

# The RJafrroc Book

Dev P. Chakraborty, PhD

2021-12-18



# Contents

<b>Preface</b>	<b>19</b>
TBA How much finished . . . . .	19
The pdf file of the book . . . . .	19
A note on the online distribution mechanism of the book . . . . .	19
Structure of the book . . . . .	20
Contributing to this book . . . . .	20
Is this book relevant to you and what are the alternatives? . . . . .	20
ToDos TBA . . . . .	21
Chapters needing heavy edits . . . . .	21
Shelved vs. removed vs. parked folders needing heavy edits . . . . .	21
Coding aids . . . . .	21
<b>Quick Start</b>	<b>25</b>
<b>1 Help</b>	<b>25</b>
1.1 TBA How much finished . . . . .	25
1.2 Getting help on the software . . . . .	25
1.3 References . . . . .	25
<b>2 JAFROC data format</b>	<b>27</b>
2.1 TBA How much finished . . . . .	27
2.2 Introduction . . . . .	27
2.3 Note to existing users . . . . .	28

2.4	Contents of Excel file . . . . .	28
2.5	The Truth worksheet . . . . .	29
2.6	The false positive (FP) ratings . . . . .	30
2.7	The true positive (TP) ratings . . . . .	32
2.8	A single reader dataset . . . . .	33
2.9	References . . . . .	33
<b>3</b>	<b>Reading the Excel data file</b>	<b>35</b>
3.1	TBA How much finished . . . . .	35
3.2	Introduction . . . . .	35
3.3	The structure of an ROC dataset . . . . .	36
3.4	Correspondence between NL member of dataset and the FP worksheet . . . . .	39
3.5	Case-index vs. caseID . . . . .	40
3.6	Correspondence between LL member of dataset and the TP worksheet . . . . .	41
3.7	References . . . . .	42
<b>4</b>	<b>Data format and reading FROC data</b>	<b>43</b>
4.1	TBA How much finished . . . . .	43
4.2	Introduction . . . . .	43
4.3	The Truth worksheet . . . . .	44
4.4	Reading the FROC dataset . . . . .	46
4.5	The false positive (FP) ratings . . . . .	47
4.6	The true positive (TP) ratings . . . . .	49
4.7	On the distribution of numbers of lesions in diseased cases . . . . .	50
4.8	Definition of <code>lesWghtDistr</code> array . . . . .	53
4.9	References . . . . .	55
<b>5</b>	<b>DBM analysis text output</b>	<b>57</b>
5.1	TBA How much finished . . . . .	57
5.2	Introduction . . . . .	57
5.3	Analyzing the ROC dataset . . . . .	57

<b>CONTENTS</b>	<b>5</b>
5.4 Explanation of the output . . . . .	57
5.5 References . . . . .	64
<b>6 OR analysis text output</b>	<b>65</b>
6.1 TBA How much finished . . . . .	65
6.2 Introduction . . . . .	65
6.3 Analyzing the ROC dataset . . . . .	65
6.4 Explanation of the output . . . . .	65
6.5 References . . . . .	69
<b>7 OR analysis Excel output</b>	<b>71</b>
7.1 TBA How much finished . . . . .	71
7.2 Introduction . . . . .	71
7.3 Generating the Excel output file . . . . .	71
7.4 References . . . . .	72
<b>ROC paradigm</b>	<b>77</b>
<b>8 Preliminaries</b>	<b>77</b>
8.1 TBA How much finished . . . . .	77
8.2 Introduction . . . . .	77
8.3 Clinical tasks . . . . .	78
8.4 Imaging device development and its clinical deployment . . . . .	81
8.5 Image quality vs. task performance . . . . .	86
8.6 Why physical measures of image quality are not enough . . . . .	87
8.7 Model observers . . . . .	88
8.8 Measuring observer performance: four paradigms . . . . .	89
8.9 Hierarchy of assessment methods . . . . .	92
8.10 Overview of the book and how to use it . . . . .	94
8.11 Summary . . . . .	96
8.12 Discussion . . . . .	96
8.13 References . . . . .	96

<b>9 The Binary Task</b>	<b>97</b>
9.1 TBA How much finished . . . . .	97
9.2 Introduction . . . . .	97
9.3 The fundamental 2x2 table . . . . .	98
9.4 Sensitivity and specificity . . . . .	99
9.5 Disease prevalence . . . . .	102
9.6 Accuracy . . . . .	103
9.7 Negative and positive predictive values . . . . .	104
9.8 Summary . . . . .	108
9.9 Discussion . . . . .	108
9.10 References . . . . .	108
<b>10 Modeling the Binary Task</b>	<b>109</b>
10.1 TBA How much finished . . . . .	109
10.2 Introduction . . . . .	109
10.3 Decision variable and decision threshold . . . . .	110
10.4 Changing the decision threshold: Example I . . . . .	113
10.5 Changing the decision threshold: Example II . . . . .	114
10.6 The equal-variance binormal model . . . . .	114
10.7 The normal distribution . . . . .	116
10.8 Analytic expressions for specificity and sensitivity . . . . .	121
10.9 Demonstration of the concepts of sensitivity and specificity . . .	125
10.10 Inverse variation of sensitivity and specificity and the need for a single FOM . . . . .	129
10.11 The ROC curve . . . . .	129
10.12 Assigning confidence intervals to an operating point . . . . .	137
10.13 Variability in sensitivity and specificity: the Beam et al study .	141
10.14 Summary . . . . .	143
10.15 References . . . . .	144

<b>11 Ratings Paradigm</b>	<b>145</b>
11.1 TBA How much finished . . . . .	145
11.2 Introduction . . . . .	145
11.3 The ROC counts table . . . . .	146
11.4 Operating points from counts table . . . . .	147
11.5 Automating all this . . . . .	151
11.6 Relation between ratings paradigm and the binary paradigm . .	154
11.7 Ratings are not numerical values . . . . .	155
11.8 A single “clinical” operating point from ratings data . . . . .	156
11.9 The forced choice paradigm . . . . .	157
11.10 Observer performance studies as laboratory simulations of clinical tasks . . . . .	159
11.11 Discrete vs. continuous ratings: the Miller study . . . . .	160
11.12 The BI-RADS ratings scale and ROC studies . . . . .	164
11.13 The controversy . . . . .	165
11.14 Discussion . . . . .	168
11.15 References . . . . .	168
<b>12 Empirical AUC</b>	<b>169</b>
12.1 TBA How much finished . . . . .	169
12.2 Introduction . . . . .	169
12.3 The empirical ROC plot . . . . .	170
12.4 Empirical operating points from ratings data . . . . .	172
12.5 AUC under the empirical ROC plot . . . . .	174
12.6 The Wilcoxon statistic . . . . .	177
12.7 Bamber’s Equivalence theorem . . . . .	177
12.8 Importance of Bamber’s theorem . . . . .	181
12.9 Discussion / Summary . . . . .	182
12.10 Appendix 5.A: Details of Wilcoxon theorem . . . . .	182
12.11 References . . . . .	183

<b>13 Binormal model</b>	<b>185</b>
13.1 TBA How much finished . . . . .	185
13.2 TBA Introduction . . . . .	185
13.3 Binormal model . . . . .	186
13.4 Binormal ROC curve . . . . .	190
13.5 Scalar threshold-independent measure . . . . .	191
13.6 Partial AUC vs. true performance . . . . .	193
13.7 Illustrative plots . . . . .	195
13.8 Geometrical argument . . . . .	197
13.9 Optimal operating point on ROC . . . . .	198
13.10 Discussion . . . . .	200
13.11 Appendix I: Density functions . . . . .	203
13.12 Appendix II: Area under binormal ROC . . . . .	203
13.13 Appendix III: Invariance property of pdfs . . . . .	207
13.14 Appendix IV: Fitting an ROC curve . . . . .	212
13.15 Appendix V: Validating fitting model . . . . .	219
13.16 References . . . . .	221
<b>14 Sources of AUC variability</b>	<b>223</b>
14.1 TBA How much finished . . . . .	223
14.2 Introduction . . . . .	223
14.3 Three sources of variability . . . . .	224
14.4 Dependence of AUC on the case sample . . . . .	226
14.5 DeLong method . . . . .	228
14.6 Bootstrap method . . . . .	232
14.7 Jackknife method . . . . .	237
14.8 Calibrated simulator . . . . .	241
14.9 Discussion . . . . .	245
14.10 References . . . . .	246

<b>CONTENTS</b>	<b>9</b>
<b>Significance Testing</b>	<b>249</b>
<b>15 Hypothesis Testing</b>	<b>249</b>
15.1 TBA How much finished . . . . .	249
15.2 Introduction . . . . .	249
15.3 Single-modality single-reader ROC study . . . . .	250
15.4 Type-I errors . . . . .	253
15.5 One vs. two sided tests . . . . .	255
15.6 Statistical power . . . . .	258
15.7 Comments . . . . .	263
15.8 Why alpha is chosen as 5% . . . . .	264
15.9 Discussion . . . . .	265
15.10 References . . . . .	266
<b>16 DBM method background</b>	<b>267</b>
16.1 TBA How much finished . . . . .	267
16.2 Introduction . . . . .	267
16.3 Random and fixed factors . . . . .	271
16.4 Reader and case populations . . . . .	272
16.5 Three types of analyses . . . . .	273
16.6 General approach . . . . .	273
16.7 Summary TBA . . . . .	275
16.8 References . . . . .	276
<b>17 Significance Testing using the DBM Method</b>	<b>277</b>
17.1 TBA How much finished . . . . .	277
17.2 The DBM sampling model . . . . .	277
17.3 Expected values of mean squares . . . . .	283
17.4 Random-reader random-case (RRRC) analysis . . . . .	284
17.5 Sample size estimation for random-reader random-case generalization . . . . .	293
17.6 Significance testing and sample size estimation for fixed-reader random-case generalization . . . . .	296

17.7 Significance testing and sample size estimation for random-reader fixed-case generalization . . . . .	297
17.8 Summary TBA . . . . .	297
17.9 Things for me to think about . . . . .	299
17.10 References . . . . .	300
<b>18 DBM method special cases</b>	<b>301</b>
18.1 TBA How much finished . . . . .	301
18.2 Fixed-reader random-case (FRRC) analysis . . . . .	301
18.3 Random-reader fixed-case (RRFC) analysis . . . . .	304
18.4 References . . . . .	305
<b>19 Introduction to the Obuchowski-Rockette method</b>	<b>307</b>
19.1 TBA How much finished . . . . .	307
19.2 Locations of helper functions . . . . .	307
19.3 Introduction . . . . .	307
19.4 Single-reader multiple-treatment . . . . .	308
19.5 Single-treatment multiple-reader . . . . .	314
19.6 Multiple-reader multiple-treatment . . . . .	315
19.7 Summary . . . . .	321
19.8 Discussion . . . . .	321
19.9 Appendix: Covariance and correlation . . . . .	321
19.10 References . . . . .	332
<b>20 Obuchowski Rockette (OR) Analysis</b>	<b>333</b>
20.1 TBA How much finished . . . . .	333
20.2 Introduction . . . . .	333
20.3 Random-reader random-case . . . . .	334
20.4 Fixed-reader random-case . . . . .	338
20.5 Random-reader fixed-case . . . . .	339
20.6 Single treatment analysis . . . . .	340

CONTENTS	11
----------	----

<b>21 Obuchowski Rockette Applications</b>	<b>341</b>
21.1 TBA How much finished . . . . .	341
21.2 Introduction . . . . .	341
21.3 Hand calculation . . . . .	342
21.4 RJafroc: dataset02 . . . . .	351
21.5 RJafroc: dataset04 . . . . .	357
21.6 RJafroc: dataset04, FROC . . . . .	363
21.7 RJafroc: dataset04, FROC/DBM . . . . .	370
21.8 Summary . . . . .	375
21.9 Discussion . . . . .	375
21.10 Tentative . . . . .	375
21.11 References . . . . .	376
<b>22 Sample size estimation for ROC studies DBM method</b>	<b>377</b>
22.1 TBA How much finished . . . . .	377
22.2 Introduction . . . . .	377
22.3 Statistical Power . . . . .	380
22.4 Formulae for fixed-reader random-case (FRRC) sample size estimation . . . . .	383
22.5 Discussion/Summary/2 . . . . .	384
22.6 References . . . . .	384
<b>23 Sample size estimation for ROC studies OR method</b>	<b>385</b>
23.1 TBA How much finished . . . . .	385
23.2 Introduction . . . . .	385
23.3 Statistical Power . . . . .	385
23.4 Formulae for fixed-reader random-case (FRRC) sample size estimation . . . . .	389
23.5 Discussion/Summary/3 . . . . .	391
23.6 References . . . . .	391

<b>FROC paradigm</b>	<b>395</b>
<b>24 The FROC paradigm</b>	<b>395</b>
24.1 TBA How much finished . . . . .	395
24.2 Introduction . . . . .	395
24.3 Location specific paradigms . . . . .	396
24.4 Visual search . . . . .	400
24.5 A pioneering FROC study in medical imaging . . . . .	403
24.6 The free-response receiver operating characteristic (FROC) plot .	405
24.7 Preview of the RSM data simulator . . . . .	406
24.8 Population and binned FROC plots . . . . .	407
24.9 Perceptual SNR . . . . .	414
24.10 The “solar” analogy: search vs. classification performance . . . . .	414
24.11 Discussion and suggestions . . . . .	418
24.12 References . . . . .	418
<b>25 Empirical plots</b>	<b>419</b>
25.1 TBA How much finished . . . . .	419
25.2 Introduction . . . . .	419
25.3 Mark rating pairs . . . . .	420
25.4 FROC notation . . . . .	421
25.5 The empirical FROC . . . . .	424
25.6 The inferred ROC plot . . . . .	427
25.7 The alternative FROC (AFROC) plot . . . . .	430
25.8 The weighted-AFROC (wAFROC) plot . . . . .	431
25.9 The AFROC1 plot . . . . .	432
25.10 The weighted-AFROC1 (wAFROC1) plot . . . . .	433
25.11 The EFROC plot . . . . .	433
25.12 Discussion . . . . .	434
25.13 References . . . . .	434

CONTENTS	13
<b>26 Empirical plot examples</b>	<b>435</b>
26.1 TBA How much finished . . . . .	435
26.2 Introduction . . . . .	435
26.3 Raw FROC/AFROC/ROC plots . . . . .	435
26.4 The chance level FROC and AFROC . . . . .	443
26.5 Location-level “true-negatives” . . . . .	445
26.6 Binned FROC/AFROC/ROC plots . . . . .	446
26.7 Structure of the binned data . . . . .	447
26.8 Summary . . . . .	451
26.9 Discussion . . . . .	451
26.10 References . . . . .	451
<b>27 FROC vs. wAFROC</b>	<b>453</b>
27.1 TBA How much finished . . . . .	453
27.2 Introduction . . . . .	453
27.3 FROC vs. wAFROC . . . . .	453
27.4 Summary of simulations . . . . .	460
27.5 Effect size comparison . . . . .	461
27.6 Performance depends on $\zeta_1$ . . . . .	462
27.7 Discussion . . . . .	463
27.8 References . . . . .	463
<b>28 Meanings of FROC figures of merit</b>	<b>465</b>
28.1 TBA How much finished . . . . .	465
28.2 Introduction . . . . .	465
28.3 Empirical AFROC FOM-statistic . . . . .	467
28.4 Empirical weighted-AFROC FOM-statistic . . . . .	468
28.5 Two Theorems . . . . .	469
28.6 Numerical illustrations . . . . .	471
28.7 Summary tables of ratings . . . . .	473
28.8 AFROC plot from first principles . . . . .	475
28.9 wAFROC plot from first principles . . . . .	478

28.10Physical interpretations . . . . .	478
28.11Discussion . . . . .	480
28.12References . . . . .	481
<b>29 Visual Search</b>	<b>483</b>
29.1 TBA How much finished . . . . .	483
29.2 Introduction . . . . .	483
29.3 Grouping and labeling ROIs . . . . .	484
29.4 Recognition vs. detection . . . . .	484
29.5 Search vs. classification . . . . .	487
29.6 Two visual search paradigms . . . . .	488
29.7 Determining where the radiologist looks . . . . .	491
29.8 The Kundel - Nodine search model . . . . .	491
29.9 Kundel-Nodine model and CAD algorithms . . . . .	496
29.10Simultaneously acquired eye-tracking and FROC data . . . . .	497
29.11Discussion / Summary . . . . .	500
29.12References . . . . .	501
<b>30 The radiological search model</b>	<b>503</b>
30.1 TBA How much finished . . . . .	503
30.2 Introduction . . . . .	503
30.3 The radiological search model . . . . .	504
30.4 RSM assumptions . . . . .	505
30.5 Summary of RSM . . . . .	506
30.6 Physical interpretation of RSM parameters . . . . .	506
30.7 Model re-parameterization . . . . .	512
30.8 Discussion / Summary . . . . .	513
30.9 References . . . . .	514

<b>CONTENTS</b>	<b>15</b>
<b>31 Radiological search model predictions</b>	<b>515</b>
31.1 TBA How much finished . . . . .	515
31.2 Introduction . . . . .	515
31.3 Inferred integer ROC ratings . . . . .	516
31.4 Constrained end-point property . . . . .	517
31.5 The RSM-predicted ROC curve . . . . .	520
31.6 The RSM-predicted FROC curve . . . . .	532
31.7 The RSM-predicted AFROC curve . . . . .	533
31.8 Discussion / Summary . . . . .	538
31.9 References . . . . .	542
<b>32 Search and classification performances</b>	<b>543</b>
32.1 TBA How much finished . . . . .	543
32.2 Introduction . . . . .	543
32.3 Quantifying search performance #rsm-search-search-performance} . . . . .	544
32.4 Quantifying classification performance . . . . .	545
32.5 Discussion / Summary . . . . .	547
32.6 References . . . . .	551
<b>33 The FROC should not be used to measure performance</b>	<b>553</b>
33.1 TBA How much finished . . . . .	553
33.2 Introduction . . . . .	553
33.3 The FROC curve is a poor descriptor of search performance . . . . .	554
33.4 Discussion / Summary . . . . .	558
33.5 References . . . . .	561
<b>34 Analyzing FROC data</b>	<b>563</b>
34.1 TBA How much finished . . . . .	563
34.2 Introduction . . . . .	563
34.3 Example 1 . . . . .	564
34.4 Plotting wAFROC and ROC curves . . . . .	566
34.5 Reporting an FROC study . . . . .	567

34.6 Crossed-treatment analysis . . . . .	568
34.7 Discussion / Summary . . . . .	570
34.8 References . . . . .	571
<b>35 FROC sample size</b>	<b>573</b>
35.1 TBA How much finished . . . . .	573
35.2 Introduction . . . . .	573
35.3 Example 1 . . . . .	575
35.4 Plotting wAFROC and ROC curves . . . . .	576
35.5 FitRsmROC usage example . . . . .	578
35.6 Discussion / Summary . . . . .	578
35.7 References . . . . .	579
<b>36 RSM fitting</b>	<b>581</b>
36.1 TBA How much finished . . . . .	581
36.2 Introduction . . . . .	581
36.3 FROC likelihood function . . . . .	583
36.4 IDCA Likelihood function . . . . .	585
36.5 ROC Likelihood function . . . . .	590
36.6 FitRsmROC implementation . . . . .	592
36.7 FitRsmROC usage example . . . . .	593
36.8 Discussion / Summary . . . . .	594
36.9 References . . . . .	595
<b>37 Three proper ROC fits</b>	<b>597</b>
37.1 TBA How much finished . . . . .	597
37.2 Introduction . . . . .	597
37.3 Applications . . . . .	598
37.4 Displaying composite plots . . . . .	599
37.5 Displaying RSM parameters . . . . .	600
37.6 Displaying CBM parameters . . . . .	602
37.7 Displaying PROPROC parameters . . . . .	603

<b>CONTENTS</b>	<b>17</b>
37.8 Overview of findings . . . . .	604
37.9 Discussion / Summary . . . . .	610
37.10 Appendices . . . . .	612
37.11 Datasets . . . . .	612
37.12 Location of PROPROC files . . . . .	615
37.13 Location of pre-analyzed results . . . . .	617
37.14 Plots for Van Dyke dataset . . . . .	619
37.15 References . . . . .	619
 <b>CAD</b>	 <b>627</b>
<b>38 Standalone CAD vs. Radiologists</b>	<b>627</b>
38.1 TBA How much finished . . . . .	627
38.2 Abstract . . . . .	627
38.3 Keywords . . . . .	628
38.4 Introduction . . . . .	628
38.5 Methods . . . . .	629
38.6 Software implementation . . . . .	636
38.7 Results . . . . .	638
38.8 Discussion . . . . .	641
38.9 Appendix . . . . .	642
38.10 References . . . . .	646
 <b>39 Optimal operating point on FROC</b>	 <b>647</b>
39.1 TBA How much finished . . . . .	647
39.2 Introduction . . . . .	647
39.3 Methods . . . . .	648
39.4 Using the method . . . . .	662
39.5 An application . . . . .	662
39.6 Discussion . . . . .	665
39.7 References . . . . .	665

<b>40 Localization - classification tasks</b>	<b>667</b>
40.1 TBA How much finished . . . . .	667
40.2 Introduction . . . . .	667
40.3 Abbreviations . . . . .	667
40.4 History and basic idea . . . . .	667
40.5 First example, File1.xlsx . . . . .	668
40.6 Second example, File2.xlsx . . . . .	670
40.7 Third example, File3.xlsx . . . . .	671
40.8 Fourth example, File4.xlsx . . . . .	671
40.9 Fifth example, File5.xlsx . . . . .	673
40.10 Precautions . . . . .	674
40.11 Discussion . . . . .	674
40.12 References . . . . .	674
<b>41 Split Plot Study Design</b>	<b>675</b>
41.1 TBA How much finished . . . . .	675
41.2 Mean Square R(T) . . . . .	675
41.3 References . . . . .	675

# Preface

- This book is currently (as of April 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 labeled (book), 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 `RJafrocBook.pdf` file.

## 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: Basics: this covers the basics of ROC methods
- Part III: Significance Testing: TBA
- Part IV: FROC paradigm: TBA

## Contributing to this book

- I appreciate constructive feedback on this document, e.g., corrections, comments, etc.
- To do this raise an Issue on the GitHub interface.
- Click on the Issues tab under dpc10ster/RJafrocBook, then click on New issue.
- When done this way, contributions from users automatically become part of the GitHub documentation/history of the book.

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

- Diagnostic imaging system evaluation
- Detection
- Detection combined with localization
- Detection combined with localization and 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

- TBA
- 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")



# Quick Start



# Chapter 1

## Help

### 1.1 TBA How much finished

30% (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 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 <sup>1</sup>.

## 2.4 Contents of Excel file

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

---

<sup>1</sup>They are needed for generalization to other data collection paradigms and for better data entry error control

<sup>2</sup>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	G	H
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			
10								
11								
12								
13								
14								
15								
16								
17								
18								

- 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					
25	2	1	3	2					

FP      TP      TRUTH      +

Average: 2.1    Count: 124    Sum: 126

- 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.

