latter are easily converted to intrinsic values). The pilot study represents an "almost" null hypothesis dataset: if a significance difference was observed one would not be going through the exercise of samples size estimation. In any case, I recommend taking the median of the three sets of parameters, over all indices, as representing the average NH dataset. The median is less sensitive to outliers than the average.

. (19.1)

Using these values ROC-AUC and wAFROC-AUC, for the NH condition, denoted and respectively, are calculated by numerical integration of the RSM predicted ROC and wAFROC curves, Chapter 17:

. (19.2)

To induce the alternative hypothesis condition one increments by . The resulting ROC-AUC and wAFROC-AUC are calculated, again by numerical integration of the RSM predicted ROC and wAFROC curves, leading to the corresponding effect-sizes (note that in each equation below one takes the difference between the AH value minus the NH value):

. (19.3)

Eqn. (19.3), evaluated for different values of , provides a calibration curve between the effect-sizes expressed in the two units, Fig. 19.4 (A). This allows one to interpolate the appropriate wAFROC effect-size corresponding to any postulated ROC effect-size.

18.3 Example 1

Empirical wAFROC-AUC and ROC-AUC for all combinations of treatments and readers, and reader-averaged AUCs for each treatment (Rdr. Avg.). The weighted AFROC results were obtained from worksheet FOMs in file FedwAfroc.xlsx. The highest rating AUC results were obtained from worksheet FOMs in file FedHrAuc.xlsx. The wAFROC-AUCs are smaller than the corresponding ROC-AUCs.

Table 19.2 shows results for RRRC analysis using the wAFROC-AUC FOM. The overall F-test of the null hypothesis that all treatments have the same reader-averaged FOM, rejected the NH: F(4, 36.8) = 7.8, p = 0.00012 . The numerator degree of freedom ndf is I - 1 = 4. Since the null hypothesis is that all treatments have the same FOM, this implies that at least one pairing of treatments yielded a significant FOM difference. The control for multiple testing is in the formulation of the null hypothesis and no further Bonferroni-like2 correction is needed. To determine which specific pairings are significantly different one examines the p-values (listed under Pr>t) in the "95% CI's FOMs, treatment difference" portion of the table. It shows that the following differences are significant at alpha = 0.05, namely "1 – 3", "1 – 5", "2 – 3", "2 – 5", "3 – 4" and "4 – 5"; these are indicated by asterisks. The values listed under the "95% CI's FOMs, each treatment" portion of the table show that treatment 4 yielded the highest FOM (0.769) followed closely by treatments 2 and 1, while treatment 5 had the least FOM (0.714), slightly worse than treatment 3. This explains why the p-value for the difference 4 - 5 is the smallest (0.00007) of all the listed p-values in the "95% CI's FOMs, each treatment" portion of the table. Each instance where the p-value for the individual treatment comparisons yields a significant p-value is accompanied by a 95% confidence interval that does not include zero. The two statements of significance, one in terms of a p-value and one in terms of a CI, are equivalent. When it comes to presenting results for treatment FOM differences, I prefer the 95% CI but some journals insist on a p-value, even when it is not significant. Note that two sequential tests are involved, an overall F-test of the NH that all treatments have the same performance and only if this yields a significant results is one justified in looking at the p-values of individual treatment pairings.

18.4 Plotting wAFROC and ROC curves

It is important to display empirical wAFROC/ROC curves, not just for publication purposes, but to get a better feel for the data. Since treatments 4 and 5 showed the largest difference, the corresponding /ROC plots for them are displayed. The code is in file mainwAfrocRocPlots.R.

The methods section should make it clear exactly how the study was conducted. The information should be enough to allow some one else to replicate the study. How many readers, how many cases, how many treatments were used. How was ground truth determined and if the FROC paradigm was used, how were true lesion locations determined? The instructions to the readers should be clearly stated in writing. Precautions to minimize reading order effects should be stated – usually this is accomplished by interleaving cases from different treatments so that the chances that cases from a particular treatment is always seen first by every reader are minimized. Additionally, images from the same case, but in different treatments, should not be viewed in the same reading session. Reading sessions are usually an hour, and the different sessions should ideally be separated by at least one day. Users generally pay minimal attention to training sessions. It is recommended that at least 25% of the total number of interpretations be training cases and cases used for training should not be used in the main study. Feedback should be provided during training session to allow the reader to become familiar with the range of difficulty levels regarding diseased and non-diseased cases in the dataset. Deception, e.g., stating a higher prevalence than is actually used, is usually not a good idea. The user-interface should be explained carefully. The best user interface is intuitive, minimizes keystrokes and requires the least explanation.