AutoSave OFF

rocCr

Share Comments

	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaselD	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				
25	4	0	73	1	2				

- 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
#> ..$ 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

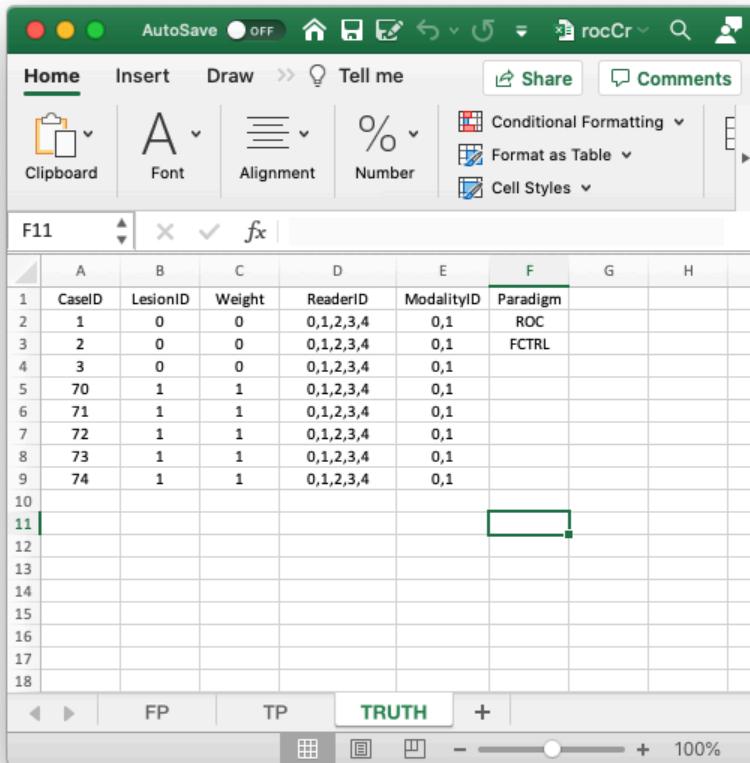
### 3.1 TBA How much finished

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



	A	B	C	D	E	F	G	H
1	CaselID	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			
10								
11								
12								
13								
14								
15								
16								
17								
18								

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 ...
#> ...$ 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.<sup>1</sup>
- 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$NL` member, with dimension [2, 5, 8, 1], contains the ratings of normal cases. The first dimension (2) is the number of treatments, the second (5) is the number of readers and the third (8) is the total number of cases. For ROC datasets the fourth dimension is always unity. The five extra values<sup>2</sup> in the third dimension, which are filled with NAs, are needed for compatibility with FROC datasets.
  - The `x$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.

<sup>1</sup>The assumptions underlying the “automatic” determination could be defeated by data entry errors.

<sup>2</sup>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					
25	2	1	3	2					

FP      TP      TRUTH      +

Average: 2.1    Count: 124    Sum: 126

- 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 always 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				
25	4	0	73	1	2				

- 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.

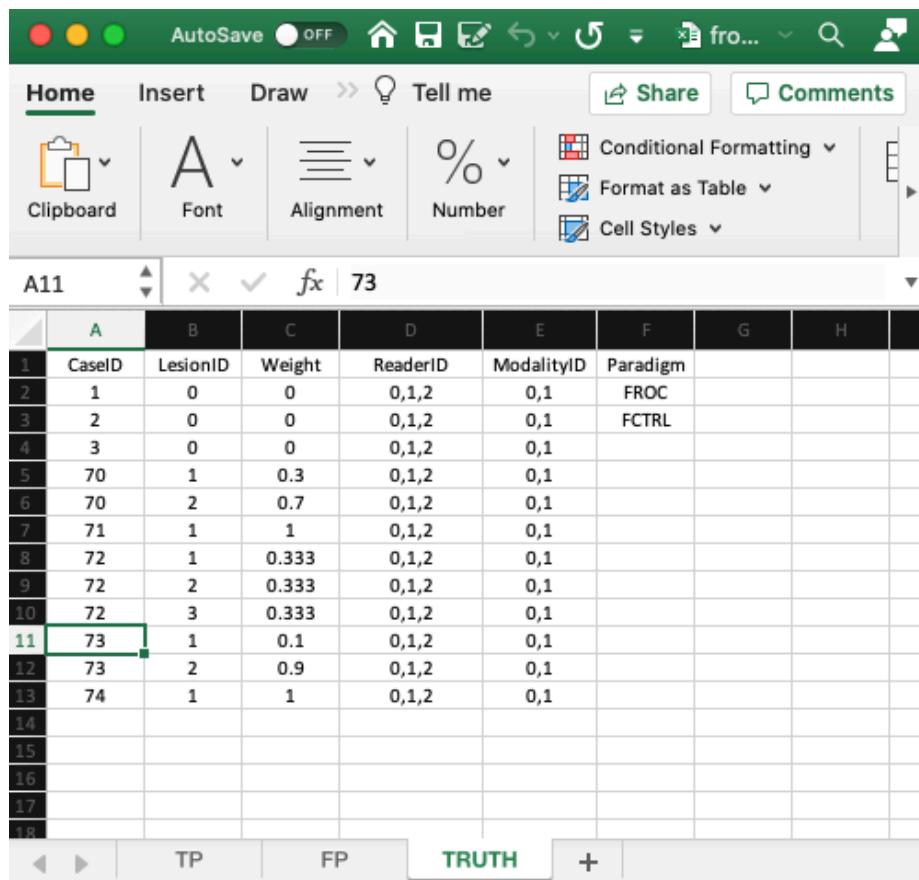
The purpose of this chapter is to:

- 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	G	H
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			
14								
15								
16								
17								
18								

The screenshot shows the Microsoft Excel interface with the 'Home' tab selected. The ribbon includes 'AutoSave OFF', 'Clipboard', 'Font', 'Alignment', 'Number', 'Share', and 'Comments'. The status bar shows 'A11', 'fx | 73', and '73'. The bottom navigation bar includes 'TP', 'FP', 'TRUTH' (which is highlighted in green), and a '+' button.

- 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.<sup>1</sup>
- **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 **LesionIDs** ‘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 **LesionIDs** 1, 2 and 3 and weights 1/3 each.
  - Diseased case 73 has two lesions, with **LesionIDs** 1, and 2 and weights 0.1 and 0.9.
  - Diseased case 74 has one lesion, with **LesionID** = 1 and **Weight** = 1.

---

<sup>1</sup>**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 ...
#>   ..$ 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. Lack-  
ing 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				

TP

- 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  $9 \times 2 \times 3 = 54$ . 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 ...
#>   ..$ LL    : num [1:4, 1:5, 1:115, 1:20] -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"
```

#### 4.7. ON THE DISTRIBUTION OF NUMBERS OF LESIONS IN DISEASED CASES51

```
#> ..$ 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 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, "lesions = ", length(which(x$lesions$perCase == el))/length(x$lesions$perCase))
#> 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 -Inf
#> [2,] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
#> [3,] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
#> [4,] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
#> [5,] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
#> [6,] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
#> [7,] 0.14285714 -Inf -Inf -Inf -Inf -Inf -Inf -Inf
#> [8,] 0.12500000 0.12500000 -Inf -Inf -Inf -Inf -Inf -Inf
#> [9,] 0.11111111 0.11111111 0.11111111 -Inf -Inf -Inf -Inf -Inf
#> [10,] 0.10000000 0.10000000 0.10000000 0.10000000 -Inf -Inf -Inf -Inf
#> [11,] 0.09090909 0.09090909 0.09090909 0.09090909 0.09090909 -Inf -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 0.05000000
#>          [,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 -Inf
#> [6,] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
#> [7,] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
#> [8,] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
#> [9,] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
#> [10,] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
#> [11,] -Inf -Inf -Inf -Inf -Inf -Inf -Inf -Inf
#> [12,] -Inf -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

## DBM analysis text output

### 5.1 TBA How much finished

50%

### 5.2 Introduction

This chapter illustrates significance testing using the DBM method.

### 5.3 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")
```

### 5.4 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
#> $ trtMeanDiff:'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 the four observers in the two treatments.
  - trtMeans is a [2x1] dataframe: the average figure of merit over all readers for each treatment.
  - trtMeanDiff 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$FOMs$trtMeanDiff
#>             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$RRRC)
#> 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
#>   ..$ StdErrr : 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
```

- FRRRC, 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$FRRRC$FTests
#>           DF        MS      FStat        p
#> Treatment  1 0.023565410 0.36395597 0.54769704
#> Error      99 0.064747968          NA         NA
ret$FRRRC$ciDiffTrt
#>                      Estimate      StdErrr DF        t      PrGTt
#> trtTREAT1-trtTREAT2 0.010854817 0.017992772 99 0.60328764 0.54769704
#>                               CILower     CIUpper
#> trtTREAT1-trtTREAT2 -0.024846746 0.04655638
ret$FRRRC$ciAvgRdrEachTrt
#>                      Estimate      StdErrr DF        CILower     CIUpper
#> trtTREAT1 0.84774989 0.027109386 99 0.79395898 0.90154079
#> trtTREAT2 0.83689507 0.027448603 99 0.78243109 0.89135905
ret$FRRRC$ciDiffTrtEachRdr
#>                      Estimate      StdErrr 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 treatment 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
```

## **5.5 References**

# Chapter 6

## OR analysis text output

### 6.1 TBA How much finished

90%

### 6.2 Introduction

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

### 6.3 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")
```

### 6.4 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<sub>2</sub>, 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<sub>1</sub> 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 0.0015227779 0.00047229915
#> trtTREAT2  3 0.00024159915 0.0015287746 0.00049497620

ret$ANOVA$IndividualRdr
#>           DF  msTEachRdr  varEachRdr  cov1EachRdr
#> rdrREADER_1  1 7.3897606e-06 0.0014771675 0.00056158020
#> rdrREADER_2  1 2.3077021e-04 0.0015186058 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<sub>1</sub>** for individual readers.

```

ret$FRC$FTests
#>              MS      Chisq DF      p
#> Treatment 0.0002356541 0.32101347 1 0.57099922
#> Error     0.0007340941      NA NA      NA
ret$FRC$ciDiffTrt
#>                  Estimate      StdErr      z      PrGTz      CILower
#> trtTREAT1-trtTREAT2 0.010854817 0.019158472 0.56658051 0.57099922 -0.026695098
#>                      CIUpper
#> trtTREAT1-trtTREAT2 0.048404732
ret$FRC$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$FRC$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$FRC$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 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$RFC$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 Excel output

### 7.1 TBA How much finished

90%

### 7.2 Introduction

This chapter illustrates significance testing using the OR method. But, instead of the perhaps unwieldy output in Chapter 6, it generates an Excel output file containing the following worksheets:

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

### 7.3 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 `ReportFileName` 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".

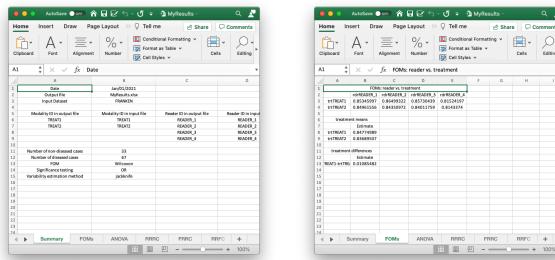


Figure 7.1: ‘Summary’ and ‘FOMs’ worksheets of Excel file ‘R/quick-start/MyResults.xlsx’

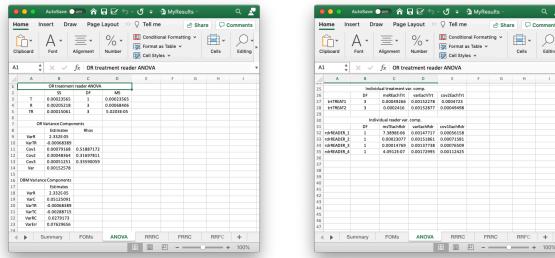


Figure 7.2: ‘ANOVA’ worksheet of Excel file ‘R/quick-start/MyResults.xlsx’

## 7.4 References

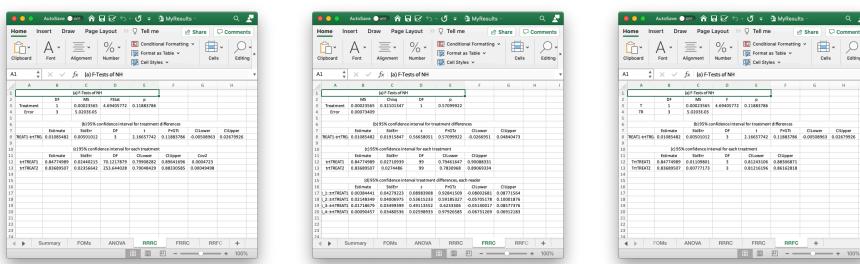


Figure 7.3: ‘RRRC’, ‘FRRC’ and ‘RRFC’ worksheets of Excel file ‘R/quick-start/MyResults.xlsx’



# **ROC paradigm**



# **Chapter 8**

## **Preliminaries**

### **8.1 TBA How much finished**

95%

### **8.2 Introduction**

The question addressed by this book is “how good are radiologists using medical imaging devices at diagnosing disease?” Observer performance measurements, widely used for this purpose, require data collection and analyses methods that fall under the rubric of what is loosely termed “ROC analysis”, where ROC is an abbreviation for Receiver Operating Characteristic (Metz, 1978). ROC analysis and its extensions form a specialized branch of science encompassing knowledge of diagnostic medical physics, perception of stimuli (commonly studied by psychologists), human observer modeling and statistics. Its importance in medical imaging is due to the evolution of technology and the need to objectively assess advances. The Food and Drug Administration, Center for Devices and Radiological Health (FDA/CDRH), which regulates medical-imaging devices, requires ROC studies as part of its device approval process . There are, conservatively, at least several hundred publications using ROC studies and a paper (Metz, 1978) by the late Prof. C.E. Metz has been cited over 1800 times. Numerous reviews and tutorial papers have appeared (Metz, 1978, Metz (1989), Kundel et al. (2008), Metz (1986)) and there are books on the statistical analysis (Zhou et al., 2002) of ROC data. However, in spite of the numbers of publications and books in this field, and in my experience, basic aspects of it are sometimes misunderstood, and lessons from the past have been sometimes forgotten, and these have seriously held back health care advances – as will be demonstrated in this book.

It is the aim of this book to describe the field in some depth while assuming little statistical background of the reader. That is a tall order. Key to accomplishing this aim is the ability to illustrate abstract statistical concepts and analysis methods with free, cross-platform, open-source software R, a programming language, and RStudio, “helper” software that makes it much easier to work with R, is very popular in the scientific community.

This chapter provides background material and an overview of the book. It starts with diagnostic interpretations occurring everyday in hospitals. The process of imaging device development by manufacturers is described, stressing the role of physical measurements in optimizing the design. Once the device is deployed, medical physicists working in hospitals use phantom quality control measurements to maintain image quality. Lacking the complexity of clinical images, phantom measurements may not correlate with clinical image quality. Model observers, that reduce the imaging process to mathematical formulae, are intended to bridge the gap. However, since they are yet restricted to simple tasks, where the location of possible lesions is known, their potential is yet to be realized. Unlike physical, phantom and model observer measurements, observer performance methods measure the net effect of the entire imaging chain, including the critical role of the radiologist. Four observer performance paradigms are described. Physical and observer performance methods are put in the context of a hierarchy of efficacy levels, where the measurements become increasingly difficult, but more clinically meaningful, as one moves to higher levels. An overview of the book is presented and suggestions are made on how to best use it.

### 8.3 Clinical tasks

In hospital based radiology departments or freestanding imaging centers, imaging studies are conducted to diagnose patients for signs of disease. Examples are chest x-rays, computerized tomography (CT) scans, magnetic resonance imaging (MRI) scans, ultrasound (US) imaging, etc. A patient does not go directly to a radiology department; rather, the patient first sees a family doctor, internist or general practitioner about an ailment. After a physical examination, perhaps augmented with non-imaging tests (blood tests, electrocardiogram, etc.) the physician may recommend an imaging study. As an example, a patient suffering from persistent cough yielding mucus and experiencing chills may be referred for chest x-rays to rule out pneumonia. In the imaging suite a radiologic technician properly positions the patient with respect to the x-ray beam. Chest x-rays are taken, usually in two projections, back to front (posterior-anterior or PA-view) and sideways (lateral or LAT-view).

Each x-ray image is a projection from, ideally a point source of x-rays, of patient anatomy in the path of the beam, onto a detector, e.g., x-ray film or digital detector. Because of differential attenuation, the shadow cast by the x-rays shows anatomical structures within the patient. The technician checks the

images for proper positioning and technical quality. A radiologist (a physician who specializes in interpreting imaging studies) interprets them and dictates a report.

Because of the referring physician's report, the radiologist knows why the patient was sent for chest x-rays in the first place, and interprets the image in that context. At the very outset one recognizes that images are not interpreted in a "vacuum", rather, for a symptomatic patient, the interpretation is done in the context of resolving a specific ailment. This is an example of a clinical task and it should explain why different specialized imaging devices are needed in a radiology department. Radiology departments in the US are usually organized according to body parts, e.g., a chest section, a breast imaging section, an abdominal imaging section, head CT, body CT, cardiac radiology, orthopedic radiology, etc. Additionally, for a given body part, different means of imaging are generally available. Examples are x-ray mammography, ultrasound and magnetic resonance imaging of the breast.

### 8.3.1 Workflow in an imaging study

The workflow in an imaging study can be summarized as follows. The patient's images are acquired. Nowadays almost all images in the US are acquired digitally, but some of the concepts are illustrated with analog images; this is not an essential distinction. The digital detector acquired image(s) are processed for optimality and displayed on one or more monitors. These are interpreted by a radiologist in the context of the clinical task implied by the referring physicians notes attached to the imaging request (such as "rule out pneumonia"). After interpreting the image(s), the radiologist makes a diagnosis, such as "patient shows no signs of disease" or "patient shows signs of disease". If signs of disease are found, the radiologist's report will contain a description of the disease and its location, extent, and other characteristics, e.g., "diffuse opacity near the bottom of the lungs, consistent with pneumonia". Alternatively, an unexpected finding can occur, such as "nodular lesion, possibly lung cancer, in the apex of the lungs". A diseased finding will trigger further imaging, e.g., a CT scan, and perhaps biopsy (excision of a small amount of tissue and examination by a pathologist to determine if it is malignant), to determine the nature and extent of the disease. In this book the terms non-diseased and diseased are used instead of "normal" and "abnormal", or "noise" and "signal plus noise", or "target absent" and "target present", etc.

So far, patients with symptoms of disease were considered. Interpreting images of asymptomatic patients involves an entirely different clinical task, termed "screening", described next.

### 8.3.2 The screening and diagnostic workup tasks

In the US, women older than 40 years are imaged at yearly intervals using a special x-ray machine designed to optimally image the breast. Here the radiologist's task is to find breast cancer, preferably when it is small and has not had an opportunity to spread, or metastasize, to other organs. Cancers found at an early stage are more likely to be treatable. Fortunately, the incidence of breast cancer is very low, about five per thousand women in the US, but, because most of the patients are non-diseased, this makes for a difficult task. The images are interpreted in context. The family history of the patient is available, the referring physician (the woman's primary care physician and / or gynecologist) has performed a physical examination of the patient, and in some cases it may be known whether the patient is at high-risk because she has a gene that predisposes her to breast cancer. The interpreting radiologist has to be MQSA-certified (Mammography Quality Standards Act) to interpret mammograms. If the radiologist finds one or more regions suspicious for breast cancer, the location of each suspicious region is recorded, as it provides a starting point for subsequent patient management. At my previous institution, The University of Pittsburgh, the images are electronically marked (annotated) on the digital images. The patient receives a dreaded letter or e-mail, perhaps preceded by a phone call from the imaging center, that she is being "recalled" for further assessment. When the woman arrives at the imaging center, further imaging, termed a diagnostic workup, is conducted. For example, magnification views, centered on the location of the suspicious region found at screening, may be performed. Magnifying the image reveals more detail. Additional x-ray projections and other types of imaging (e.g., ultrasound, MRI and perhaps breast CT – still in the research stage) may be used to resolve ambiguity regarding true disease status. If the suspicious region is determined to be benign, the woman goes home with the good news. This is the most common outcome. If ambiguity remains, a somewhat invasive procedure, termed a needle biopsy, is performed whereby a small amount of tissue is extracted from the suspicious region and sent to the pathology laboratory for final determination of malignancy status by a pathologist. Even here, the more common outcome is that the biopsy comes back negative for malignancy. About ten percent of women who are screened by experts are recalled for unnecessary diagnostic workups, in the sense that the diagnostic workup and / or biopsy end up showing no signs of cancer. These recalls cause some physical and much emotional trauma, and result in increased health care costs. About four of every five cancers are detected by experts, i.e., about 1 in 5 is missed. All of these numbers are for experts – there is considerable variability in skill-levels between MQSA-certified radiologists. If cancer is found radiation, chemotherapy or surgery may be initiated to treat the patient. Further imaging is usually performed to determine the response to therapy (has the tumor shrunk?).

The practice of radiology, and patients served by this discipline, has benefited tremendously from technological innovations. How these innovations are devel-

oped and adopted by radiology departments is the next topic.

## 8.4 Imaging device development and its clinical deployment

Roentgen's 1895 discovery of x-rays found almost immediate clinical applications and started the new discipline of radiology. Initially, two developments were key: optimizing the production of x-rays, as the process is very inefficient, and efficiently detecting the photons that pass through the imaged anatomy: these photons form the radiological image. Consequently, initial developments were in x-ray tube and screen-film detector technologies. Over many decades these have matured and new modalities have emerged, examples of which are CT in the late 1960s, MRI in the 1970s, computed radiography and digital imaging in the late 1980s.

### 8.4.1 Physical measurements

There is a process to imaging device development and deployment into clinical practice. The starting point is to build a prototype of the new imaging device. The device is designed in the context of a clinical need and is based on physical principles suggesting that the device, perhaps employing new technology or new ideas, should be an improvement over what is already available, generically termed the conventional modality. The prototype is actually the end-point of much research involving engineers, imaging scientists and radiologists.

The design of the prototype is optimized by physical measurements. For example, images are acquired of a block of Lucite<sup>TM</sup>, termed a "phantom", with thickness equivalent in x-ray penetrability to an average patient. Ideally, the images would be noise free, but x-ray quantum fluctuations and other sources of noise influence the final image and cause them to have noise, termed radiographic mottle[16-18]. For conventional x-rays, the kind one might see the doctor putting up on a viewing panel (light box) in old movies, the measurement employs a special instrument called a microdensitometer, which essentially digitizes narrow strips of the film. The noise is quantified by the standard deviation of the digitized pixel values. This is compared to that expected based on the number of photons used to make the image; the latter number can be calculated from knowledge of the x-ray beam spectrum and the thickness of the phantom. If the measured noise equals the expected noise (if it is smaller, there is obviously something wrong with the calculation of the expected noise and / or the measurement), image quality is said to be quantum limited. Since a fundamental limit, dictated by the underlying imaging physics, has been reached, further noise reduction is only possible by increasing the number of photons. The latter can be accomplished trivially by increasing the exposure time, which, of course,

increases radiation dose to the patient. Therefore, as far as image noise is concerned, in this scenario, the system is ideal and no further noise optimization is needed. In my experience teaching imaging physics to radiology residents, the preceding sentences cause confusion. In particular, the terms limited and ideal seem to be at odds, but the residents eventually understand it. The point is that if one is up against a fundamental limit, then things are ideal in the sense that they can get no better (physicists do have a sense of humor). In practice this level of perfection is never reached, as the screen-film system introduces its own noise, due to the granularity of the silver halide crystals that form the photographic emulsion and other factors – ever tried digitizing an old slide? Furthermore, there could be engineering limitations preventing attainment of the theoretical limit. Through much iteration, the designer reaches a point at which it is decided that the noise is about as low as it is going to get.

Noise is but one factor limiting image quality. Another factor is spatial resolution – the ability of an imaging system to render sharp edges and/or resolve closely spaced small objects. For this measurement, one increases the number of photons (to minimize noise), or uses a thinner Lucite™ block superposed on an object with a sharp edge, e.g., a razor blade. When the resulting image is scanned with a microdensitometer, the trace should show an abrupt transition as one crosses the edge of the phantom. In practice, the transition is rounded or spread out, resembling a sigmoid function. This is due to several factors. The finite size of the focal spot producing the x-rays produces a penumbra effect, which blurs the edge. The spread of light, within the screen due to its finite thickness, also blurs the edge. The screen absorbs photons and converts them to visible light to which film is exquisitely sensitive. Without the screen, the exposure would have to increase about thousand fold. One can make the screen only so thin, because then it would lack the ability to stop the x-rays that have penetrated the phantom. These photons contain information regarding the imaged anatomy. Ideally, all photons that form the radiological image should be stopped in the detector. Again, an optimization process is involved until the equipment designer is convinced that a fundamental limit has been reached or engineering limitations prevent further improvement.

Another factor affecting image quality is contrast – the ability of the imaging system to depict different levels of x-ray penetration. A phantom consisting of a step wedge, with varying thickness of Lucite™ is imaged and the image scanned with a microdensitometer. The resulting trace should show distinct steps as one crosses the different thickness parts of the step-wedge phantom (termed large area contrast, to distinguish it from the blurring occurring at the edges between the steps). The more steps that can be visualized, the better the system. The digital term for this is the gray-scale. For example, an 8-bit gray scale can depict 256 shades of gray. Once again design considerations and optimization is used to arrive at the design of the prototype.

The preceding is a simplified description of possible physical measurements. In fact, it is usual to measure the spatial frequency dependence of resolution,

noise and overall photon usage efficiency[19, 20]. These involve quantities named modulation transfer function (MTF), noise power spectrum (NPS) and detective quantum efficiency (DQE), each of which is a function of spatial frequency ( $f$ , in cycles per mm). The frequency dependence is important in understanding, during the development process, the factors limiting image quality.

After an optimized prototype has been made it needs approval from the FDA/CDRH for pre-clinical usage. This involves submitting information about the results of the physical measurements and making a case that the new design is indeed an improvement over existing methods. However, since none of the physical measurements involved radiologists interpreting actual patient images produced by the prototype, observer performance measurements are needed before machines based on the prototype can be marketed. Observer performance measurements, in which the prototype is compared to an existing standard, involve a group of about five or six radiologists interpreting a set of patient images acquired on the prototype and on the conventional modality. The truth (is the image of a diseased patient?) is unknown to them but is known to the researcher, i.e., the radiologist is “blinded” to the truth. The radiologists’ decisions, classified by the investigator as correct or incorrect, are used to determine the average performance of the radiologists on the prototype and on the existing standard. Specialized statistical analysis is needed to determine if the difference in performance is in the correct direction and “statistically significant”, i.e., unlikely to be due to chance. The measurements are unique in the sense that the entire imaging chain is being evaluated. In order to get a sufficiently large and representative sample of patients and radiologists, such studies are generally performed in a multi-institutional setting[21]. If the prototype’s performance equals or exceeds that of the existing standard, it is approved for clinical usage. At this point, the manufacturer can start marketing the device to radiology departments. This is a simplified description of the device approval process. Most imaging companies have experts in this area that help them negotiate a necessarily more complex process.

#### 8.4.2 Quality Control and Image quality optimization

Once the imaging device is sold to a radiology department, both routine quality control (QC) and continuous image quality optimization are needed to assure proper utilization of the machine over its life span. The role of QC is to maintain image quality at an established standard. Initial QC measurements, termed acceptance testing[22-24], are made to establish base-line QC parameters and a medical physicist establishes a program of systematic checks to monitor them. The QC measurements are relatively simple, typically taking a few hours of technologist time, that look for changes in monitored variables. The role of continuous image quality optimization, which is the bread-and-butter of a diagnostic medical physicist, is to resolve site-specific image quality issues. The manufacturer cannot anticipate every issue that may arise when their equipment

is used in the field, and it takes a medical physicist, working in collaboration with the equipment manufacturer, technologists and radiologists, to continually optimize the images and solve specific image quality related problems. Sometimes the result is a device that performs better than what the manufacturer was able to achieve. One example, from my experience, is the optimization, using special filters and an air-gap technique, of a chest x-ray machine in the 1980s by Prof. Gary T. Barnes, a distinguished medical physicist and the late Prof. Robert Fraser, a famous chest radiologist[25]. The subsequent evaluation of this machine vs. a prototype digital chest x-ray machine by the same manufacturer, Picker International, was my entry into the field of observer performance [26].

A good example of QC is the use of the American College of Radiology Mammography Quality Standards Act (ACR-MQSA) phantom to monitor image quality of mammography machines[27-29]. The phantom consists of a (removable) wax insert in an acrylic holder; the latter provides additional absorption and scattering material to more closely match the attenuation and beam hardening of an average breast. Embedded in the wax insert are target objects consisting of 6 fibrils, five groups of microcalcifications, each containing six specks, and five spherical objects of different sizes, called masses. An image of the phantom, Fig. 8.1 (A) is obtained daily, before the first patient is imaged, and is inspected by a technologist, who records the number of target objects of different types that are visible. There is a pass-fail criterion and if the image fails then patients cannot be imaged on that machine until the problem is corrected. At this point, the medical physicist is called in to investigate.

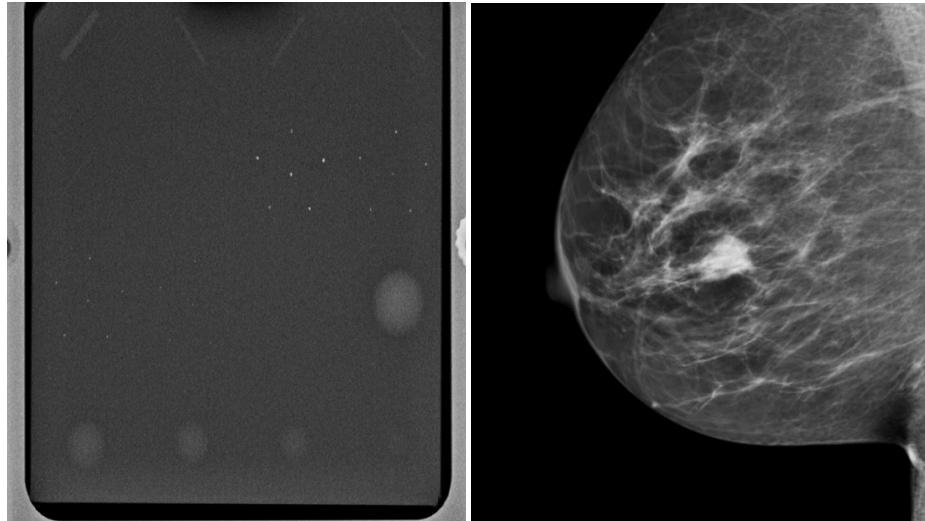


Figure 8.1: (A) Image of an ACR phantom, (B) Clinical image.

#### 8.4. IMAGING DEVICE DEVELOPMENT AND ITS CLINICAL DEPLOYMENT 85

Fig. 8.1 (A – B): (A) Image of an American College of Radiology mammography accreditation phantom. The phantom contains target objects consisting of six fibrils, five groups of microcalcifications, and five nodule-like objects. An image of the phantom is obtained daily, before the first patient is imaged, and is inspected by a technologist, who records the number of target objects of different types that are visible. On his 27" iMac monitor, I see four fibrils, three speck groups and four masses, which would be graded as a "pass". This is greatly simplified version of the test. The scoring accounts for irregular fibril or partially visible masses borders, etc., all of which is intended to get more objectivity out of the measurement. (B) A breast image showing an invasive cancer, located roughly in the middle of the image. Note the lack of similarity between the two images (A) and (B). The breast image is much more complex and there is more information, and therefore more to go wrong than with the phantom image. Moreover, there is variability between patients in contrast to the fixed image in (A). In my clinical experience, the phantom images interpreted visually are a poor predictor of clinical image quality.

One can perhaps appreciate the subjectivity of the measurement. Since the target locations are known, the technologist can claim to have detected it and the claim cannot be disproved; unless a claim is falsifiable, it is not science. While the QC team is trained to achieve repeatable measurements, I have shown TBA [30-34] that computer analysis of mammography phantom images (CAMPPI) can achieve far greater precision and repeatability than human observer readings. Commercial software is currently available from various vendors that perform proprietary analysis of phantom images for various imaging systems (e.g., mammography machines, CT scanners, MRI scanners, ultrasound, etc.).

Fig. 8.1 (B) shows a mammogram with a mass-like cancer visible near its center. It is characterized by complex anatomical background, quite unlike the uniform background in the phantom image in Fig. 8.1 (A). In mammography 30% of retrospectively visible lesions are missed at initial screening and radiologist variability can be as large as 40% [35]. QC machine parameters (e.g., kVp, the kilovoltage accuracy) are usually measured to 1% accuracy. It is ironic that the weak link, in the sense of greatest variability, is the radiologist but quality control and much effort is primarily focused on measuring/improving the physical parameters of the machine. This comment is meant to motivate clinical medical physicists, most of who are focused on QC, to become more aware about observer performance methods, where achieving better than 5% accuracy is quite feasible[36]. The author believes there should be greater focus on improving radiologist performance, particularly those with marginal performance. Efforts in this direction, using ROC methods, are underway in the UK [37, 38] by Prof Alistair Gale and colleagues.

## 8.5 Image quality vs. task performance

In this book, “image quality” is defined as the fidelity of the image with respect to some external gold standard of what the ideal image should look like, while “task performance” is how well a radiologist, using the image, accomplishes a given clinical task. For example, if one had an original Rembrandt and a copy, the image quality of the copy is perfect if even an expert appraiser cannot distinguish it from the original. The original painting is the “gold standard”. If an expert can distinguish the copy from the original, its image quality is degraded. The amount of degradation is related to the ease with which the expert can detect the fraud.

A radiological image is the result of x-rays interactions within the patient and the image receptor. Here it is more difficult to define a gold standard. If it exists at all, the gold standard is expected to depend on what the image is being used for, i.e., the diagnostic task. An image suitable for soft-tissue disease diagnosis may not be suitable for diagnosis of bone disease. This is the reason why CT scanners have different soft-tissue and bone window/level settings. With clinical images, a frequently used approach is to have an expert rank-order the images, acquired via different methods, with respect to “clinical appropriateness” or “clinical image quality”. The quotes are used to emphasize that these terms are hard to define objectively. In this approach, the gold standard is in the mind of the expert. Since experts have typically interpreted tens of thousands of images in the past, and have lived with the consequences of their decisions, there is considerable merit to using them to judge clinical image quality. However, experts do disagree and biases cannot be ruled out. This is especially true when a new imaging modality is introduced. The initial introduction of computed radiography (CR) was met with some resistance in the US among technologists, who had to learn a different way of obtaining the images that disrupted their workflow. There was also initial resistance from more experienced radiologists, who were uncomfortable with the appearance of the new images, i.e., their gold standard was biased in favor of the modality – plain films – that they were most familiar. The author is aware of at least one instance where CR had to be imposed by “diktat” from the Chairman of the department. Some of us are more comfortable reading printed material than viewing it on a computer screen, so this type of bias is understandable.

Another source of bias is patient variability, i.e., the gold standard depends on the patient. Some patients are easier to image than others are in the sense that their images are “cleaner”, i.e., they depict anatomical structures that are known to be present more clearly. X-rays pass readily through a relatively slim patient (e.g., an athlete) and there are fewer scattered photons which degrade image quality[39, 40], than when imaging a larger patient (e.g., an NFL linebacker). The image of the former will be clearer, the ribs, the heart shadow, the features of the lungs, etc., will be better visualized (i.e., closer to what is expected based on the anatomy) than the image of the linebacker. Similar differences

## 8.6. WHY PHYSICAL MEASURES OF IMAGE QUALITY ARE NOT ENOUGH<sup>87</sup>

exist in the ease of imaging women with dense breasts, containing a larger fraction of glandular tissue compared to women with fatty breasts. By imaging appropriately selected patients, one can exploit these facts to make one's favorite imaging system look better. [Prof. Harold Kundel, one of my mentors, used to say: "Tell me which modality you want to come out better and I will prepare a set of patient images to help you make your case".]

### 8.6 Why physical measures of image quality are not enough

Both high spatial resolution and low noise are desirable characteristics. However, imaging systems do not come unambiguously separated as high spatial resolution and low noise vs. low spatial resolution and high noise. There is generally an intrinsic imaging physics dictated tradeoff between spatial resolution and noise. Improving one makes the other worse. For example, if the digital image is smoothed with, for example, with a spatial filter, then noise will be smaller, because of the averaging of neighboring pixels, but the ability to resolve closely spaced structures will be compromised. Therefore, a more typical scenario is deciding whether the decreased noise justifies the accompanying loss of spatial resolution. Clearly the answer to this depends on the clinical task: if the task is detecting relatively large low contrast nodules, then some spatial smoothing may actually be beneficial, but if the task involves detecting small microcalcifications, often the precursors of cancer in the breast, then the smoothing will tend to reduce their visibility.

The problem with physical measures of image quality lies in relating them to clinical performance. Phantom images have little resemblance to clinical images, compare Fig. 8.1 (A) and (B). X-ray machines generally have automatic exposure control: the machines use a brief exposure to automatically sense the thickness of the patient from the detected x-rays. Based on this, the machine chooses the best combinations of technical factors (kVp and tube charge) and image processing. The machine has to be put in a special manual override mode to obtain reasonable images of phantoms, as otherwise the exposure control algorithm, which expects patient anatomy, is misled by the atypical nature of the "patient", compared to typical patient anatomy, into producing very poor phantom images. This type of problem makes it difficult to reproduce problems encountered using clinical images with phantom images. It has been my general experience that QC failures often lag clinical image quality reported problems: more often than not, clinical image quality problems are reported before QC measurements indicate a problem. This is not surprising since clinical images, e.g., Fig. 8.1 (B) are more complex and have more information[41], both in the clinical and in the information theoretic sense[42], than the much simpler phantom image shown in Fig. 8.1 (A), so there is more that can go wrong with clinical images than with phantom images. Manufacturers now design an-

thropomorphic phantoms whose images resemble human x-rays. Often these phantoms provide the option of inserting target objects at random locations; this is desired to get more objectivity out of the measurement. Now, if the technologist claims to have found the target, the indicated location can be used to determine if the target was truly detected.

To circumvent the possibility that changes in physical measurements on phantoms may not sensitively track changes in clinical image interpretations by radiologists, a measurement needs to include both the complexity of clinical images and radiologists as part of the measurement. Because of variability in both patient images and radiologist interpretations, such measurements are expected to be more complicated than QC measurements, so to be clear, I am not advocating observer performance studies as part of QC. However, they could be built into a continuous quality improvement program, perhaps performed annually. Before giving an overview of the more complex methods, an alternative modeling driven approach, that is widely used, is described next.

## 8.7 Model observers

If one can adequately simulate (or model) the entire imaging process, then one can design mathematical measurements that can be used to decide if a new imaging system is an improvement over a conventional imaging system. Both new and conventional systems are modeled (i.e., reduced to formulae that can be evaluated). The field of model observers[43] is based on assuming this can be done. The FDA/CDRH has a research program called VICTRE: Virtual Imaging Clinical Trials for Regulatory Evaluation. Since everything is done on a computer, the method does not require time-consuming studies involving radiologists.

A simple example may elucidate the process (for more details one should consult the extensive literature on model observers). Suppose one simulates image noise by sampling a Gaussian random number generator and filling up the pixels in the image with the random samples. This simulates a non-diseased image. The number of such images could be quite large, e.g., 1000, limited only by one's patience. A second set of simulated diseased images is produced in which one samples a random number generator to create non-diseased images, as before, but this time one adds a small low-contrast but noiseless disk, possibly with Gaussian edges, to the center of each image. The procedure yields two sets of images, 1000 with noise only backgrounds and 1000 with different noise backgrounds and the superposed centered low contrast disk. One constructs a template whose shape is identical to that of the superposed disk (i.e., one does not simply measure peak contrast at the center of the lesion; rather the shape-dependent contrast of the disk is taken into account). One then calculates the cross-correlation of the template with each of the superposed disks[30, 44]. The cross correlation is the sum of the products of pixel values of corresponding

pixels, one drawn from the template and the other drawn from the matching position on the disk image. [Details of this calculation are in Online Appendix 12.B of Chapter 12.] Because of random noise, the cross-correlations from different simulated diseased cases will not be identical, and one averages the 1000 values. Next one applies the template to the centers of the non-diseased images and computes the cross correlations as before. Because of the absence of the disk, the values will be smaller (assuming positive disk contrast). The difference between the average of the cross-correlations at disk locations and the average at disk-absent locations is the numerator of a signal to noise ratio (SNR) like quantity. The denominator is the standard deviation of the cross-correlations at disk-free locations. To be technical, the procedure yields the signal-to-noise-ratio (SNR) of the non-pre-whitening ideal observer[45]. It is an ideal mathematical “observer” in the sense that for white noise no human observer can surpass this level of performance[46, 47].

Suppose the task is to evaluate two image-processing algorithms. One applies each algorithm to the 2000 images described above and measures SNR for each algorithm. The one yielding the higher SNR, after accounting for variability in the measurements, is the superior algorithm.

Gaussian noise images are not particularly “clinical” in appearance. If one filters the noise appropriately, one can produce simulated images that are similar to non-diseased backgrounds observed in mammography[48-50]. Other techniques exist for simulating statistically characterized lumpy backgrounds that are a closer approximation to some medical images[51].

Having outlined one of the alternatives, one is ready for the methods that form the subject matter of this book.

## 8.8 Measuring observer performance: four paradigms

Observer performance measurements come in different “flavors”, types or paradigms. In the current context, a paradigm is an agreed-upon method for collecting the data. A given paradigm can lend itself to different analyses. In historical order the paradigms are: (1) the receiver operating characteristic (ROC) paradigm [1, 2, 7, 52, 53]; (2) the free-response ROC (FROC) paradigm [54, 55]; (3) the location ROC (LROC) paradigm [56, 57] and (4) the region of interest (ROI) paradigm [58]. Each paradigm assumes that the truth is known independently of the modalities to be compared. This implies that one cannot use diagnoses from one of the modalities to define truth – if one did, the measurement would be biased in favor of the modality used to define truth. It is also assumed that the true disease status of the image is known to the researcher but the radiologist is “blinded” to this information.

In the ROC paradigm the observer renders a single decision per image. The decision could be communicated using a binary scale (ex. 0 or 1) or declared by use of the terms “negative” or “positive,” abbreviations of “negative for disease” (the radiologist believes the patient is non-diseased) and “positive for disease” (the radiologist believes the patient is diseased), respectively. Alternatively, the radiologist could give an ordered numeric label, termed a rating, to each case where the rating is a number with the property that higher values correspond to greater radiologist’s confidence in presence of disease. A suitable ratings scale could be the consecutive integers 1 through 6, where “1” is “definitely non-diseased” and “6” is “definitely diseased”.