In publications, the paradigm used to collect the data (ROC, FROC, etc.) and the figure of merit used for analysis should be stated. If FROC, the proximity criterion should be stated. The analysis should state the NH and the alpha of the test, and the desired generalization. The software used and appropriate references should be cited. The results of the overall F-test, the p-value, the observed F-statistic and its degrees of freedom should be stated. If the NH is not rejected, one should cite the observed inter-treatment FOM differences, confidence intervals and p-values and ideally provide preliminary sample size estimates. This information could be useful to other researchers attempting to conduct a larger study. If the NH is rejected, a table of inter-treatment FOM differences such as Table 19.3 should be summarized. Reader averaged plots of the relevant operating characteristics for each treatment should be provided. In FROC studies it is recommended to vary the proximity criterion, perhaps increasing it by a factor of 2, to test if the final conclusions (is NH rejected and if so which treatment is highest) are unaffected.

Assuming the study has been done properly and with sufficiently large number of cases, the results should be published in some form, even if the NH is not rejected. The dearth of datasets to allow reasonable sample size calculations is

a real problem in this field. The dataset set should be made available, perhaps on Research Gate, or if communicated to me, they will be included in the Online Appendix material. Datasets acquired via NIH or other government funding must be made available upon request, in an easily decipherable format. Subsequent users of these datasets must cite the original source of the data. Given the high cost of publishing excess pages in some journals, an expanded version, if appropriate for clarity, should be made available using online posting avenues.

Crossed-treatment analysis This analysis was developed for a particular application6 in which nodule detection in an anthropomorphic chest phantom in computed tomography (CT) was evaluated as a function of tube charge and reconstruction method. The phantom was scanned at 4 values of mAs and images were reconstructed with adaptive iterative dose reduction 3D (AIDR3D) and filtered back projection (FBP). Thus there are two treatment factors and the factors are crossed since for each value of the mAs factor there were two values of the reconstruction algorithm factor. Interest was in determining if whether performance depends on mAs and/or reconstruction method.

In a typical analysis of MRMC ROC or FROC study, treatment is considered as a single factor with I levels, where I is usually small. The figure of merit for treatment i (i =1,2,...,I) and reader j (j =1,2,...,J) is denoted ; the case set index is suppressed. MRMC analysis compares the observed magnitude of the difference in reader-averaged figures of merit between treatments i and i', , to the estimated standard deviation of the difference. For example, the reader-averaged difference in figures of merit is , where the dot symbol represents the average over the corresponding (reader) index. The standard deviation of the difference is estimated using the DBMH or the ORH method, using for example jackknifing to determine the variance components and/or covariances. With I levels, the number of distinct i vs. i' comparisons is I (I −1)/2. If the current study were analyzed in this manner, where I =8 (4 levels of mAs and two image reconstruction methods), then this would imply 28 comparisons. The large number of comparisons leads to loss of statistical power in detecting the effect of a specific pair of treatments, and, more importantly, does not inform one of the main points of interest: whether performance depends on mAs and/or reconstruction method. For example, in standard analysis the two reconstruction algorithms might be compared at different mAs levels, and one is in the dark as to which factor (algorithm or mAs) caused the observed significant difference.

Unlike conventional ROC type studies, the images in this study are defined by two factors. The first factor, tube charge, had four levels: 20, 40, 60 and 80 mAs. The second factor, reconstruction method, had two levels: FBP and AIDR3D. The figure of merit is represented by , where represents the levels of the first factor (mAs), and represents the levels of the second factor (reconstruction method), . Two sequential analyses were performed: (i) mAs analysis, where the figure of merit was averaged over (the reconstruction index); and

(ii) reconstruction analysis, where the figure of merit was averaged over (the mAs index). For example, the mAs analysis figure of merit is , where the dot represents the average over the reconstruction index, and the corresponding reconstruction analysis figure of merit is , where the dot represents the average over the mAs index. Thus in either analysis, the figure of merit is dependent on a single treatment factor, and therefore standard DBMH or ORH methods apply.