If data is acquired on a binary scale, then the performance of the radiologist can be plotted as a single operating point on an ROC plot. The x-axis of the plot is false positive fraction (FPF), i.e., the fraction of non-diseased cases incorrectly diagnosed as diseased. The y-axis of the plot is true positive fraction (TPF), i.e., the fraction of diseased cases correctly diagnosed as diseased. Models have been developed to fit binary or multiple rating datasets. These models predict continuous curves, or operating characteristics, along which an operating point can move by varying the radiologist’s reading style. The reading style is related to the following concept: based on the evidence in the image, how predisposed is a radiologist to declaring a case as diseased. A “lenient”, “lax” or “liberal” reporting style radiologist is very predisposed even with scant evidence. A “strict” or “conservative” reporting style radiologist requires more evidence before declaring a patient as diseased. This brief introduction to the ROC was given to explain the term “operating characteristic” in ROC. The topic is addressed in more detail in Chapter 02.

In the FROC paradigm the observer marks and rates all regions in the image that are sufficiently suspicious for disease. A mark is the location of the suspicious region and the rating is an ordered label, characterizing the degree of suspicion attached to the suspicious region. In the LROC paradigm the observer gives an overall ROC-type rating to the image, and indicates the location of the most suspicious region in the image. In the ROI paradigm the researcher divides each image into a number of adjacent non-overlapping regions of interest (ROIs) that cover the clinical area of interest. The radiologist’s task is to evaluate each ROI for presence of disease and give an ROC-type rating to it.

### 8.8.1 Basic approach to the analysis

The basic approach is to obtain data, according to one of the above paradigms, from a group of radiologists interpreting a common set of images in one or more modalities. The way the data is collected, and the structure of the data, depends on the selected paradigm. The next step is to adopt an objective measure of performance, termed a figure of merit (FOM) and a procedure for estimating it for each modality-reader combination. Assuming two modalities, e.g., a new modality and the conventional one, one averages FOM over all readers within

each modality. If the difference between the two averages (new modality minus the conventional one) is positive, that is an indication of improvement. Next comes the statistical part: is the difference large enough so as to be unlikely to be due to chance. This part of the analysis, termed significance testing, yields a probability, or p-value, that the observed difference or larger could result from chance even though the modalities have identical performances. If the p-value is very small, that it is taken as evidence that the modalities are not identical in performance, and if the difference is in the right direction, the new modality is judged better.

### 8.8.2 Historical notes

The term “receiver operating characteristic” (ROC) traces its roots to the early 1940s. The “receiver” in ROC literally denoted a pulsed radar receiver that detects radio waves bounced off objects in the sky, the obvious military application being to detect enemy aircraft. Sometimes the reflections were strong compared to receiver electronic noise and other sources of noise and the operator could confidently declare that the reflection indicated the presence of aircraft and the operator was correct. This combination of events was termed a true positive (TP). At other times the aircraft was present but due to electronic noise and reflections off clouds, the operator was not confident enough to declare “aircraft present” and this combination of events was termed a false negative (FN). Two other types of decisions can be discerned when there was no aircraft in the field of view: (1) the operator mistook reflections from clouds or perhaps a flock of large birds and declared “aircraft present”, termed a false positive (FP). (2) The operator did not declare “aircraft present” because the reflected image was clear of noise or false reflections and the operator felt confident in a negative decision, termed a true negative (TN). Obviously, it was desirable to maximize correct decisions (TPs and TNs) while minimizing incorrect decisions (FNs and FPs). Scientists working on this problem analyzed it as a generic signal detection problem, where the signal was the aircraft reflection and the noise was everything else. A large field called signal detection theory (SDT) emerged[59]. However, even at this early stage, it must have been apparent to the researchers that the problem was incomplete in a key respect: when the operator detects a suspicious signal, there is a location (specifically an azimuth and altitude associated with it). The operator could be correct in stating “aircraft present” but direct the interceptors to the wrong location. Additionally, there could be multiple enemy aircraft present, but the operator is only allowed the “aircraft present” and “aircraft absent” responses, which fail to allow for multiplicity of suspected aircraft locations. This aspect was not recognized, to the best of my knowledge, until Egan coined the term “free-response” in the auditory detection context[54].

Having briefly introduced the different paradigms, two of which, namely the ROC and the FROC, will be the focus of this book, it is appropriate to see how

Table 8.1: FrybackThornbury hierarchy of efficacies.

Level Designation	Essential Characteristic
1. Technical efficacy	Engineering measures: MTF, NPS, DQE
2. Diagnostic accuracy efficacy	Sensitivity, specificity, ROC or FROC area
3. Diagnostic thinking efficacy	Positive and negative predictive values
4. Therapeutic efficacy	Treatment benefits from imaging test?
5. Patient outcome efficacy	Patients benefit from imaging test?
6. Societal efficacy	Society benefits from imaging test?

these measurements fit in with the different types of measurements possible in assessing imaging systems.

## 8.9 Hierarchy of assessment methods

The methods described in this book need to be placed in context of a six-level hierarchy of assessment methods[7, 60]. The cited paper by Fryback and Thornbury on “The Efficacy of Diagnostic Imaging” is a highly readable account, which also gives a more complete overview of this field, including key contributions by Yerushalmy[61] and Lusted[62]. The term efficacy is defined generically as “the probability of benefit to individuals in a defined population from a medical technology applied for a given medical problem under ideal conditions of use”. Demonstration of efficacy at each lower level is a necessary but not sufficient condition to assure efficacy at higher level. The different assessment methods are, in increasing order of efficacy : technical, diagnostic accuracy, diagnostic thinking, therapeutic, patient outcome and societal, Table 8.1.

Table 8.1: Fryback and Thornbury proposed hierarchy of assessment methods. Demonstration of efficacy at each lower level is a necessary but not sufficient condition to assure efficacy at higher level. [MTF = modulation transfer function; NPS(f) = noise power spectra as a function of spatial frequency f; DQE(f) = detective quantum efficiency]

The term “clinical relevance” is used rather loosely in the literature. The author is not aware of an accepted definition of “clinical relevance” apart from its obvious English language meaning. As a working definition I have proposed [63] that the clinical relevance of a measurement be defined as its hierarchy-level. A level-5 patient outcome measurement (do patients, on the average, benefit from the imaging study) is clinically more relevant than a technical measurement like noise on a uniform background phantom or an ROC study. This is because it relates directly to the benefit, or lack thereof, to a group of patients (it is impossible to define outcome efficacy at the individual patient level – at the

patient level outcome is a binary random variable, e.g., 1 if the outcome was good or 0 if the outcome was bad).

One could make physical measurements ad-infinitum, but one cannot (yet) predict the average benefit to patients. Successful virtual clinical trials would prove me wrong. ROC studies are more clinically relevant than physical measurements, and it is more likely that a modality with higher performance will yield better outcomes, but it is not a foregone conclusion. Therefore, higher-level measurements are needed.

However, the time and cost of the measurement increases rapidly with the hierarchy level. Technical efficacy, although requiring sophistical mathematical methods, take relatively little time. ROC and FROC, both of which are level-2 diagnostic accuracy measurements, take more time, often a few months to complete. However, since ROC measurements include the entire imaging chain and the radiologist, they are more clinically relevant than technical measurements, but they do not tell us the effect on diagnostic thinking. After the results of “live” interpretations are available, e.g., patients are diagnosed as diseased or non-diseased, what does the physician do with the information. Does the physician recommend further tests or recommends immediate treatment. This is where the level-3 measurements come in, which measure the effect on diagnostic thinking. Typical level-3 measurements are positive predictive value (PPV) and negative predictive value (NPV). PPV is the probability that the patient is actually diseased when the diagnosis is diseased and NPV is the probability that the patient is actually non-diseased when the diagnosis is non-diseased. These are discussed in more detail in Chapter 02.

Unlike level-2 measurements, PPV and NPV depend on disease prevalence. As an example consider breast cancer which (fortunately) has low prevalence, about 0.005. Before the image is interpreted and lacking any other history, the mammographer knows only there is a five in 1000 chance that the woman has breast cancer. After the image is interpreted, the mammographer has more information. If the image was interpreted as diseased, the confidence in presence of cancer increases. For an expert mammographer typical values of sensitivity and specificity are 80% and 90%, respectively (these terms will be explained in the next chapter; sensitivity is identical to true positive fraction and specificity is 1-false positive fraction). It will be shown (in Chapter 02, §2.9.2) that for this example PPV is only 0.04. In other words, even though an expert interpreted the screening mammogram as diseased, the chance that the patient actually has cancer is only 4%. Obviously more tests are needed before one knows for sure if the patient has cancer – this is the reason for the recall and the subsequent diagnostic workup referred to in §1.2.2. The corresponding NPV is 0.999. Negative interpretations by experts are definitely good news for the affected patients and these did not come directly from an ROC study, or physical measurements, rather they came from actual “live” clinical interpretations. Again, NPV and PPV are defined as averages over a group of patients. For example, the 4% chance of cancer following a positive diagnosis is good news, on the average.

An unlucky patient could be one of the four-in-100-patients that has cancer following a positive screening diagnosis.

While more relevant than ROC, level-3 measurements like PPV and NPV are more difficult to conduct than ROC studies [18] – they involve following, in real time, a large cohort of patients with images interpreted under actual clinical conditions. Level 4 and higher measurements, namely therapeutic, patient outcome and societal, are even more difficult and are sometimes politically charged, as they involve cost benefit considerations.

## 8.10 Overview of the book and how to use it

For the most part the book follows the historical development, i.e., it starts with chapters on ROC methodology, chapters on significance testing, chapters on FROC methodology, chapters on advanced topics and appendices. Not counting Chapter 01, the current chapter, the book is organized five Parts (A - E).

### 8.10.1 Overview of the book

#### 8.10.1.1 Part A: The ROC paradigm

Part A describes the ROC (receiver operating characteristic) paradigm. Chapter 02 describes the binary decision task. Terminology that is important to master, such as accuracy, sensitivity, specificity, disease prevalence, positive and negative predictive values is introduced. Chapter 03 introduces the important concepts of decision variable, the reporting threshold, and how the latter may be manipulated by the researcher and it introduces the ROC curve. Chapter 04 reviews the widely used ratings method for acquiring ROC data. Chapter 06 introduces the widely used binormal model for fitting ratings data. The chapter is heavy on mathematical and computational aspects, as it is intended to take the mystery out of these techniques, which are used in subsequent chapters. The data fitting method, pioneered by Dorfman and Alf in 1969, is probably one of the most used algorithms in ROC analysis. Chapter 07 describes sources of variability affecting any performance measure, and how they can be estimated.

#### 8.10.1.2 Part B: The statistics of ROC analysis

Part B describes the specialized statistical methods needed to analyze ROC data, in particular how to analyze data originating from multiple readers interpreting the same cases in multiple modalities. Chapter 08 introduces hypothesis-testing methodology, familiar to statisticians, and the two types of errors that the researcher wishes to control, the meaning of the ubiquitous p-value and statistical power. Chapter 09 focuses on the Dorfman-Berbaum-Metz method,

with improvements by Hillis. Relevant formulae, mostly from publications by Prof. Steven Hillis, are reproduced without proofs (it is my understanding that Dr. Hillis is working on a book on his specialty, which should nicely complement the minimalistic-statistical description approach adopted in this book). Chapter 10 describes the Obuchowski-Rockette method of analyzing MRMC ROC data, with Hillis' improvements. Chapter 11 describes sample size estimation in an ROC study.

#### 8.10.1.3 Part C: The FROC paradigm

Part C is unique to this book. Anyone truly wishing to understand human observer visual search performance needs to master it. The payoff is that the concepts, models and methods described here apply to almost all clinical tasks. Chapter 17 and Chapter 18 are particularly important. These were difficult chapters to write and they will take extra effort to comprehend. However, the key findings presented in these chapters and their implications should strongly influence future observer performance research. If the potential of the findings is recognized and used to benefit patients, by even one reader, I will consider this book a success. Chapter 19 describes how to analyze FROC data and report the results.

#### 8.10.1.4 Part D: Advanced topics

Some of the chapters in Part D are also unique to this book. Chapter 20 discusses proper ROC curve fitting and software. The widely used bivariate binormal model, developed around 1980, but never properly documented, is explained in depth, and a recent extension of it that works with any dataset is described in Chapter 21. Also described is a method for comparing (standalone) CAD to radiologists, Chapter 22. Standalone CAD performance is rarely measured, which is a serious mistake, for which we are all currently paying the price. It does not work for masses in mammography[64-66]. In the UK CAD is not used, instead they rely on double readings by experts, which is actually the superior approach, given the current low bar used in the US for CAD to be considered a success. Chapter 23, co-authored by Mr. Xuetong Zhai, a graduate student, describes validation of the CAD analysis method described in Chapter 22. It describes constructing a single-modality multiple-reader ratings data simulator. The method is extendible to multiple-modality multiple-reader datasets.

#### 8.10.1.5 Part E: Appendices (TBA)

Part E contains two online chapters. Online Chapter 24 is a description of 14 datasets, all but 2 of them collected by me over years of collaborations with researchers who conducted the studies and on which I helped with analysis and sometimes with manuscript preparation. The datasets provide a means to

demonstrate analysis techniques and to validate fitting methods. Finally, Online Chapter 25, co-authored by Mr. Xuetong Zhai, is a user-manual for the RJafroc package. Since RJafroc is used extensively in the book, this is expected to be a useful “go-to” chapter for the reader. The choice to put these chapters online is to allow me to update the datasets with new files as they become available and to update the documentation of RJafroc as new features are added.

### **8.10.2 How to use the book**

Each chapter consists of the physical book chapter that one is reading. Additionally, there are good chances that the online directory corresponding to this book will contain two directories, one called software and the other called Supplementary Material. The software directory contains “ready to run” code that is referenced in the book chapter text. When one sees such a reference in a chapter, the reader should open the relevant file and run it. Detailed directions are provided in the Online Appendix corresponding to each chapter.

Those new to the field should read the chapters in sequence. It is particularly important to master Part A. Part B presents the statistical analysis at a level accessible to the expected readers of this book, namely the user community. The only way to really understand this part is to apply the described methods and codes to the online datasets. Understanding the formulae in this part, especially those relating to statistical hypothesis testing, requires statistical expertise, which could lead the average reader in unproductive directions. It is best to accept the statisticians’ formulae and confirm that they work. How to determine if a method “works” will be described. Readers with prior experience in the field may wish to “skim” chapters. If they do, it is strongly recommended that they at least run and understand the software examples. This will prepare them for the more complex code in later chapters.

This concludes the introduction of the book.

## **8.11 Summary**

## **8.12 Discussion**

## **8.13 References**

# Chapter 9

## The Binary Task

### 9.1 TBA How much finished

85%

### 9.2 Introduction

In the previous chapter four observer performance paradigms were introduced: the receiver operating characteristic (ROC), the free-response ROC (FROC), the location ROC (LROC) and the region of interest (ROI). The next few chapters focus on the ROC paradigm, where each case is rated for confidence in presence of disease. While a multiple point rating scale is generally used, in this chapter it is assumed that the ratings are binary, and the allowed values are “1” vs. “2”. Equivalently, the ratings could be “non-diseased” vs. “diseased”, “negative” vs. “positive”, etc. In the literature this method of data acquisition is also termed the “yes/no” procedure (Green and Swets, 1966; Egan, 1975). The reason for restricting, for now, to the binary task is that the multiple rating task can be shown to be equivalent to a number of simultaneously conducted binary tasks. Therefore, understanding the simpler method is a good starting point.

Since the truth is also binary, this chapter could be named the binary-truth binary-decision task. The starting point is a  $2 \times 2$  table summarizing the outcomes in such studies and useful fractions that can be defined from the counts in this table, the most important ones being true positive fraction (TPF) and false positive fraction (FPF). These are used to construct measures of performance, some of which are desirable from the researcher’s point of view, but others are more relevant to radiologists. The concept of disease prevalence is introduced and used to formulate relations between the different types of measures. An

Table 9.1: Truth Table.

	T=1	T=2
D=1	TN	FN
D=2	FP	TP

R example of calculation of these quantities is given that is only slightly more complicated than the demonstration in the prior chapter.

### 9.3 The fundamental 2x2 table

In this book, the term case is used for images obtained for diagnostic purposes, of a patient; often multiple images of a patient, sometimes from different modalities, are involved in an interpretation; all images of a single patient, that are used in the interpretation, are collectively referred to as a case. A familiar example is the 4-view presentation used in screening mammography, where two views of each breast are available for viewing.

Let  $D$  represent the radiologist's decision, with  $D = 1$  representing the decision "case is non-diseased" and  $D = 2$  representing the decision "case is diseased". Let  $T$  denote the truth with  $T = 1$  representing "case is actually non-diseased" and  $T = 2$  representing "case is actually diseased". Each decision, one of two values, will be associated with one of two truth states, resulting in an entry in one of 4 cells arranged in a  $2 \times 2$  layout, termed the decision vs. truth table, Table 9.1, which is of fundamental importance in observer performance. The cells are labeled as follows. The abbreviation  $TN$ , for true negative, represents a  $D = 1$  decision on a  $T = 1$  case.  $FN$ , for false negative, represents a  $D = 1$  decision on a  $T = 2$  case (also termed a "miss").  $FP$ , for false positive, represents a  $D = 2$  decision on a  $T = 1$  case (a "false-alarm") and  $TP$ , for true positive, represents a  $D = 2$  decision on a  $T = 2$  case (a "hit").

Table 9.2 shows the numbers of decisions in each of the four categories defined in Table 9.1. Specifically,  $n(TN)$  is the number of true negative decisions,  $n(FN)$  is the number of false negative decisions, etc. The last row is the sum of the corresponding columns. The sum of the number of true negative decisions  $n(TN)$  and the number of false positive decisions  $n(FP)$  must equal the total number of non-diseased cases, denoted  $K_1$ . Likewise, the sum of the number of false negative decisions  $n(FN)$  and the number of true positive decisions  $n(TP)$  must equal the total number of diseased cases, denoted  $K_2$ . The last column is the sum of the corresponding rows. The sum of the number of true negative  $n(TN)$  and false negative  $n(FN)$  decisions is the total number of negative decisions, denoted  $n(N)$ . Likewise, the sum of the number of false positive  $n(FP)$  and true positive  $n(TP)$  decisions is the total number of positive decisions, denoted  $n(P)$ . Since each case yields a decision, the bottom-right corner cell is

Table 9.2: Cell counts.

	T=1	T=2	RowSums
D=1	n(TN)	n(FN)	n(N)=n(TN)+n(FN)
D=2	n(FP)	n(TP)	n(P)=n(FP)+n(TP)
ColSums	$K_1 = n(TN) + n(FP)$	$K_2 = n(FN) + n(TP)$	$K = K_1 + K_2 = n(N) + n(P)$

$n(N) + n(P)$ , which must also equal  $K_1 + K_2$ , the total number of cases  $K$ . These statements are summarized in Eqn. (9.1).

$$\left. \begin{array}{l} K_1 = n(TN) + n(FP) \\ K_2 = n(FN) + n(TN) \\ n(N) = n(TN) + n(FN) \\ n(P) = n(TP) + n(FP) \\ K = K_1 + K_2 = n(N) + n(P) \end{array} \right\} \quad (9.1)$$

## 9.4 Sensitivity and specificity

The notation  $P(D|T)$  indicates the probability of diagnosis D given truth state T (the vertical bar symbol is used to denote a conditional probability, i.e., what is to the left of the vertical bar depends on the condition appearing to the right of the vertical bar being true).

$$P(D|T) = P(\text{diagnosis is D} | \text{truth is T}) \quad (9.2)$$

Therefore the probability that the radiologist will diagnose “case is diseased” when the case is actually diseased is  $P(D = 2|T = 2)$ , which is the probability of a true positive  $P(TP)$ .

$$P(TP) = P(D = 2|T = 2) \quad (9.3)$$

Likewise, the probability that the radiologist will diagnose “case is non-diseased” when the case is actually diseased is  $P(D = 1|T = 2)$ , which is the probability of a false negative  $P(FN)$ .

$$P(FN) = P(D = 1|T = 2) \quad (9.4)$$

The corresponding probabilities for non-diseased cases,  $P(TN)$  and  $P(FP)$ , are defined by:

$$\left. \begin{aligned} P(TN) &= P(D = 1|T = 1) \\ P(FP) &= P(D = 2|T = 1) \end{aligned} \right\} \quad (9.5)$$

Since the diagnosis must be either  $D = 1$  or  $D = 2$ , for each truth state the probabilities on non-diseased and diseased cases must sum to unity:

$$\left. \begin{aligned} P(D = 1|T = 1) + P(D = 2|T = 1) &= 1 \\ P(D = 1|T = 2) + P(D = 2|T = 2) &= 1 \end{aligned} \right\} \quad (9.6)$$

Equivalently, these equations can be written:

$$\left. \begin{aligned} P(TN) + P(FP) &= 1 \\ P(FN) + P(TP) &= 1 \end{aligned} \right\} \quad (9.7)$$

Comments:

- An easy way to remember Eqn. (9.7) is to start by writing down the probability of one of the four probabilities, e.g.,  $P(TN)$ , and “reversing” both terms inside the parentheses, i.e.,  $T \Rightarrow F$ , and  $N \Rightarrow P$ . This yields the term  $P(FP)$  which when added to the previous probability,  $P(TN)$ , yields unity, i.e., the 1st equation in Eqn. (9.7).
- Because there are two equations in four unknowns, only two of the four probabilities, one per equation, are independent. By tradition these are chosen to be  $P(D = 1|T = 1)$  and  $P(D = 2|T = 2)$ , i.e.,  $P(TN)$  and  $P(TP)$ , which happen to be the probabilities of correct decisions on non-diseased and diseased cases, respectively. The two basic probabilities are so important that they have names:  $P(D = 2|T = 2) = P(TP)$  is termed sensitivity (Se) and  $P(D = 1|T = 1) = P(TN)$  is termed specificity (Sp):

$$\left. \begin{aligned} Se &= P(TP) = P(D = 2|T = 2) \\ Sp &= P(TN) = P(D = 1|T = 1) \end{aligned} \right\} \quad (9.8)$$

The radiologist can be regarded as a diagnostic “test” yielding a binary decision under the binary truth condition. More generally, any test (e.g., a blood test for HIV) yielding a binary result (positive or negative) under a binary truth condition is said to be sensitive if it correctly detects the diseased condition most of the time. The test is said to be specific if it correctly detects the non-diseased condition most of the time. Sensitivity is how correct the test is at detecting a diseased condition, and specificity is how correct the test is at detecting a non-diseased condition.

### 9.4.1 Reasons for the names sensitivity and specificity

It is important to understand the reason for these names and an analogy may be helpful. Most of us are sensitive to temperature, especially if the choice is between ice-cold vs. steaming hot. The sense of touch is said to be sensitive to temperature. One can imagine some neurological condition rendering a person hypersensitive to temperature, such that the person responds “hot” no matter what is being touched. For such a person the sense of touch is not very specific, as it is unable to distinguish between the two temperatures. This person would be characterized by unit sensitivity (since the response is “hot” to all steaming hot objects) and zero specificity (since the response is never “cold” to ice-cold objects). Likewise, a different neurological condition could render a person hypersensitive to cold, and the response is “cold” no matter what is being touched. Such a person would have zero sensitivity (since the response is never “hot” when touching steaming hot) and unit specificity (since the response is “cold” when touching ice-cold). Already one suspects that there is an inverse relation between sensitivity and specificity.

### 9.4.2 Estimating sensitivity and specificity

Sensitivity and specificity are the probabilities of correct decisions, over diseased and non-diseased cases, respectively. The true values of these probabilities would require interpreting all diseased and non-diseased cases in the entire population of cases. In reality, one has a finite sample of cases and the corresponding quantities, calculated from this finite sample, are termed estimates. Population values are fixed, and in general unknown, while estimates are random variables. Intuitively, an estimate calculated over a larger number of cases is expected to be closer to the true or population value than an estimate calculated over a smaller number of cases.

Estimates of sensitivity and specificity follow from counting the numbers of TP and TN decisions in Table 2.2 and dividing by the appropriate denominators. For sensitivity, the appropriate denominator is the number of actually diseased cases, namely  $K_2$ , and for specificity, the appropriate denominator is the number of actually non-diseased cases, namely  $K_1$ . The estimation equations for sensitivity and specificity are (estimates are denoted by the “hat” or circumflex symbol  $\widehat{\cdot}$ ):

$$\left. \begin{aligned} \widehat{\text{Se}} &= \widehat{P(TP)} = \frac{n(TP)}{K_2} \\ \widehat{\text{Sp}} &= \widehat{P(TN)} = \frac{n(TN)}{K_1} \end{aligned} \right\} \quad (9.9)$$

The ratio of the number of TP decisions to the number of actually diseased cases is termed true positive fraction  $\widehat{TPF}$ , which is an estimate of sensitivity, or equivalently, an estimate of  $P(TP)$ . Likewise, the ratio of the number of TN

decisions to the number of actually non-diseased cases is termed true negative fraction  $\widehat{TNF}$ , which is an estimate of specificity, or equivalently, an estimate of  $P(\widehat{TN})$ . The complements of  $\widehat{TPF}$  and  $\widehat{TNF}$  are termed false negative fraction  $\widehat{FNF}$  and false positive fraction  $\widehat{FPF}$ , respectively.

## 9.5 Disease prevalence

Disease prevalence, often abbreviated to prevalence, is defined as the actual or true probability that a randomly sampled case is of a diseased patient, i.e., the fraction of the entire population that is diseased. It is denoted  $P(D|pop)$  when patients are randomly sampled from the population (“pop”) and otherwise it is denoted  $P(D|lab)$ , where the condition “lab” stands for a laboratory study, where cases may be artificially enriched, and thus not representative of the population value:

$$\left. \begin{array}{l} P(D|pop) = P(T = 2|pop) \\ P(D|lab) = P(T = 2|lab) \end{array} \right\} \quad (9.10)$$

Since the patients must be either diseased or non-diseased, it follows with either sampling method, that:

$$\left. \begin{array}{l} P(T = 1|pop) + P(T = 2|pop) = 1 \\ P(T = 1|lab) + P(T = 2|lab) = 1 \end{array} \right\} \quad (9.11)$$

If a finite number of patients are sampled randomly from the population the fraction of diseased patients in the sample is an estimate of true disease prevalence.

$$P(\widehat{D}|pop) = \frac{K_2}{K_1+K_2} \Big|_{pop} \quad (9.12)$$

It is important to appreciate the distinction between true (population) prevalence and laboratory prevalence. As an example, true disease prevalence for breast cancer is about five per 1000 patients in the US, but most mammography studies are conducted with comparable numbers of non-diseased and diseased cases:

$$\left. \begin{array}{l} P(\widehat{D}|pop) \sim 0.005 \\ P(\widehat{D}|lab) \sim 0.5 \gg P(\widehat{D}|pop) \end{array} \right\} \quad (9.13)$$

## 9.6 Accuracy

Accuracy is defined as the fraction of all decisions that are in fact correct. Denoting it by  $\widehat{Ac}$  one has for the corresponding estimate:

$$\widehat{Ac} = \frac{n(TN) + n(TP)}{n(TN) + n(TP) + n(FP) + n(FN)} \quad (9.14)$$

The numerator is the total number of correct decisions and the denominator is the total number of decisions. An equivalent expression is:

$$\widehat{Ac} = \widehat{Sp}\widehat{P}(\overline{!D}) + \widehat{Se}\widehat{P}(D) \quad (9.15)$$

The exclamation mark symbol is used to denote the “not” or negation operator. For example,  $P(\overline{!D})$  means the probability that the patient is not diseased. Eqn. (9.15) applies equally to laboratory or population studies, *provided sensitivity and specificity are estimated consistently*. One cannot combine a population estimate of prevalence with a laboratory measurement of sensitivity and / or specificity.

Eqn. (9.15) can be understood from the following argument.  $\widehat{Sp}$  is the fraction of correct (i.e., negative) decisions on non-diseased cases. Multiplying this by  $\widehat{P}(\overline{!D})$  yields  $\widehat{Sp}\widehat{P}(\overline{!D})$ , the fraction of correct negative decisions on all cases. Similarly,  $\widehat{Sp}$  is the fraction of correct positive decisions on all cases. Therefore, their sum is the fraction of (all, i.e., negative and positive) correct decisions on all cases. A formal mathematical derivation follows. The terms on the right hand side of Eqn. (9.9) can be “turned around” yielding:

$$\left. \begin{aligned} n(TP) &= K_2 \widehat{Se} \\ n(TN) &= K_1 \widehat{Sp} \end{aligned} \right\} \quad (9.16)$$

Therefore,

$$\begin{aligned} \widehat{Ac} &= \frac{n(TN) + n(TP)}{K} \\ &= \frac{K_1 \widehat{Sp} + K_2 \widehat{Se}}{K} \\ &= \widehat{Sp}\widehat{P}(\overline{!D}) + \widehat{Se}\widehat{P}(D) \end{aligned} \quad (9.17)$$

## 9.7 Negative and positive predictive values

Sensitivity and specificity have desirable characteristics insofar as they reward the observer for correct decisions on actually diseased and actually non-diseased cases, respectively, so these quantities are expected to be independent of disease prevalence; one is dividing by the relevant denominator, so increased numbers of non-diseased cases are balanced by a corresponding increased number of correct decisions on non-diseased cases, and likewise for diseased cases. However, radiologists interpret cases in a “mixed” situation where cases could be positive or negative for disease and disease prevalence plays a crucial role in their decision-making – this point will be clarified shortly. Therefore, a measure of performance that is desirable from the researcher’s point of view is not necessarily desirable from the radiologist’s point of view. It should be obvious that if most cases are non-diseased, i.e., disease prevalence is close to zero, specificity, being correct on non-diseased cases, is more important to the radiologist than sensitivity. Otherwise, the radiologist would figuratively be crying “wolf” most of the time. The radiologist who makes too many FPs would discover it from subsequent clinical audits or daily case conferences, which are held in most large imaging departments. There is a cost to unnecessary false positives – the cost of additional imaging and / or needle biopsy to rule out cancer, not to mention the pain and emotional trauma inflicted on the patient. Conversely, if disease prevalence is high, then sensitivity, being correct on diseased cases, is more important to the radiologist than specificity. With intermediate disease prevalence a weighted average of sensitivity and specificity, where the weighting involves disease prevalence, would appear to be desirable from the radiologist’s point of view.

The radiologist is less interested in the normalized probability of a correct decision on non-diseased cases. Rather interest is in the probability that a patient diagnosed as non-diseased is actually non-diseased. The reader should notice how the two probability definitions are “turned around” - more on this below. Likewise, the radiologist is less interested in the normalized probability of correct decisions on diseased cases; rather interest is in the probability that a patient diagnosed as diseased is actually diseased. These are termed negative and positive predictive values, respectively, and denoted NPV and PPV.

NPV is defined as the probability, given a non-diseased diagnosis, that the patient is actually non-diseased:

$$NPV = P(T = 1|D = 1) \quad (9.18)$$

PPV is defined as the probability, given a diseased diagnosis, that the patient is actually diseased:

$$PPV = P(T = 2|D = 2) \quad (9.19)$$

Note that both equations are “turned around” from the definition of specificity and sensitivity, Eqn. (9.8), i.e., specificity =  $P(D = 1|T = 1)$  and sensitivity =  $P(D = 2|T = 2)$ .

For now we focus on NPV. To estimate NPV one divides the number of correct negative decisions  $n(TN)$  by the total number of negative decisions  $n(N)$ . The latter is the sum of the number of correct negative decisions  $n(TN)$  and the number of incorrect negative decisions  $n(FN)$ . Therefore,

$$\widehat{NPV} = \frac{n(TN)}{n(TN) + n(FN)} \quad (9.20)$$

Dividing the numerator and denominator by the total number of negative cases, one gets:

$$\widehat{NPV} = \frac{\widehat{P(TN)}}{\widehat{P(TN)} + \widehat{P(FN)}} \quad (9.21)$$

The estimate of the probability of a TN equals the estimate of true negative fraction  $1 - \widehat{FPF}$  multiplied by the estimate that the patient is non-diseased, i.e.,  $\widehat{P(!D)}$ :

$$\widehat{P(TN)} = \widehat{P(!D)}(1 - \widehat{FPF}) \quad (9.22)$$

Explanation: A similar logic to that used earlier applies:  $(1 - \widehat{FPF})$  is the probability of being correct on non-diseased cases. Multiplying this by the estimate of probability of disease absence yields the estimate of  $\widehat{P(TN)}$ .

Likewise, the estimate of the probability of a FN equals the estimate of false negative fraction, which is  $(1 - \widehat{TPF})$ , multiplied by the estimate of the probability that the patient is diseased, i.e.,  $(\widehat{P(D)})$ :

$$\widehat{P(FN)} = \widehat{P(D)}(1 - \widehat{TPF}) \quad (9.23)$$

Putting this all together, one has:

$$\widehat{NPV} = \frac{\widehat{P(!D)}(1 - \widehat{FPF})}{(\widehat{P(!D)}(1 - \widehat{FPF}) + (\widehat{P(D)}(1 - \widehat{TPF}))} \quad (9.24)$$

For the population,

$$NPV = \frac{P(!D)(1 - FPF)}{(P(!D)(1 - FPF) + (P(D)(1 - TPF))} \quad (9.25)$$

Likewise, it can be shown that  $PPV$  is given by:

$$PPV = \frac{P(D)(TPF)}{P(D)(TPF) + P(!D)FPF} \quad (9.26)$$

The equations defining  $NPV$  and  $PPV$  are actually special cases of Bayes' theorem (Larsen and Marx, 2001). The general theorem is:

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(A)P(B|A)}{P(A)P(B|A) + P(!A)P(B|!A)} \end{aligned} \quad (9.27)$$

An easy way to remember Eqn. (9.27) is to start with the numerator on the right hand side, which is the “reversed” form of the desired probability on the left hand side, multiplied by an appropriate probability. For example, if the desired probability is  $P(A|B)$ , one starts with the “reversed” form, i.e.,  $P(B|A)$ , multiplied by  $P(A)$ . This yields the numerator. The denominator is the sum of two probabilities: the probability of B given A, i.e.,  $P(B|A)$ , multiplied by  $P(A)$  plus the probability of B given  $\neg A$ , i.e.,  $P(B|\neg A)$ , multiplied by  $P(\neg A)$ .

### 9.7.1 Example calculation of $PPV$ , $NPV$ and accuracy

- Typical disease prevalence in the US in screening mammography is 0.005.
- A typical operating point, for an expert mammographer, is  $FPF = 0.1$ ,  $TPF = 0.8$ . What are  $NPV$  and  $PPV$ ?

```
# disease prevalence in
# USA screening mammography
prevalence <- 0.005 # Line 3
FPF <- 0.1 # typical operating point
TPF <- 0.8 # do:
specificity <- 1-FPF
sensitivity <- TPF
NPV <- (1-prevalence)*(specificity) /
  ((1-prevalence)*(specificity) + # Line 8
   prevalence*(1-sensitivity))
PPV <- prevalence*sensitivity/ # Line 10
  (prevalence*sensitivity +
   (1-prevalence)*(1-specificity))
cat("NPV = ", NPV, "\nPPV = ", PPV, "\n")
#> NPV = 0.9988846
#> PPV = 0.03864734
```

```

accuracy <- (1-prevalence)*
(specificity)+(prevalence)*(sensitivity)
cat("accuracy = ", accuracy, "\n")
#> accuracy = 0.8995

```

- Line 3 initializes the variable `prevalence`, the disease prevalence, to 0.005.
- Line 4 assigns 0.1 to FPF and line 5 assigns 0.8 to TPF.
- Lines 6 and 7 initialize the variables specificity and sensitivity, respectively.
- Line 8 calculates NPV using Eqn. (9.25).
- Line 9 calculates PPV using Eqn. (9.26).

### 9.7.2 Comments

If a woman has a negative diagnosis, chances are very small that she has breast cancer: the probability that the radiologist is incorrect in the negative diagnosis is  $1 - \text{NPV} = 0.0011154$ . Even if she has a positive diagnosis, the probability that she actually has cancer is still only 0.0386473. That is why following a positive screening diagnosis the woman is recalled for further imaging, and if that reveals cause for reasonable suspicion, then additional imaging is performed, perhaps augmented with a needle biopsy to confirm actual disease status. If the biopsy turns out positive, only then is the woman referred for cancer therapy. Overall, accuracy is 0.8995. The numbers in this illustration are for expert radiologists. In practice there is wide variability in radiologist performance.

### 9.7.3 PPV and NPV are irrelevant to laboratory tasks

According to the hierarchy of assessment methods described in (book) Chapter 01, Table 1.1, PPV and NPV are level- 3 measurements, which are calculated from “live” interpretations (recall that the higher the level the greater the clinical relevance). In the clinic, the radiologist adjusts the operating point to achieve a balance between sensitivity and specificity. The balance depends critically on the known disease prevalence. Based on geographical location and type of practice, the radiologist over time develops an idea of actual disease prevalence, or it can be found in various databases. For example, a breast-imaging clinic that specializes in imaging high-risk women will have higher disease prevalence than the general population and the radiologist is expected to err more on the side of reduced specificity because of the expected benefit of increased sensitivity. However, in the context of a laboratory study, where one uses enriched case sets, the concepts of NPV and PPV are meaningless. For example, it would be rather difficult to perform a laboratory study with 10,000 randomly sampled women, which would ensure about 50 actually diseased patients, which is large enough to get a reasonably precise estimate of sensitivity (estimating specificity is inherently more precise because most women are actually non-diseased).

Rather, in a laboratory study one uses enriched data sets where the numbers of diseased-cases is much larger than in the general population, Eqn. (9.13). The radiologist cannot interpret these cases pretending that the actual prevalence is very low. Negative and positive predictive values, while they can be calculated from laboratory data, have very little, if any, clinical meanings, since they have no effect on radiologist thinking. As noted in (book) Chapter 01 the purpose of level-3 measurements is to determine the effect on radiologist thinking. There are no diagnostic decisions riding on laboratory ROC interpretations of retrospectively acquired patient images. However, PPV and NPV do have clinical meanings when calculated from very large population based “live” studies. For example, the (Fenton et al., 2007) study sampled 684,956 women and used the results of “live” interpretations of their images. In contrast, laboratory ROC studies are typically conducted with 50-100 non-diseased and 50-100 diseased cases. A study using about 300 cases total would be considered a “large” ROC study.

## 9.8 Summary

This chapter introduced the terms sensitivity (identical to TPF), specificity (the complement of FPF), disease prevalence, and positive and negative predictive values and accuracy. It is shown that, due to its strong dependence on disease prevalence, accuracy is a relatively poor measure of performance. Radiologists generally have a good, almost visceral, understanding of positive and negative predictive values, as these terms are relevant in the clinical context, being in effect, their “batting averages”. A caveat on the use of PPV and NPV calculated from laboratory studies is noted; these quantities only make sense in the context of “live” clinical interpretations.

## 9.9 Discussion

## 9.10 References

# Chapter 10

## Modeling the Binary Task

### 10.1 TBA How much finished

85%

### 10.2 Introduction

Chapter 10 introduced measures of performance associated with the binary decision task. Described in this chapter is a 2-parameter statistical model for the binary task, in other words it shows how one can predict quantities like sensitivity and specificity based on the values of the parameters of a statistical model. It introduces the fundamental concepts of a decision variable and a decision threshold (the latter is one of the parameters of the statistical model) that pervade this book, and shows how the decision threshold can be altered by varying experimental conditions. The receiver-operating characteristic (ROC) plot is introduced which shows how the dependence of sensitivity and specificity on the decision threshold is exploited by a measure of performance that is independent of decision threshold, namely the area AUC under the ROC curve. AUC turns out to be related to the other parameter of the model.

The dependence of variability of the operating point on the numbers of cases is explored, introducing the concept of random sampling and how the results become more stable with larger numbers of cases, or larger sample sizes. These are perhaps intuitively obvious concepts but it is important to see them demonstrated, Online Appendix 3.A. Formulae for 95percent confidence intervals for estimates of sensitivity and specificity are derived and the calculations are shown explicitly,

## 10.3 Decision variable and decision threshold

The model for the binary task involves three assumptions: (i) the existence of a decision variable associated with each case, (ii) the existence of a case-independent decision threshold for reporting individual cases as non-diseased or diseased and (iii) the adequacy of training session(s) in getting the observer to a steady state. In addition, common to all models is that the observer is “blinded” to the truth, while the researcher is not.

### 10.3.1 Existence of a decision variable

**Assumption 1:** Each case presentation is associated with the occurrence (or realization) of a specific value of a random scalar sensory variable yielding a unidirectional measure of evidence of disease. The two italicized phrases introduce important terms.

- By sensory variable one means one that is sensed internally by the observer (in the cognitive system, associated with the brain) and as such is not directly measurable in the traditional physical sense. A physical measurement, for example, might consist of measuring a voltage difference across two points with a voltmeter. The term “latent” is often used to describe the sensory variable because it turns out that transforming this variable by an arbitrary monotonic non-decreasing transformation has no effect on the ROC – this will become clearer later. Alternative terms are “psychophysical variable”, “perceived variable”, “perceptual variable” or “confidence level”. The last term is the most common. It is a subjective variable since its value is expected to depend on the observer: the same case shown to different observers could evoke different values of the sensory variable. Since one cannot measure it anyway, it would be a very strong assumption to assume that the two sensations are identical. In this book the term “latent decision variable”, or simply “decision variable” is used, which hopefully gets away from the semantics and focuses instead on what the variable is used for, namely making decisions. The symbol  $Z$  will be used for it and specific realized values are termed  $z$ -samples. It is a random in the sense that it varies randomly from case to case; unless the cases are similar in some respect, for example, two variants of the same case under different image processing conditions, or images of twins; in these instances the corresponding decision variables are expected to be correlated. In the binary paradigm model to be described, the decision variables corresponding to different cases are assumed mutually independent.
- The latent decision variable rank-orders cases with respect to evidence for presence of disease. Unlike a traditional rank-ordering scheme, where “1” is the highest rank, the scale is inverted with larger values corresponding

to greater evidence of disease. Without loss of generality, one assumes that the decision variable ranges from  $-\infty$  to  $+\infty$ , with large positive values indicative of strong evidence for presence of disease, and large negative values indicative of strong evidence for absence of disease. The zero value indicates no evidence for presence or absence of disease. [The  $-\infty$  to  $+\infty$  scale is not an assumption. The decision variable scale could just as well range from  $a$  to  $b$ , where  $a < b$ ; with appropriate rescaling of the decision variable, there will be no changes in the rank-orderings, and the scale will extend from  $-\infty$  to  $+\infty$ .] Such a decision scale, with increasing values corresponding to increasing evidence of disease, is termed positive-directed.

### 10.3.2 Existence of a decision threshold

**Assumption 2:** In the binary decision task the radiologist adopts a single and fixed (i.e., case-independent) decision threshold and states: “case is diseased” if the decision variable is greater than or equal to  $\zeta$ , i.e.,  $Z \geq \zeta$ , and “case is non-diseased” if the decision variable is smaller than  $\zeta$ , i.e.,  $Z < \zeta$ .

- The decision threshold is a fixed value used to separate cases reported as diseased from cases reported as non-diseased.
- Unlike the random Z-sample, which varies from case to case, the decision threshold is held fixed for the duration of the study. In some of the older literature<sup>2</sup> the decision threshold is sometimes referred to as “response bias”. The author hesitates to use the term “bias” which has a negative connotation, whereas, in fact, the choice of decision threshold depends on rational assessment of costs and benefits of different outcomes.
- The choice of decision threshold depends on the conditions of the study: perceived or known disease prevalence, cost-benefit considerations, instructions regarding dataset characteristics, personal interpreting style, etc. There is a transient “learning curve” during which observer is assumed to find the optimal threshold and henceforth holds it constant for the duration of the study. The learning is expected to stabilize during a sufficiently long training interval.
- Data should only be collected in the fixed threshold state, i.e., at the end of the training session.
- If a second study is conducted under different conditions, the observer will determine, after a new training session, the optimal threshold for the new conditions and henceforth hold it constant for the duration of the second study, etc.

From assumption #2, it follows that:

$$1 - Sp = FPF = P(Z \geq \zeta | T = 1) \quad (10.1)$$

$$Se = TPF = P(Z \geq \zeta | T = 2) \quad (10.2)$$

**Explanation:**  $P(Z \geq \zeta | T = 1)$  is the probability that the Z-sample for a non-diseased case is greater than or equal to  $\zeta$ . According to assumption #2 these cases are incorrectly classified as diseased, i.e., they are FP decisions and the corresponding probability is false positive fraction  $FPF$ , which is the complement of specificity  $Sp$ . Likewise,  $P(Z \geq \zeta | T = 2)$  denotes the probability that the Z-sample for a diseased case is greater than or equal to  $\zeta$ . These cases are correctly classified as diseased, i.e., these are TP decisions and the corresponding probability is true positive fraction  $TPF$ , which is sensitivity  $Se$ .

There are several concepts implicit in Eqn. (10.1) and Eqn. (10.2).

- The Z-samples have an associated probability distribution; this is implicit in the notation  $P(Z \geq \zeta | T = 2)$  and  $P(Z \geq \zeta | T = 1)$ . Diseased-cases are not homogenous; in some, disease is easy to detect, perhaps even obvious, in others the signs of disease are subtler, and in some, the disease is almost impossible to detect. Likewise, non-diseased cases are not homogenous.
- The probability distributions depend on the truth state  $T$ . The distribution of the Z-samples for non-diseased cases is in general different from that for the diseased cases. Generally, the distribution for  $T = 2$  is shifted to the right of that for  $T = 1$  (assuming a **positive-directed** decision variable scale). Later, specific distributional assumptions will be employed to obtain analytic expressions for the right hand sides of Eqn. (10.1) and Eqn. (10.2).
- The equations imply that via choice of the decision threshold  $\zeta$ ,  $Se$  and  $Sp$  are under the control of the observer. The lower the decision threshold the higher the sensitivity and the lower the specificity, and the converses are also true. Ideally both sensitivity and specificity should be large, i.e., unity (since they are probabilities they cannot exceed unity). The tradeoff between sensitivity and specificity says, essentially, that there is no “free lunch”. In general, the price paid for increased sensitivity is decreased specificity and vice-versa.

### 10.3.3 Adequacy of the training session

**Assumption 3:** The observer has complete knowledge of the distributions of actually non-diseased and actually diseased cases and makes rational decision based on this knowledge. Knowledge of the probabilistic distributions is consistent with not knowing for sure which distribution a specific sample came from, i.e., the “blindedness” assumption common to all observer performance studies.

How an observer can be induced to change the decision threshold is the subject of the following two examples.

## 10.4 Changing the decision threshold: Example I

Suppose that in the first study a radiologist interprets a set of cases subject to the instructions that it is rather important to identify actually diseased cases and not to worry about misdiagnosing actually non-diseased cases. One way to do this would be to reward the radiologist with \$10 for each TP decision but only \$1 for each TN decision. For simplicity, assume there is no penalty imposed for incorrect decisions (FPs and FNs) and the case set contains equal numbers of non-diseased and diseased cases, and the radiologist is informed of these facts. It is also assumed that the radiologist is allowed to reach a steady state and responds rationally to the payoff arrangement. Under these circumstances, the radiologist is expected to set the decision threshold at a small value so that even slight evidence of presence of disease is enough to result in a “case is diseased” decision. The low decision threshold also implies that considerable evidence of lack of disease is needed before a “case is non-diseased” decision is rendered. The radiologist is expected to achieve relatively high sensitivity but specificity will be low. As a concrete example, if there are 100 non-diseased cases and 100 diseased cases, assume the radiologist makes 90 TP decisions; since the threshold for presence of presence of disease is small, this number is close to the maximum possible value, namely 100. Assume further that 10 TN decisions are made; since the implied threshold for evidence of absence of disease is large, this number is close to the minimum possible value, namely 0. Therefore, sensitivity is 90percent and specificity is 10percent. The radiologist earns  $90 \times \$10 + 10 \times \$1 = \$910$  for participating in this study.

Next, suppose the study is repeated with the same cases but this time the payoff is \$1 for each TP decision and \$10 for each TN decision. Suppose, further, that sufficient time has elapsed between the two study sessions that memory effects can be neglected. Now the roles of sensitivity and specificity are reversed. The radiologist’s incentive is to be correct on actually non-diseased cases without worrying too much about missing actually diseased cases. The radiologist is expected to set the decision threshold at a large value so that considerable evidence of disease-presence is required to result in a “case is diseased” decision, but even slight evidence of absence of disease is enough to result in a “case is non-diseased” decision. This radiologist is expected to achieve relatively low sensitivity but specificity will be higher. Assume the radiologist makes 90 TN decisions and 10 TP decisions, earning \$910 for the second study. The corresponding sensitivity is 10percent and specificity is 90percent.

The incentives in the first study caused the radiologist to accept low specificity in order to achieve high sensitivity; the incentives in the second study caused the radiologist to accept low sensitivity in order to achieve high specificity.

## 10.5 Changing the decision threshold: Example II

Suppose one asks the same radiologist to interpret a set of cases, but this time the reward for a correct decision is always \$1, regardless of the truth state of the case, and as before, there is no penalty for incorrect decisions. However, the radiologist is told that disease prevalence is only 0.005 and that this is the actual prevalence, i.e., the experimenter is not deceiving the radiologist in this regard. [Even if the experimenter attempts to deceive the radiologist, by claiming for example that there are roughly equal numbers of non-diseased and diseased cases, after interpreting a few tens of cases the radiologist will know that a deception is involved. Deception in such studies is generally not a good idea, as the observer's performance is not being measured in a "steady state condition" – the observer's performance will change as the observer "learns" the true disease prevalence.] In other words, only five out of every 1000 cases are actually diseased. This information will cause the radiologist to adopt a high threshold for diagnosing disease-present thereby becoming more reluctant to state: "case is diseased". By simply diagnosing all cases as non-diseased, without using any case information, the radiologist will be correct on every disease absent case and earn \$995, which is very close to the maximum \$1000 the radiologist can earn by using case information to the full and being correct on disease-present and disease-absent cases.

The example is not as contrived as might appear at first sight. However, in screening mammography, the cost of missing a breast cancer, both in terms of loss of life and a possible malpractice suite, is usually perceived to be higher than the cost of a false positive. This can result in a shift towards higher sensitivity at the expense of lower specificity.

If a new study were conducted with a highly enriched set of cases, where the disease prevalence is 0.995 (i.e., only 5 out of every 1000 cases are actually non-diseased), then the radiologist would adopt a low threshold. By simply calling every case "non-diseased", the radiologist earns \$995.

These examples show that by manipulating the relative costs of correct vs. incorrect decisions and / or by varying disease prevalence one can influence the radiologist's decision threshold. These examples apply to laboratory studies. Clinical interpretations are subject to different cost-benefit considerations that are generally not under the researcher's control: actual (population) disease prevalence, the reputation of the radiologist, malpractice, etc.

## 10.6 The equal-variance binormal model

Here is the model for the Z-samples. Using the notation  $N(\mu, \sigma^2)$  for the normal (or "Gaussian") distribution with mean  $\mu$  and variance  $\sigma^2$ , it is assumed: 1. The

Z-samples for non-diseased cases are distributed  $N(0, 1)$ . 2. The Z-samples for diseased cases are distributed  $N(\mu, 1)$  with  $\mu > 0$ . 3. A case is diagnosed as diseased if its Z-sample  $\geq$  a constant threshold  $\zeta$ , and non-diseased otherwise.

The constraint  $\mu > 0$  is needed so that the observer's performance is at least as good as chance. A large negative value for this parameter would imply an observer so predictably bad that the observer is good; one simply reverses the observer's decision ("diseased" to "non-diseased" and vice versa) to get near-perfect performance .

The model described above is termed the equal-variance binormal model. [If the common variance is not unity, one can re-scale the decision axis to achieve unit-variance without changing the predictions of the model.] A more general model termed the unequal-variance binormal model is generally used for modeling human observer data, discussed later, but for the moment, one does not need that complication. The equal-variance binormal model is defined by:

$$\left. \begin{array}{l} Z_{k_t t} \sim N(\mu_t, 1) \\ \mu_1 = 0 \\ \mu_2 = \mu \end{array} \right\} \quad (10.3)$$

In Eqn. (10.3) the subscript  $t$  denotes the truth, sometimes referred to as the "gold standard", with  $t = 1$  denoting a non-diseased case and  $t = 2$  denoting a diseased case. The variable  $Z_{k_t t}$  denotes the random Z-sample for case  $k_t t$ , where  $k_t$  is the index for cases with truth state  $t$ ; for example  $k_1 1 = 21$  denotes the 21st non-diseased case and  $k_2 2 = 3$  denotes the 3rd diseased case. To explicate  $k_1 1 = 21$  further, the label  $k_1$  indexes the case while the label 1 indicates the truth of the case. The label  $k_t$  ranges from  $1, 2, \dots, K_t$ , where  $K_t$  is the total number of cases with disease state  $t$ .

The author departs from usual convention, see for example paper by Hillis, which labels the cases with a single index  $k$ , which ranges from 1 to  $K_1 + K_2$ , and one is left guessing as to the truth-state of each case. Also, the proposed notation extends readily to the FROC paradigm where two states of truth have to be distinguished, one at the case level and one at the location level.

The first line in Eqn. (10.3) states that  $Z_{k_t t}$  is a random sample from the  $N(\mu_t, 1)$  distribution, which has unit variance regardless of the value of  $t$  (this is the reason for naming it the equal-variance binormal model). The remaining lines in Eqn. (10.3) defines  $\mu_1$  as zero and  $\mu_2$  as  $\mu$ . Taken together, these equations state that non-diseased case Z-samples are distributed  $N(0, 1)$  and diseased case Z-samples are distributed  $N(\mu, 1)$ . The name binormal arises from the two normal distributions underlying this model. It should not be confused with bivariate, which identifies a single distribution yielding two values per sample, where the two values could be correlated. In the binormal model, the samples from the two distributions are assumed independent of each other.

A few facts concerning the normal (or Gaussian) distribution are summarized next.

## 10.7 The normal distribution

In probability theory, a probability density function (pdf), or density of a continuous random variable, is a function giving the relative chance that the random variable takes on a given value. For a continuous distribution, the probability of the random variable being exactly equal to a given value is zero. The probability of the random variable falling in a range of values is given by the integral of this variable's pdf function over that range. For the normal distribution  $N(\mu, \sigma^2)$  the pdf is denoted  $\phi(z|\mu, \sigma)$ .

By definition,

$$\phi(z|\mu, \sigma) = P(z < Z < z + dz | Z \sim N(\mu, \sigma^2)) \quad (10.4)$$

The right hand side of Eqn. (10.4) is the probability that the random variable  $Z$ , sampled from  $N(\mu, \sigma^2)$ , is between the fixed limits  $z$  and  $z + dz$ . For this reason  $\phi(z|\mu, \sigma)$  is termed the probability density function. The special case  $\phi(z|0, 1)$  is referred to as the **unit normal distribution**; it has zero mean and unit variance and the corresponding pdf is denoted  $\phi(z)$ . The defining equation for the pdf of this distribution is:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad (10.5)$$

The integral of  $\phi(t)$  from  $-\infty$  to  $z$ , as in Eqn. (10.6), is the probability that a sample from the unit normal distribution is less than  $z$ . Regarded as a function of  $z$ , this is termed the cumulative distribution function (CDF) and is denoted, in this book, by the symbol  $\Phi$  (sometimes the term probability distribution function is used for what we are terming the CDF). The function  $\Phi(z)$ , specific to the unit normal distribution, is defined by:

$$\Phi(z) = \int_{-\infty}^z \phi(t) dt \quad (10.6)$$

Fig. 10.1 shows plots, as functions of  $z$ , of the CDF and the pdf for the unit normal distribution. Since  $z$ -samples outside  $\pm 3$  are unlikely, the plotted range, from  $-3$  to  $+3$  includes most of the distribution. The pdf is the familiar bell-shaped curve, centered at zero; the corresponding R function is `dnorm()`, i.e., density of the normal distribution. The CDF  $\Phi(z)$  increases monotonically from 0 to unity as  $z$  increases from  $-\infty$  to  $+\infty$ . It is the sigmoid (S-shaped) shaped curve in Fig. 10.1; the corresponding R function is `pnorm()`.

The sigmoid shaped curve is the CDF, or cumulative distribution function, of the  $N(0,1)$  distribution, while the bell-shaped curve is the corresponding pdf, or probability density function. The dashed line corresponds to the reporting threshold  $\zeta$ . The area under the pdf to the left of  $\zeta$  equals the value of CDF at the selected  $\zeta$ , i.e.,  $0.841$  (`pnorm(1) = 0.841`).