The mAs analysis determines whether tube charge is a significant factor and in this analysis the number of possible comparisons is only six. The reconstruction analysis determines whether AIDR3D offers any advantage over FBP and in this analysis the number of possible comparisons is only one. Multiple testing on the same dataset increases the probability of Type I error, therefore a Bonferroni correction is applied by setting the threshold for declaring significance at 0.025; this is expected to conservatively maintain the overall probability of a Type I error at $= 0.05$. Crossed-treatment analysis is used to describe this type of analysis of ROC/FROC data, which yields clearer answers on which of the two factors effects performance. The averaging over the other treatment has the effect of increasing the power of the study in detecting differences in each of the two factors.

Since the phantom is unique, and conclusions are only possible that are specific to this one phantom, the case (or image) factor was regarded as fixed. For this reason only results of random-reader fixed-case analyses are reported.

18.5 `FitRsmROC` usage example

18.6 Discussion / Summary

Over the years, there have been several attempts at fitting FROC data. Prior to the RSM-based ROC curve approach described in this chapter, all methods were aimed at fitting FROC curves, in the mistaken belief that this approach was using all the data. The earliest was my FROCFIT software 36. This was followed by Swensson's approach 37, subsequently shown to be equivalent to my earlier work, as far as predicting the FROC curve was concerned 11. In the meantime, CAD developers, who relied heavily on the FROC curve to evaluate their algorithms, developed an empirical approach that was subsequently put on a formal basis in the IDCA method 12.

This chapter describes an approach to fitting ROC curves, instead of FROC curves, using the RSM. On the face of it, fitting the ROC curve seems to be ignoring much of the data. As an example, the ROC rating on a non-diseased case is the rating of the highest-rated mark on that image, or negative infinity if the case has no marks. If the case has several NL marks, only the highest rated one is used. In fact the highest rated mark contains information about the

other marks on the case, namely they were all rated lower. There is a statistical term for this, namely sufficiency 38. As an example, the highest of a number of samples from a uniform distribution is a sufficient statistic, i.e., it contains all the information contained in the observed samples. While not quite the same for normally distributed values, neglect of the NLs rated lower is not as bad as might seem at first.

18.7 References

Bibliography

Bunch, P. C., Hamilton, J. F., Sanderson, G. K., and Simmons, A. H. (1977). A free response approach to the measurement and characterization of radiographic observer performance. In *Application of Optical Instrumentation in Medicine VI*, volume 127, pages 124–135. International Society for Optics and Photonics.

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

Chakraborty, D. P. (2002). Statistical power in observer performance studies: A comparison of the ROC and free-response methods in tasks involving localization. *Acad. Radiol.*, 9(2):147–156.

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

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

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, 2 edition.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 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. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.

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. A. (1998). Sample size calculations in studies of test accuracy. *Statistical Methods in Medical Research*, 7(4):371–392.

Obuchowski, N. A. (2000). Sample size tables for receiver operating characteristic studies. *Am. J. Roentgenol.*, 175(3):603–608.

Obuchowski, N. A. and Rockette, H. (1995). Hypothesis testing of the diagnostic accuracy for multiple diagnostic tests: An ANOVA approach with dependent observations. *Communications in Statistics: Simulation and Computation*, 24:285–308.

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

Roe, C. A. and Metz, C. (1997b). Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: Validation with computer simulation. *Acad Radiol*, 4:298–303.

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

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

Starr, S., Metz, C., Lusted, L., Sharp, P., and Herath, K. (1977). Comments on the generalization of receiver operating characteristic analysis to detection and localization tasks. *Physics in Medicine & Biology*, 22(2):376.

Starr, S. J., Metz, C. E., Lusted, L. B., and Goodenough, D. J. (1975). Visual detection and localization of radiographic images. *Radiology*, 116(3):533–538.

Swensson, R. G. (1996). Unified measurement of observer performance in detecting and localizing target objects on images. *Medical physics*, 23(10):1709–1725.

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

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

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

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

Zanca, F., Jacobs, J., Van Ongeval, C., Claus, F., Celis, V., Geniets, C., Provost, V., Pauwels, H., Marchal, G., and Bosmans, H. (2009). Evaluation of clinical image processing algorithms used in digital mammography. *Medical Physics*, 36(3):765–775.

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