```
x <- seq(-3,3,0.01)
pdfData <- data.frame(z = x, pdfcdf = dnorm(x))
cdfData <- data.frame(z = x, pdfcdf = pnorm(x))
pdfcdfPlot <- ggplot(
  mapping = aes(x = z, y = pdfcdf)) +
  geom_line(data = pdfData) +
  geom_line(data = cdfData) +
  geom_vline(xintercept = 1, linetype = 2) +
  xlab(label = "z") + ylab(label = "pdf/CDF")
print(pdfcdfPlot)
```

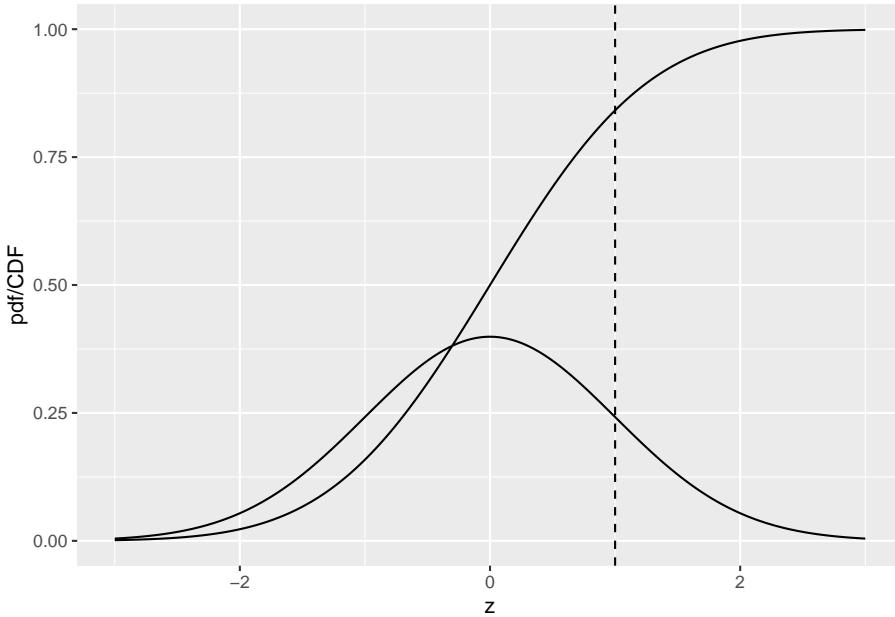


Figure 10.1: pdf-CDF plots for unit normal.

A related function is the inverse of Eqn. (10.6). Suppose the left hand side of Eqn. (10.6) is denoted  $p$ , which is a probability in the range 0 to 1.

$$p = \Phi(z) = \int_{-\infty}^z \phi(t)dt \quad (10.7)$$

The inverse of  $\Phi(z)$  is that function which when applied to  $p$  yields the upper limit  $z$  in Eqn. (10.6), i.e.,

$$\Phi^{-1}(p) = z \quad (10.8)$$

Since  $p = \Phi(z)$  it follows that

$$\Phi(\Phi^{-1}(z)) = z \quad (10.9)$$

This nicely satisfies the property of an inverse function. The inverse function is known in statistical terminology as the quantile function, implemented in R as the `qnorm()` function. Think of `pnorm()` as a probability and `qnorm()` as value on the z-axis.

To summarize, `norm` implies the unit normal distribution, `p` denotes a probability distribution function or CDF, `q` denotes a quantile function and `d` denotes a density function; this convention is used with all distributions in R.

```
qnorm(0.025)
#> [1] -1.959964
qnorm(1-0.025)
#> [1] 1.959964
pnorm(qnorm(0.025))
#> [1] 0.025
qnorm(pnorm(-1.96))
#> [1] -1.96
```

The first command `qnorm(0.025)` demonstrates the identity:

$$\Phi^{-1}(0.025) = -1.959964 \quad (10.10)$$

The next command `qnorm(1-0.025)` demonstrates the identity:

$$\Phi^{-1}(1 - 0.025) = +1.959964 \quad (10.11)$$

The last two commands demonstrate that `pnorm` and `qnorm`, applied in either order, are inverses of each other.

Eqn. (10.10) means that the (rounded) value -1.96 is such that the area under the pdf to the left of this value is 0.025. Similarly, Eqn. (10.11) means that the (rounded) value +1.96 is such that the area under the pdf to the left of

this value is  $1 - 0.025 = 0.975$ . In other words,  $-1.96$  captures, to its left, the 2.5th percentile of the unit-normal distribution, and  $1.96$  captures, to its left, the 97.5th percentile of the unit-normal distribution, Fig. 10.2. Since between them they capture 95percent of the unit-normal pdf, these two values can be used to estimate 95percent confidence intervals.

```

mu <- 0; sigma <- 1
zeta <- -qnorm(0.025)
step <- 0.1

LL<- -3
UL <- mu + 3*sigma

x.values <- seq(zeta,UL,step)
cord.x <- c(zeta, x.values,UL)
cord.y <- c(0,dnorm(x.values),0)

z <- seq(LL, UL, by = step)
curveData <- data.frame(z = z, pdfs = dnorm(z))
shadeData <- data.frame(z = cord.x, pdfs = cord.y)
shadedTails <- ggplot(mapping = aes(x = z, y = pdfs)) +
  geom_polygon(data = shadeData, color = "grey", fill = "grey")

zeta <- qnorm(0.025)
x.values <- seq(LL, zeta,step)
cord.x <- c(LL, x.values,zeta)
cord.y <- c(0,dnorm(x.values),0)
shadeData <- data.frame(z = cord.x, pdfs = cord.y)
shadedTails <- shadedTails +
  geom_polygon(
    data = shadeData, color = "grey", fill = "grey") +
  xlab(label = "z")
shadedTails <- shadedTails +
  geom_line(data = curveData, color = "black")
print(shadedTails)

```

If one knows that a variable is distributed as a unit-normal random variable, then the observed value minus 1.96 defines the lower limit of its 95percent confidence interval, and the observed value plus 1.96 defines the upper limit of its 95percent confidence interval.

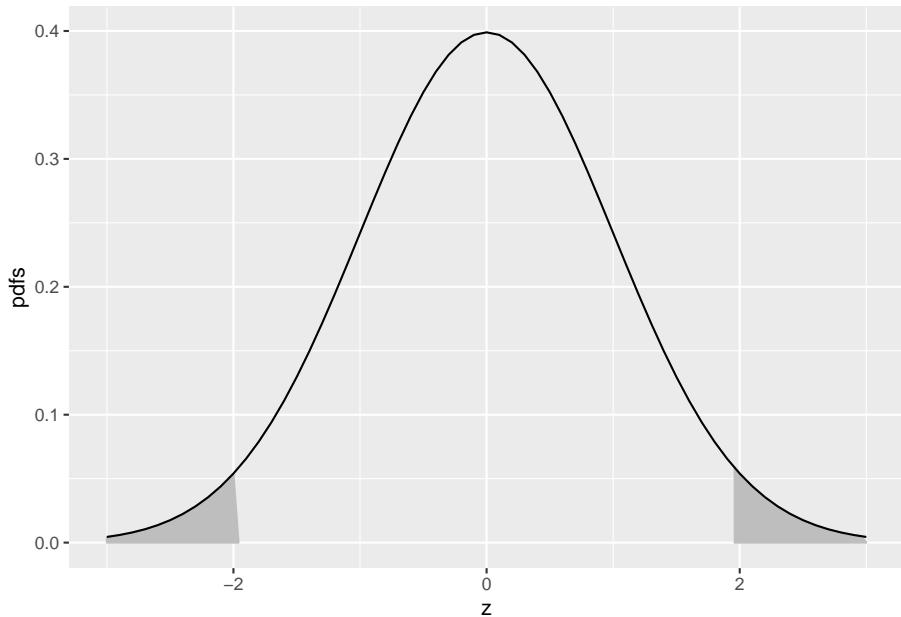


Figure 10.2: Illustrating that 95percent of the total area under the unit normal pdf is contained in the range  $|Z| < 1.96$ , which can be used to construct a 95percent confidence interval for an estimate of a suitably normalized statistic. The area contained in each shaded tail is 2.5percent.

## 10.8 Analytic expressions for specificity and sensitivity

Specificity corresponding to threshold  $\zeta$  is the probability that a Z-sample from a non-diseased case is smaller than  $\zeta$ . By definition, this is the CDF corresponding to the threshold  $\zeta$ . In other words:

$$Sp(\zeta) = P(Z_{k_11} < \zeta | Z_{k_11} \sim N(0, 1)) = \Phi(\zeta) \quad (10.12)$$

The expression for sensitivity can be derived tediously by starting with the fact that  $Z_{k_22}$  and then using calculus to obtain the probability that a z-sample for a disease-present case exceeds  $\zeta$ . A quicker way is to consider the random variable obtaining by shifting the origin to  $\mu$ . A little thought should convince the reader that  $Z_{k_22} - \mu$  must be distributed as  $N(0, 1)$ . Therefore, the desired probability is (the last step follows from the identity in Eqn. (3.7), with z replaced by  $\zeta - \mu$  :

$$\begin{aligned} Se(\zeta) \\ &= P(Z_{k_22} \geq \zeta) \\ &= P((Z_{k_22} - \mu) \geq (\zeta - \mu)) \\ &= 1 - P((Z_{k_22} - \mu) < (\zeta - \mu)) \\ &= 1 - \Phi(\zeta - \mu) \end{aligned} \quad (10.13)$$

A little thought (based on the definition of the CDF function and the symmetry of the unit-normal pdf function) should convince the reader that:

$$1 - \Phi(\zeta) = -\Phi(\zeta)1 - \Phi(\zeta - \mu) = \Phi(\mu - \zeta) \quad (10.14)$$

Instead of carrying the “1 minus” around, one can use the more compact notation. Summarizing, the analytical formulae for the specificity and sensitivity for the equal-variance binormal model are:

$$Sp(\zeta) = \Phi(\zeta)Se(\zeta) = \Phi(\mu - \zeta) \quad (10.15)$$

In these equations, the threshold  $\zeta$  appears with different signs because specificity is the area under a pdf to the **left** of a threshold, while sensitivity is the area to the **right**.

**As probabilities, both sensitivity and specificity are restricted to the range 0 to 1.** The observer’s performance could be characterized by specifying sensitivity and specificity, i.e., a pair of numbers. If both sensitivity and specificity of an imaging system are greater than

the corresponding values for another system, then the 1st system is unambiguously better than the 2nd. But what if sensitivity is greater for the 1st but specificity is greater for the 2nd? Now the comparison is ambiguous. It is difficult to unambiguously compare two pairs of performance indices. Clearly, a scalar measure is desirable that combines sensitivity and specificity into a single measure of diagnostic performance.

The parameter  $\mu$  satisfies the requirements of a scalar figure of merit (FOM). Eqn. (10.15) can be solved for  $\mu$  as follows. Inverting the equations yields:

$$\zeta = \Phi^{-1}(Sp(\zeta)) \quad \mu - \zeta = \Phi^{-1}(Se(\zeta)) \quad (10.16)$$

Eliminating  $\zeta$  yields:

$$\mu = \Phi^{-1}(Sp(\zeta)) + \Phi^{-1}(Se(\zeta)) \quad (10.17)$$

This is a useful relation, as it converts a *pair* of numbers that is hard to compare between two modalities, in the sense described above, into a *single* FOM. Now it is almost trivial to compare two modalities: the one with the higher  $\mu$  wins. In reality, the comparison is not trivial since like sensitivity and specificity,  $\mu$  has to be estimated from a finite dataset and is therefore subject to sampling variability.

```
options(digits=3)
mu <- 3; sigma <- 1
zeta <- 1
step <- 0.1

lowerLimit<- -1 # lower limit
upperLimit <- mu + 3*sigma # upper limit

z <- seq(lowerLimit, upperLimit, by = step)
pdfs <- dnorm(z)
seqNor <- seq(zeta,upperLimit,step)
cord.x <- c(zeta, seqNor,upperLimit)
# need two y-coords at each end point of range;
# one at zero and one at value of function
cord.y <- c(0,dnorm(seqNor),0)
curveData <- data.frame(z = z, pdfs = pdfs)
shadeData <- data.frame(z = cord.x, pdfs = cord.y)
shadedPlots <- ggplot(mapping = aes(x = z, y = pdfs)) +
  geom_line(data = curveData, color = "blue") +
  geom_polygon(data = shadeData, color = "blue", fill = "blue")
```

```

crossing <- uniroot(function(x) dnorm(x) - dnorm(x, mu, sigma),
                      lower = 0, upper = 3)$root
crossing <- max(c(zeta, crossing))
seqAbn <- seq(crossing, upperLimit, step)
cord.x <- c(seqAbn, rev(seqAbn))
# reason for reverse
# we want to explicitly define the polygon
# we dont want R to close it

cord.y <- c()
for (i in seq(1, length(cord.x)/2)) {
  cord.y <- c(cord.y, dnorm(cord.x[i], mu, sigma))
}
for (i in seq(1, length(cord.x)/2)) {
  cord.y <- c(cord.y, dnorm(cord.x[length(cord.x)/2+i]))
}
pdःfs <- dnorm(z, mu, sigma)
curveData <- data.frame(z = z, pdःfs = pdःfs)
shadeData <- data.frame(z = cord.x, pdःfs = cord.y)
shadedPlots <- shadedPlots +
  geom_line(data = curveData, color = "red") +
  geom_polygon(data = shadeData, color = "red", fill = "red")
seqAbn <- seq(zeta, upperLimit, step)
for (i in seqAbn) {
  # define xs and ys of two points, separated only along y-axis
  vlineData <- data.frame(x1 = i,
                           x2 = i,
                           y1 = 0,
                           y2 = dnorm(i, mu, sigma))
  # draw vertical line between them
  shadedPlots <- shadedPlots +
    geom_segment(aes(x = x1, xend = x2, y = y1, yend = y2),
                 data = vlineData, color = "red")
}
shadedPlots <- shadedPlots + xlab(label = "z-sample")
print(shadedPlots)

```

Fig. 10.3 shows the equal-variance binormal model for  $\mu = 3$  and  $\zeta = 1$ . The blue-shaded area, including the “common” portion with the vertical red lines, is the probability that a z-sample from a non-diseased case exceeds  $\zeta = 1$ , which is the complement of specificity, i.e., it is false positive fraction, which is  $1 - \text{pnorm}(1) = 0.159$ . The red shaded area, including the “common” portion with the vertical red lines, is the probability that a z-sample from a diseased case exceeds  $\zeta = 1$ , which is sensitivity or true positive fraction, which is  $\text{pnorm}(3-1) = 0.977$ .

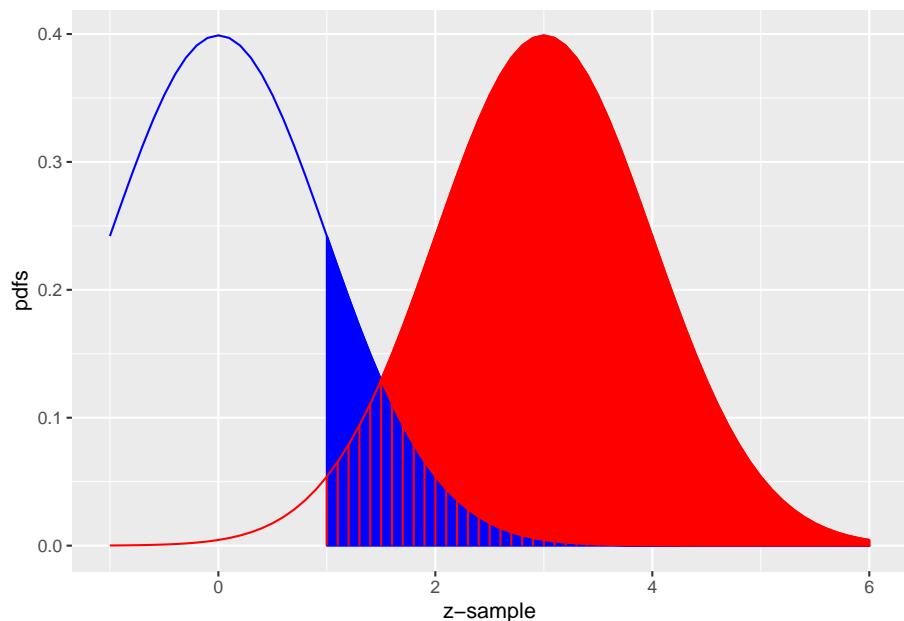


Figure 10.3: The equal-variance binormal model for  $\mu = 3$  and  $\zeta = 1$ ; the blue curve, centered at zero, corresponds to the pdf of non-diseased cases and the red one, centered at  $\mu = 3$ , corresponds to the pdf of diseased cases. The left edge of the blue shaded region represents the threshold  $\zeta$ , currently set at unity. The red shaded area, including the common portion with the vertical red lines, is sensitivity. The blue shaded area including the common portion with the vertical red lines is 1-specificity.

Demonstrated next are these concepts using R examples.

## 10.9 Demonstration of the concepts of sensitivity and specificity

### 10.9.1 Estimating mu from a finite sample

The following code simulates 9 non-diseased and 11 diseased cases. The  $\mu$  parameter is 1.5 and  $\zeta$  is  $\mu/2$ . Shown are the calculations of sensitivity and specificity and the value of estimated  $\mu$ .

```
mu <- 1.5
zeta <- mu/2
seed <- 100 # line 4
K1 <- 9
K2 <- 11
ds <- simulateDataset(K1, K2, mu, zeta, seed)

cat("seed = ", seed,
    "\nK1 = ", K1,
    "\nK2 = ", K2,
    "\nSpecificity = ", ds$Sp,
    "\nSensitivity = ", ds$Se,
    "\nEst. of mu = ", ds$mu, "\n")
#> seed = 100
#> K1 = 9
#> K2 = 11
#> Specificity = 0.889
#> Sensitivity = 0.909
#> Est. of mu = 2.56
```

Since this is a finite sample, the estimate of  $\mu$  is not exactly equal to the true value. In fact, all of the estimates, sensitivity, specificity and  $\mu$  are subject to sampling variability.

### 10.9.2 Changing the seed variable: case-sampling variability

No matter how many times one runs the above code, one always sees the same output shown above. This is because at line 4 one sets the `seed` of the random number generator to a fixed value, namely 100. This is like having a perfectly reproducible reader repeatedly interpreting the same cases – one always gets the same results. Change the `seed` to 101. One should see:

```

seed <- 101 # change
ds <- simulateDataset(K1, K2, mu, zeta, seed)

cat("seed = ", seed,
    "\nK1 = ", K1,
    "\nK2 = ", K2,
    "\nSpecificity = ", ds$Sp,
    "\nSensitivity = ", ds$Se,
    "\nEst. of mu = ", ds$mu, "\n")
#> seed = 101
#> K1 = 9
#> K2 = 11
#> Specificity = 0.778
#> Sensitivity = 0.545
#> Est. of mu = 0.879

```

Changing `seed` is equivalent to sampling a completely new set of patients. This is an example of case sampling variability. The effect is quite large (`Se` fell from 0.909 to 0.545 and estimated `mu` fell from 2.56 to 0.879!) because the size of the relevant case set,  $K_2 = 11$  for sensitivity, is rather small, leading to large variability.

### 10.9.3 Increasing the numbers of cases

Here we increase  $K_1$  and  $K_2$ , by a factor of 10 each, and return the `seed` to 100.

```

K1 <- 90 # change
K2 <- 110 # change
seed <- 100 # change
ds <- simulateDataset(K1, K2, mu, zeta, seed)

cat("seed = ", seed,
    "\nK1 = ", K1,
    "\nK2 = ", K2,
    "\nSpecificity = ", ds$Sp,
    "\nSensitivity = ", ds$Se,
    "\nEst. of mu = ", ds$mu, "\n")
#> seed = 100
#> K1 = 90
#> K2 = 110
#> Specificity = 0.778
#> Sensitivity = 0.836
#> Est. of mu = 1.74

```

Next we change `seed` to 101.

```
seed <- 101 # change
ds <- simulateDataset(K1, K2, mu, zeta, seed)

cat("seed = ", seed,
    "\nK1 = ", K1,
    "\nK2 = ", K2,
    "\nSpecificity = ", ds$Sp,
    "\nSensitivity = ", ds$Se,
    "\nEst. of mu = ", ds$mu, "\n")
#> seed = 101
#> K1 = 90
#> K2 = 110
#> Specificity = 0.811
#> Sensitivity = 0.755
#> Est. of mu = 1.57
```

Notice that now the values are less sensitive to `seed`. Table 10.1 illustrates this trend with ever increasing sample sizes (the reader should confirm the listed values).

```
results <- array(dim = c(9,6))
mu <- 1.5
zeta <- mu/2
results[9,] <- c(Inf, Inf, NA, pnorm(zeta), pnorm(mu-zeta), mu)
K1_arr <- c(9, 9, 90, 90, 900, 900, 9000, 9000, NA)
K2_arr <- c(11, 11, 110, 110, 1100, 1100, 11000, 11000, NA)
seed_arr <- c(100,101,100,101,100,101,100,101,NA)
for (i in 1:8) {
  ds <- simulateDataset(K1_arr[i], K2_arr[i], mu, zeta, seed_arr[i])
  results[i,] <- c(K1_arr[i], K2_arr[i], seed_arr[i], ds$Sp, ds$Se, ds$mu)
}
df <- as.data.frame(results)
colnames(df) <- c("K1", "K2", "seed", "Se", "Sp", "mu")
```

As the numbers of cases increase, the sensitivity and specificity converge to a common value, around 0.773 and the estimate of the separation parameter converges to the known value.

```
pnorm(0.75) # example 1
#> [1] 0.773
2*qnorm(pnorm(zeta)) # example 2
#> [1] 1.5
```

Table 10.1: Effect of sample size and seed on estimates of sensitivity, specificity and the mu-parameter.

K1	K2	seed	Se	Sp	mu
9	11	100	0.889	0.909	2.556
9	11	101	0.778	0.545	0.879
90	110	100	0.778	0.836	1.744
90	110	101	0.811	0.755	1.571
900	1100	100	0.764	0.761	1.430
900	1100	101	0.807	0.759	1.569
9000	11000	100	0.774	0.772	1.496
9000	11000	101	0.771	0.775	1.498
Inf	Inf	NA	0.773	0.773	1.500

Because the threshold is halfway between the two distributions, as in this example, sensitivity and specificity are identical. In words, with two unit variance distributions separated by 1.5, the area under the diseased distribution (centered at 1.5) above 0.75, namely sensitivity, equals the area under the non-diseased distribution (centered at zero) below 0.75, namely specificity, and the common value is  $\Phi(0.75) = 0.773$ , yielding the last row of Table 10.1, and example 1 in the above code snippet. Example 2 in the above code snippet illustrates Eqn. (10.17). The factor of two arises since in this example sensitivity and specificity are identical.

From Table 10.1, for the same numbers of cases but different seeds, comparing pairs of sensitivity and specificity values is more difficult as two pairs of numbers (i.e., four numbers) are involved. Comparing a single pair of  $\mu$  values is easier as only two numbers are involved. The tendency of the pairs to become independent of case sample is discernible with fewer cases with  $\mu$ , around 90/110 cases, than with sensitivity and specificity pairs. The numbers in the table might appear disheartening in terms of the implied numbers of cases needed to detect a difference in specificity. Even with 200 cases, the difference in specificity for two seed values is 0.081, which is actually a large effect considering that the scale extends from 0 to 1.0. A similar comment applies to differences in sensitivity. The situation is not quite that bad. One uses an area measure that combines sensitivity and specificity yielding less variability in the combined measure. One uses the ratings paradigm, which is more efficient than the binary one used in this chapter. Finally, one takes advantage of correlations that exist between the interpretations in matched-case matched-reader interpretations in two modalities that tend to decrease variability in the AUC-difference even further (most applications of ROC methods involved detecting differences in AUCs not absolute values).

## 10.10 Inverse variation of sensitivity and specificity and the need for a single FOM

The variation of sensitivity and specificity is modeled in the binormal model by the threshold parameter  $\zeta$ . From Eqn. (10.12), specificity at threshold  $\zeta$  is  $\Phi(\zeta)$  and the corresponding expression for sensitivity is  $\Phi(\mu - \zeta)$ . Since the threshold  $\zeta$  appears with a minus sign, the dependence of sensitivity on  $\zeta$  will be the opposite of the corresponding dependence of specificity on  $\zeta$ . In Fig. 10.3, the left edge of the blue shaded region represents the threshold  $\zeta = 1$ . As  $\zeta = 1$  is moved towards the left, specificity decreases but sensitivity increases. Specificity decreases because less of the non-diseased distribution lies to the left of the new threshold, in other words fewer non-diseased cases are correctly diagnosed as non-diseased. Sensitivity increases because more of the diseased distribution lies to the right of the new threshold, in other words more diseased cases are correctly diagnosed as diseased. If an observer has higher sensitivity than another observer, but lower specificity, it is difficult to unambiguously compare them. It is not impossible (Skaane et al., 2013). The unambiguous comparison is difficult for the following reason. Assuming the second observer can be coaxed into adopting a lower threshold, thereby decreasing specificity to match that of the first observer, then it is possible that the second observer's sensitivity, formerly smaller, could now be greater than that of the first observer. A single figure of merit is desirable to the sensitivity - specificity analysis. It is possible to leverage the inverse variation of sensitivity and specificity by combining them into a single scalar measure, as was done with the  $\mu$  parameter in the previous section, Eqn. (10.17). An equivalent way is by using the area under the ROC plot, discussed next.

## 10.11 The ROC curve

The receiver operating characteristic (ROC) is defined as the plot of sensitivity (y-axis) vs. 1-specificity (x-axis). Equivalently, it is the plot of TPF (y-axis) vs. FPF (x-axis). From Eqn. (10.15) it follows that:

$$\begin{aligned} FPF(\zeta) &= 1 - Sp(\zeta) \\ &= \Phi(-\zeta) \\ TPF(\zeta) &= Se(\zeta) \\ &= \Phi(\mu - \zeta) \end{aligned} \tag{10.18}$$

Specifying  $\zeta$  selects a particular operating point on this plot and varying  $\zeta$  from  $+\infty$  to  $-\infty$  causes the operating point to trace out the ROC curve from the origin  $(0,0)$  to  $(1,1)$ . Specifically, as  $\zeta$  is decreased from  $+\infty$  to  $-\infty$ , the

operating point rises from the origin (0,0) to the end-point (1,1). In general, as  $\zeta$  increases, the operating point moves down the curve, and conversely, as  $\zeta$  decreases the operating point moves up the curve. The operating point  $O(\zeta|\mu)$  for the equal variance binormal model is (the notation assumes the  $\mu$  parameter is fixed and  $\zeta$  is varied by the observer in response to interpretation conditions):

$$O(\zeta | \mu) = (\Phi(-\zeta), \Phi(\mu - \zeta)) \quad (10.19)$$

The operating point predicted by the above equation lies exactly on the theoretical ROC curve. This condition can only be achieved with very large numbers of cases, so that sampling variability is very small. In practice, with finite datasets, the operating point will almost never be exactly on the theoretical curve.

The ROC curve is the locus of the operating point for fixed  $\mu$  and variable  $\zeta$ . Fig. 10.4 shows examples of equal-variance binormal model ROC curves for different values of  $\mu$ . Each curve is labeled with the corresponding value of  $\mu$ . Each has the property that TPF is a monotonically increasing function of FPF and the slope decreases monotonically as the operating point moves up the curve. As  $\mu$  increases the curves get progressively upward-left shifted, approaching the top-left corner of the ROC plot. In the limit  $\mu = \infty$  the curve degenerates into two line segments, a vertical one connecting the origin to (0,1) and a horizontal one connecting (0,1) to (1,1) – the ROC plot for a perfect observer.

```
mu <- 0;zeta <- seq(-5, mu + 5, 0.05)
FPF <- pnorm(-zeta)
rocPlot <- ggplot(mapping = aes(x = FPF, y = TPF))
for (mu in 0:3){
  TPF <- pnorm(mu-zeta)
  curveData <- data.frame(FPF = FPF, TPF = TPF)
  rocPlot <- rocPlot +
    geom_line(data = curveData, size = 2) +
    xlab("FPF") + ylab("TPF" ) +
    theme(axis.title.y = element_text(size = 25,face="bold"),
          axis.title.x = element_text(size = 30,face="bold")) +
    annotate("text",
             x = pnorm(-mu/2) + 0.07,
             y = pnorm(mu/2),
             label = paste0("mu == ", mu),
             parse = TRUE, size = 8)
  next
}
rocPlot <- rocPlot +
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0))
```

```

rocPlot <- rocPlot +
  geom_abline(slope = -1,
              intercept = 1,
              linetype = 3,
              size = 2)
print(rocPlot)

```

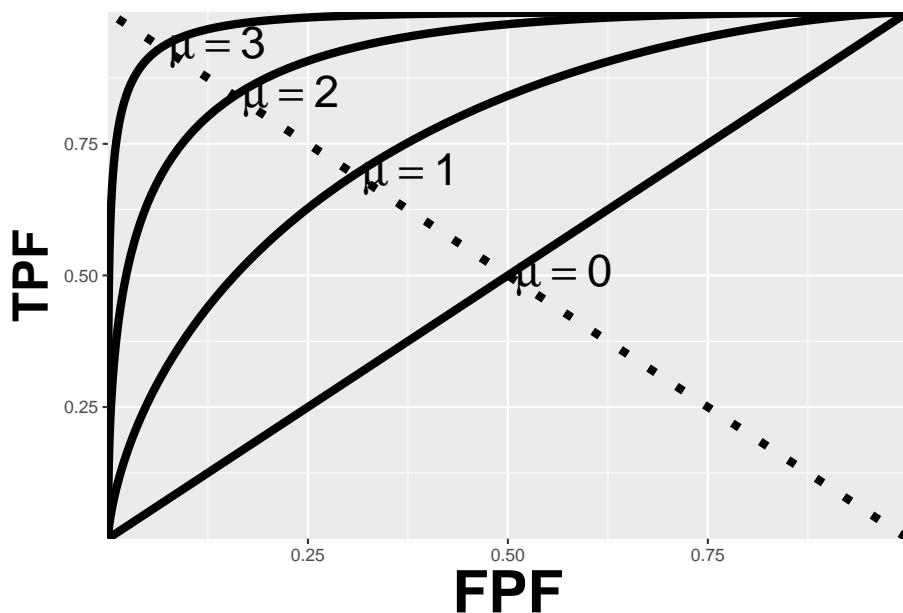


Figure 10.4: ROC plots predicted by the equal variance binormal model for different values of  $\mu$ . As  $\mu$  increases the intersection of the curve with the negative diagonal moves closer to the ideal operating point,  $(0,1)$  at which sensitivity and specificity are both equal to unity.

### 10.11.1 The chance diagonal

In Fig. 10.4 the ROC curve for  $\mu = 0$  is the positive diagonal of the ROC plot, termed the chance diagonal. Along this curve  $TPF = FPF$  and the observer's performance is at chance level. In the equal variance binormal model, for  $\mu = 0$ , the pdf of the diseased distribution is identical to that of the non-diseased distribution: both are centered at the origin. Therefore, no matter the choice of threshold  $\zeta$ ,  $TPF = FPF$ . Setting  $\mu = 0$  in Eqn. (10.18) yields:

$$TPF(\zeta) = FPF(\zeta) = \Phi(-\zeta)$$

In this special case, the red and blue curves in Fig. 10.3 coincide. The observer is unable to find any difference between the two distributions. This can happen if the cancers are of such low visibility so that diseased cases are indistinguishable from non-diseased ones, or the observer's skill level is so poor that the observer is unable to make use of distinguishing characteristics between diseased and non-diseased cases that do exist, and which experts exploit.

### 10.11.2 The guessing observer

If the cases are indeed impossibly difficult and/or the observer has zero skill at discriminating between them, the observer has no option but to guess. This rarely happens in the clinic, as too much is at stake and this paragraph is intended to make a pedagogical point that the observer can move the operating point along the chance diagonal. If there is no special incentive, the observer tosses a coin and if the coin lands head up, the observer states: "case is diseased" and otherwise states: "case is non-diseased". When this procedure is averaged over many non-diseased and diseased cases, it will result in the operating point (0.5, 0.5). [Many cases are assumed as otherwise, due to sampling variability, the operating point will not be on the theoretical ROC curve.] To move the operating point downward, e.g., to (0.1, 0.1) the observer randomly selects an integer number between 1 and 10, equivalent to a 10-sided "coin". Whenever a one "shows up", the observer states "case is diseased" and otherwise the observer states "case is non-diseased". To move the operating point to (0.2, 0.2) whenever a one or two "shows up", the observer states "case is diseased" and otherwise the observer states "case is non-diseased". One can appreciate that simply by changing the probability of stating "case is diseased" the observer can place the operating point anywhere on the chance diagonal, but wherever the operating point is placed, it will satisfy  $TPF = FPF$ .

### 10.11.3 Symmetry with respect to negative diagonal

A characteristic of the ROC curves shown in Fig. 10.4 is that they are symmetric with respect to the negative diagonal, defined as the straight line joining (0,1) and (1,0) which is shown as the dotted straight line in Fig. 10.4. The symmetry property is due to the equal variance nature of the binormal model and is not true for models considered in later chapters. The intersection between the ROC curve and the negative diagonal corresponds to  $\zeta = \mu/2$ , in which case the operating point is:

$$\begin{aligned} FPF(\zeta) &= \Phi(-\mu/2) \\ TPF(\zeta) &= \Phi(\mu/2) \end{aligned} \tag{10.20}$$

The first equation implies:

$$1 - FPF(\zeta) = 1 - \Phi(-\mu/2) = \Phi(\mu/2)$$

Therefore,

$$TPF(\zeta) = 1 - FPF(\zeta) \tag{10.21}$$

This equation describes a straight line with unit intercept and slope equal to minus 1, which is the negative diagonal. Since TPF = sensitivity and FPF = 1- specificity, another way of stating this is that at the intersection with the negative diagonal, sensitivity equals specificity.

#### 10.11.4 Area under the ROC curve

The area AUC (abbreviation for area under curve) under the ROC curve suggests itself as a measure of performance that is independent of threshold and therefore circumvents the ambiguity issue of comparing sensitivity/specificity pairs, and has other advantages. It is defined by the following integrals:

$$\begin{aligned} A_{z;\sigma=1} &= \int_0^1 TPF(\zeta) d(FPF(\zeta)) \\ &= \int_0^1 FPF(\zeta) d(TPF(\zeta)) \end{aligned} \tag{10.22}$$

Eqn. (10.22) has the following equivalent interpretations:

- The first form performs the integration using thin vertical strips, e.g., extending from  $x$  to  $x + dx$ , where for convenience  $x$  is a temporary symbol for FPF. The area can be interpreted as the average TPF over all possible values of FPF.
- The second form performs the integration using thin horizontal strips, e.g., extending from  $y$  to  $y + dy$ , where for convenience  $y$  is a temporary symbol for TPF. The area can be interpreted as the average FPF over all possible values of TPF.

By convention, the symbol  $A_z$  is used for the area under the binormal model predicted ROC curve. In Eqn. (10.22), the extra subscript  $\sigma = 1$  is necessary to distinguish it from another one corresponding to the unequal variance binormal model to be derived later. It can be shown that:

$$A_{z;\sigma=1} = \Phi\left(\frac{\mu}{\sqrt{2}}\right) \quad (10.23)$$

Since the ROC curve is bounded by the unit square, AUC must be between zero and one. If  $\mu$  is non-negative, the area under the ROC curve must be between 0.5 and 1. The chance diagonal, corresponding to  $\mu = 0$ , yields  $A_{z;\sigma=1} = 0.5$ , while the perfect ROC curve, corresponding to infinite yields unit area. Since it is a scalar quantity, AUC can be used to less-ambiguously quantify performance in the ROC task than is possible using sensitivity - specificity pairs.

### 10.11.5 Properties of the equal-variance binormal model ROC curve

- a. The ROC curve is completely contained within the unit square. This follows from the fact that both axes of the plot are probabilities.
- b. The operating point rises monotonically from (0,0) to (1,1).
- c. Since  $\mu$  is positive, the slope of the equal-variance binormal model curve at the origin (0,0) is infinite and the slope at (1,1) is zero, and the slope along the curve is always non-negative and decreases monotonically as the operating point moves up the curve.
- d. AUC is a monotone increasing function of  $\mu$ . It varies from 0.5 to 1 as  $\mu$  varies from zero to infinity.

### 10.11.6 Comments

Property (b): since the operating point coordinates can both be expressed in terms of  $\Phi$  functions, which are monotone in their arguments, and in each case the argument appears with a negative sign, it follows that as  $\zeta$  is lowered both TPF and FPF increase. In other words, the operating point corresponding to  $\zeta - d\zeta$  is to the upper right of that corresponding  $\zeta$  to (assuming  $d\zeta > 0$ ).

Property (c): The slope of the ROC curve can be derived by differentiation ( $\mu$  is constant):

$$\begin{aligned} \frac{d(TPF)}{d(FPF)} &= \frac{d(\Phi(\mu - \zeta))}{d(\Phi(-\zeta))} \\ &= \frac{\phi(\mu - \zeta)}{\phi(-\zeta)} \\ &= \exp(\mu(\zeta - \mu/2)) \propto \exp(\mu\zeta) \end{aligned} \quad \left. \right\} \quad (10.24)$$

The above derivation uses the fact that the differential of the CDF function yields the pdf function, i.e.,

$$d\Phi(\zeta) = P(\zeta < Z < \zeta + d\zeta) = \phi(\zeta)d\zeta$$

Since the slope of the ROC curve can be expressed as a power of  $e$ , it is always non-negative. Provided  $\mu > 0$ , then, in the limit  $\zeta \rightarrow \infty$ , the slope at the origin approaches  $\infty$ . Eqn. (10.24) also implies that in the limit  $\zeta \rightarrow -\infty$  the slope of the ROC curve at the end-point (1,1) approaches zero, i.e., the slope is a monotone increasing function of  $\zeta$ . As  $\zeta$  decrease from  $+\infty$  to  $-\infty$ , the slope decreases monotonically from  $+\infty$  to 0.

Fig. 10.5 is the ROC curve for the equal-variance binormal model for  $\mu$ . The entire curve is defined by  $\zeta$ . Specifying a particular value of  $\zeta$  corresponds to specifying a particular point on the ROC curve. In Fig. 3.5 the open circle corresponds to the operating point (0.159, 0.977) defined by  $\zeta = 1$ ;  $\text{pnorm}(-1) = 0.159$ ;  $\text{pnorm}(3-1) = 0.977$ . The operating point lies exactly on the curve, as this is a predicted operating point.

```
mu <- 3;zeta <- seq(-4,mu+3,0.05)
FPF <- pnorm(-zeta)
TPF <- pnorm(mu -zeta)
FPF <- c(1, FPF, 0);TPF <- c(1, TPF, 0)
curveData <- data.frame(FPF = FPF, TPF = TPF)
OpX <- pnorm(-1)
OpY <- pnorm(mu-1)
pointData <- data.frame(FPF = OpX, TPF = OpY)
rocPlot <- ggplot(
  mapping = aes(x = FPF, y = TPF)) +
  xlab("FPF") + ylab("TPF" ) +
  geom_line(data = curveData, size = 2) +
  geom_point(data = pointData, size = 5) +
  theme(axis.title.y = element_text(size = 25,face="bold"),
        axis.title.x = element_text(size = 30,face="bold")) +
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0))
print(rocPlot)
```

### 10.11.7 Physical interpretation of the mu-parameter

As a historical note,  $\mu$  is equivalent (Macmillan and Creelman, 1991) to a signal detection theory variable denoted  $d'$  in the literature (pronounced “dee-prime”). It can be thought of as the *perceptual signal to noise ratio* (pSNR) of diseased cases relative to non-diseased ones. It is a measure of reader expertise and / or

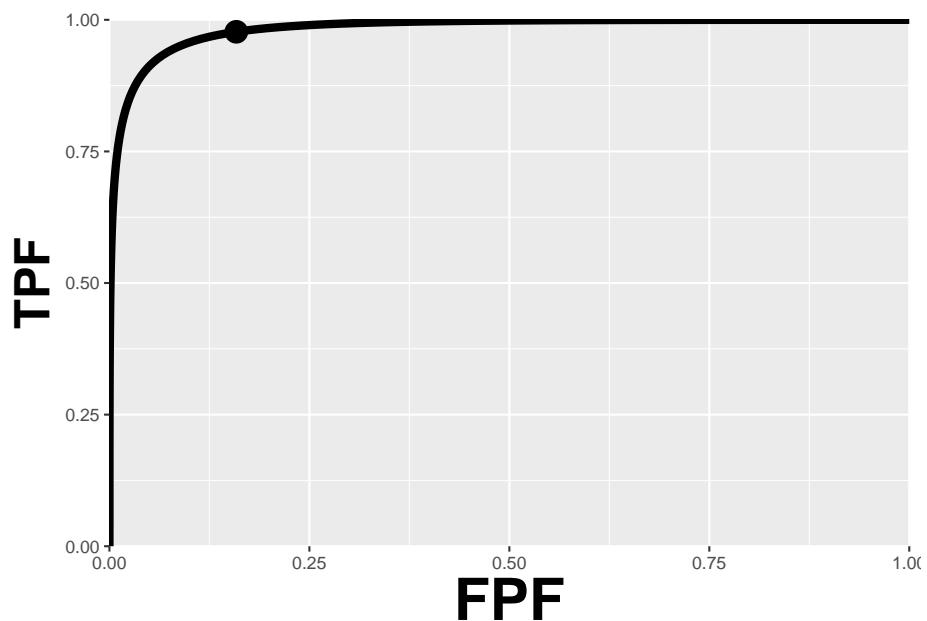


Figure 10.5: ROC curve predicted by equal variance binormal model for  $\mu = 3$ . The circled operating point corresponds to  $\zeta = 1$ . The operating point falls exactly on the curve, as these are analytical results. Due to sampling variability, with finite numbers of cases, this is not observed in practice.

ease of detectability of the disease. SNR is a term widely used in engineering, specifically in signal detection theory (Green and Swets, 1966; Egan, 1975), it dates to the early 1940s when one had the problem (USAirForce, 1947) of detecting faint radar reflections from a plane against a background of noise. The reader may be aware of the “rule-of-thumb” that if SNR exceeds three the target is likely to be detected. It will be shown later that the area under the ROC curve is the probability that a diseased case Z-sample is greater than that of a non-diseased one. The following code snippet shows that for  $\mu = 3$ , the probability of detection is 98.3 percent.

```
pnorm(3/sqrt(2))
#> [1] 0.983
```

For electrical signals, SNR can be measured with instruments but, in the context of decisions, measured is the perceptual SNR. Physical characteristics that differentiate non-diseased from diseased cases, and how well they are displayed will affect it; in addition the eye-sight of the observer is an obvious factor; not so obvious is how information is processed by the cognitive system, and the role of the observer’s experience in making similar decisions (i.e., expertise).

## 10.12 Assigning confidence intervals to an operating point

- The notation in the following equations follows that introduced in Chapter 02.
- A  $(1-\alpha)$  confidence interval (CI) of a statistic is the range that is expected to contain the true value of the statistic with probability  $(1 - \alpha)$ .
- It should be clear that a 99 percent CI is wider than a 95 percent CI, and a 90percentCI is narrower; in general, the higher the confidence that the interval contains the true value, the wider the range of the CI.
- Calculation of a parametric confidence interval requires a distributional assumption (non-parametric estimation methods, which use resampling methods, are described later). With a distributional assumption, the method being described now, the parameters of the distribution can be estimated, and since the distribution accounts for variability, the needed confidence interval estimate follows.
- With TPF and FPF, each of which involves a ratio of two integers, it is convenient to assume a *binomial* distribution for the following reason:
- The diagnosis “non-diseased” vs. “diseased” is a Bernoulli trial, i.e., one whose outcome is binary.
- A Bernoulli trial is like a coin-toss, a special coin whose probability of landing “diseased” face up is  $p$ , which is not necessarily 0.5 as with a real coin.

- It is a theorem in statistics that the total number of Bernoulli outcomes of one type, e.g.,  $n(FP)$ , is a binomial-distributed random variable, with success probability  $\widehat{FPF}$  and trial size  $K_1$ . The circumflex denotes an estimate.

$$n(FP) \sim B(K_1, \widehat{FPF}) \quad (10.25)$$

In Eqn. (10.25),  $B(n, p)$  denotes the binomial distribution with success probability  $p$  and trial size  $n$ :

$$\left. \begin{array}{l} k \sim B(n, p) \\ k = 0, 1, 2, \dots, n \end{array} \right\} \quad (10.26)$$

Eqn. (10.26) states that  $k$  is a random sample from the binomial distribution  $B(n, p)$ . For reference, the probability mass function pmf of  $B(n, p)$  is defined by (the subscript  $Bin$  denotes a binomial distribution):

$$\text{pmf}_{Bin}(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k} \quad (10.27)$$

For a discrete distribution, one has probability *mass* function; in contrast, for a continuous distribution one has a probability *density* function.

The binomial coefficient  $\binom{n}{k}$  appearing in Eqn. (10.27), to be read as “ $n$  pick  $k$ ”, is defined by:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (10.28)$$

From the properties of the binomial distribution the variance of  $n(FP)$  is given by:

$$\sigma_{n(FP)}^2 = K_1 \widehat{FPF} (1 - \widehat{FPF}) \quad (10.29)$$

It follows that  $FPF$  has mean  $\widehat{FPF}$  and variance  $\sigma_{FPF}^2$  given by (using theorem  $Var(aX) = a^2 Var(X)$ , where  $a$  is a constant, equal to  $1/K_1$  in this case):

$$\sigma_{FPF}^2 = \frac{\widehat{FPF} (1 - \widehat{FPF})}{K_1} \quad (10.30)$$

For large  $K_1$  the distribution of  $FPF$  approaches a normal distribution as follows:

$$\widehat{FPF} \sim N\left(\widehat{FPF}, \sigma_{FPF}^2\right)$$

This immediately allows us to write down the confidence interval for  $\widehat{FPF}$ , i.e.,  $\pm z_{\alpha/2}$  around  $\widehat{FPF}$ .

$$CI_{1-\alpha}^{FPF} = \left(\widehat{FPF} - z_{\alpha/2}\sigma_{FPF}, \widehat{FPF} + z_{\alpha/2}\sigma_{FPF}\right) \quad (10.31)$$

In Eqn. (10.31),  $z_{\alpha/2}$  is the upper  $\alpha/2$  quantile of the unit normal distribution, i.e., the area to the *right* under the unit normal distribution pdf from  $z_{\alpha/2}$  to  $\infty$  equals  $\alpha/2$ . It is the complement (i.e., plus goes to minus) of  $\Phi^{-1}(\alpha/2)$  introduced earlier; the difference is that the latter uses the area to the *left*. The following code might help.

```
alpha <- 0.05
# this is z_{\alpha/2}, the upper \alpha/2 quantile
qnorm(1-alpha/2)
#> [1] 1.96
# this is \Phi^{-1}(\alpha/2), the lower \alpha/2 quantile
qnorm(alpha/2)
#> [1] -1.96
```

Here is the definition of  $z_{\alpha/2}$ :

$$\left. \begin{aligned} z_{\alpha/2} &= \Phi^{-1}(1 - \alpha/2) \\ \alpha/2 &= \int_{z_{\alpha/2}}^{\infty} \phi(z) dz \\ &= 1 - \Phi(z_{\alpha/2}) \end{aligned} \right\} \quad (10.32)$$

The normal approximation is adequate if both of the following two conditions are both met:  $K_1 \widehat{FPF} > 10$  and  $K_1(1 - \widehat{FPF}) > 10$ . This means, essentially, that  $\widehat{FPF}$  is not too close to zero or 1.

Similarly, an approximate symmetric  $(1 - \alpha)$  confidence interval for TPF is:

$$CI_{1-\alpha}^{TPF} = \left(\widehat{TPF} - z_{\alpha/2}\sigma_{TPF}, \widehat{TPF} + z_{\alpha/2}\sigma_{TPF}\right) \quad (10.33)$$

In Eqn. (10.33),

$$\sigma_{TPF}^2 = \frac{\widehat{TPF}(1 - \widehat{TPF})}{K_2} \quad (10.34)$$

The confidence intervals are largest when the probabilities (FPF or TPF) are close to 0.5 and decrease inversely as the square root of the relevant number of cases. The symmetric binomial distribution based estimates can stray outside the allowed range (0 to 1). Exact confidence intervals<sup>9</sup> that are asymmetric around the central value and which are guaranteed to be in the allowed range can be calculated: it is implemented in R in function `binom.test()` and used below (The approximate confidence intervals can exceed the allowed ranges, but the exact confidence intervals do not):

```

options(digits=3)
seed <- 100; set.seed(seed)
alpha <- 0.05; K1 <- 99; K2 <- 111; mu <- 5; zeta <- mu/2
cat("alpha = ", alpha,
    "\nK1 = ", K1,
    "\nK2 = ", K2,
    "\nmu = ", mu,
    "\nzeta = ", zeta, "\n")
#> alpha = 0.05
#> K1 = 99
#> K2 = 111
#> mu = 5
#> zeta = 2.5
z1 <- rnorm(K1)
z2 <- rnorm(K2) + mu
nTN <- length(z1[z1 < zeta])
nTP <- length(z2[z2 >= zeta])
Sp <- nTN/K1; Se <- nTP/K2
cat("Specificity = ", Sp,
    "\nSensitivity = ", Se, "\n")
#> Specificity = 0.99
#> Sensitivity = 0.991

# Approx binomial tests
cat("approx 95percent CI on Specificity = ",
    -abs(qnorm(alpha/2))*sqrt(Sp*(1-Sp)/K1)+Sp,
    +abs(qnorm(alpha/2))*sqrt(Sp*(1-Sp)/K1)+Sp, "\n")
#> approx 95percent CI on Specificity = 0.97 1.01

# Exact binomial test
ret <- binom.test(nTN, K1, p = nTN/K1)
cat("Exact 95percent CI on Specificity = ",
    as.numeric(ret$conf.int), "\n")
#> Exact 95percent CI on Specificity = 0.945 1

# Approx binomial tests
cat("approx 95percent CI on Sensitivity = ",

```

```

-abs(qnorm(alpha/2))*sqrt(Se*(1-Se)/K2)+Se,
+abs(qnorm(alpha/2))*sqrt(Se*(1-Se)/K2)+Se, "\n")
#> approx 95percent CI on Sensitivity = 0.973 1.01

# Exact binomial test
ret <- binom.test(nTP, K2, p = nTP/K2)
cat("Exact 95percent CI on Sensitivity = ",
    as.numeric(ret$conf.int), "\n")
#> Exact 95percent CI on Sensitivity = 0.951 1

```

Note the usage of the *absolute* value of the `qnorm()` function; `qnorm` is the lower quantile function for the unit normal distribution, identical to  $\Phi^{-1}(0.025)$ , i.e., about -1.96, and  $z_{\alpha/2}$  is the upper quantile.

## 10.13 Variability in sensitivity and specificity: the Beam et al study

In this study (Beam et al., 1996) fifty accredited mammography centers were randomly sampled in the United States. “Accredited” is a legal/regulatory term implying, among other things, that the radiologists interpreting the breast cases were “board certified” by the American Board of Radiology. One hundred eight (108) certified radiologists from these centers gave blinded interpretation to a common set of 79 randomly selected enriched screening cases containing 45 cases with cancer and the rest normal or with benign lesions. Ground truth for these women had been established either by biopsy or by 2-year follow-up (establishing truth is often the most time consuming part of conducting an ROC study). The observed range of sensitivity (TPF) was 53percent and the range of FPF was 63percent; the corresponding range for AUC was 21percent, Table 10.2.

```

results <- array(dim = c(3,3))
results[1,] <- c(46.7, 100, 53.3)
results[2,] <- c(36.3, 99.3, 63.0)
results[3,] <- c(0.74, 0.95, 0.21)
df <- as.data.frame(results)
rownames(df) <- c("Sensitivity", "Specificity", "AUC")
colnames(df) <- c("Min", "Max", "Range")

```

In Fig. 10.6, a schematic of the data, if one looks at the points labeled (B) and (C) one can mentally construct a smooth ROC curve that starts at (0,0), passes roughly through these points and ends at (1,1). In this sense, the intrinsic performances (i.e., AUCs or equivalently the parameter) of the two radiologists are similar. The only difference between them is that radiologist (B) is using

Table 10.2: The variability of 108 radiologists on a common dataset of screening mammograms. Note the reduced variability when one uses AUC, which accounts for variations in reporting thresholds (AUC variability range is 21percent compared to 53percent for sensitivity and 63percent for specificity).

	Min	Max	Range
Sensitivity	46.70	100.00	53.30
Specificity	36.30	99.30	63.00
AUC	0.74	0.95	0.21

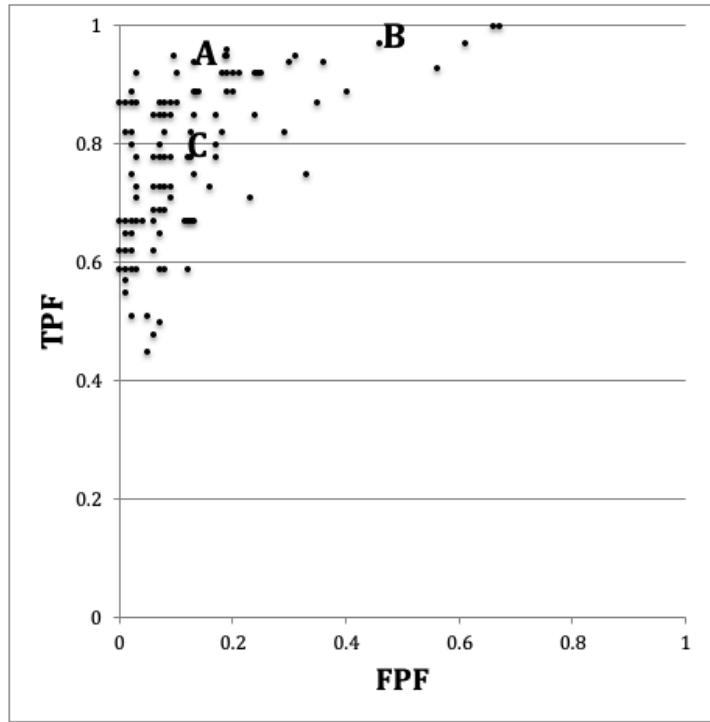


Figure 10.6: Schematic, patterned from the Beam et al study, showing the ROC operating points of 108 mammographers. Wide variability in sensitivity (40percent) and specificity (45percent) are evident. Radiologists (B) and (C) appear to be trading sensitivity for specificity and vice versa, while radiologist A's performance is intrinsically superior. See summary of important principles below.

lower threshold relative to the radiologist (C). Radiologist (C) is more concerned with minimizing FPs while radiologist (B) is more concerned with maximizing sensitivity. By appropriate feedback radiologist (C) can perhaps be induced to change the threshold to that of radiologist (B), or they both could be induced to achieve a happy compromise. An example of feedback might be: “you are missing too many cancers and this could get us all into trouble; worry less about reduced specificity and more about increasing your sensitivity”. In contrast, radiologist (A) has intrinsically greater performance (B) or (C). No change in threshold is going to get the other two to a similar level of performance as radiologist A. Extensive training will be needed to bring the under-performing radiologists to the expert level represented by radiologist A.

Fig. 10.6 and Table 10.2 illustrate several important principles. 1. Since an operating point is characterized by two values, unless both numbers are higher (e.g., radiologist A vs. B or C), it is difficult to unambiguously compare them. 2. While sensitivity and specificity depend on the reporting threshold, the area under the ROC plot is independent of it. Using the area under the ROC curve one can unambiguously compare two readers. 3. Combining sensitivity and the complement of specificity into a single AUC measure yields the additional benefit of lower variability. In Fig. 10.6, the range for sensitivity is 53 percent while that for specificity is 63 percent. In contrast, the range for AUC is only 21 percent. This means that much of the observed variations in sensitivity and specificity are due to variations in thresholds, and using AUC eliminates this source of variability. Decreased variability of a measure is a highly desirable characteristic as it implies the measurement is more precise, making it easier to detect genuine changes between readers and / or modalities.

## 10.14 Summary

TBA ## Discussion{#binary-task-model-discussion} The concepts of sensitivity and specificity are of fundamental importance and are widely used in the medical imaging literature. However, it is important to realize that sensitivity and specificity do not provide a complete picture of diagnostic performance, since they represent performance at a particular threshold. As demonstrated in Fig. 3.6, expert observers can and do operate at different points, and the reporting threshold depends on cost-benefit considerations, disease prevalence and personal reporting styles. If using sensitivity and specificity the dependence on reporting threshold often makes it difficult to unambiguously compare observers. Even if one does compare them, there is loss of statistical power (equivalent to loss of precision of the measurement) due to the additional source of variability introduced by the varying thresholds.

The ROC curve is the locus of operating points as the threshold is varied. It and AUC are completely defined by the parameter of the equal variance binormal model. Since both are independent of reporting threshold , they overcome the

ambiguity inherent in comparing sensitivity/specificity pairs. Both are scalar measures of performance. AUC is widely used in assessing imaging systems. It should impress the reader that a subjective internal sensory perception of disease presence and an equally subjective internal threshold can be translated into an objective performance measure, such as the area under an ROC curve or equivalently, the parameter. The latter has the physical meaning of a perceptual signal to noise ratio.

The ROC curve predicted by the equal variance binormal model has a useful property, namely, as the threshold is lowered, its slope decreases monotonically. The predicted curve never crosses the chance diagonal, i.e., the predicted ROC curve is “proper”. Unfortunately, as one will see later, most ROC datasets are inconsistent with this model: rather, they are more consistent with a model where the diseased distribution has variance greater than unity. The consequence of this is an “improper” ROC curve, where in a certain range, which may be difficult to see when the data is plotted on a linear scale, the predicted curve actually crosses the chance diagonal and then its slope increases as it hooks up to reach (1,1). The predicted worse than chance performance is unreasonable. Models of ROC curves have been developed that do not have this unreasonable behavior: Chapter 17, Chapter 18 and Chapter 20.

The properties of the unit normal distribution and the binomial distribution were used to derive parametric confidence intervals for sensitivity and specificity. These were compared to exact confidence intervals. An important study was reviewed showing wide variability in sensitivity and specificity for radiologists interpreting a common set of cases in screening mammography, but smaller variability in areas under the ROC curve. This is because much of the variability in sensitivity and specificity is due to variation of the reporting threshold, which does not affect the area under the ROC curve. This is an important reason for preferring comparisons based on area under the ROC curve to those based on comparing sensitivity/specificity pairs.

This chapter has been demonstrated the equal variance binormal model with R examples. These were used to illustrate important concepts of case-sampling variability and its dependence on the numbers of cases. Again, while relegated for organizational reasons to online appendices, these appendices are essential components of the book. Most of the techniques demonstrated there will be reused in the remaining chapters. The motivated reader can learn much from studying the online material and running the different main-level functions contained in the software-directory corresponding to this chapter.

## 10.15 References

# Chapter 11

## Ratings Paradigm

### 11.1 TBA How much finished

80%

### 11.2 Introduction

In Chapter 9 the binary paradigm and associated concepts (e.g., sensitivity, specificity) were introduced. Chapter 9 introduced the concepts of a random scalar decision variable, or z-sample for each case, which is compared, by the observer to a fixed reporting threshold  $\zeta$ , resulting in two types of decisions. It described a statistical model, characterized by two unit-variance normal distributions separated by  $\mu$ , for the binary task. The concept of an underlying receiver operating characteristic (ROC) curve with the reporting threshold defining an operating point on the curve was introduced and the advisability of using the area under the curve as a measure of performance, which is independent of reporting threshold, was stressed.

In this chapter the more commonly used ratings method will be described, which yields greater definition to the underlying ROC curve than just one operating point obtained in the binary task, and moreover, is more efficient. In this method, the observer assigns a rating to each case. Described first is a typical ROC counts table and how operating points (i.e., pairs of FPF and TPF values) are calculated from the counts data. A labeling convention for the operating points is introduced. Notation is introduced for the observed integers in the counts table and the rules for calculating operating points are expressed as formulae and implemented in R. The ratings method is contrasted to the binary method, in terms of efficiency and practicality. A theme occurring repeatedly in this book, that the ratings are not numerical values but rather they are ordered

Table 11.1: Representative counts table.

	$r = 5$	$r = 4$	$r = 3$	$r = 2$	$r = 1$
non-diseased	1	2	8	19	30
diseased	22	12	5	6	5

labels is illustrated with an example. A method of collecting ROC data on a 6-point scale is described that has the advantage of yielding an unambiguous single operating point. The forced choice paradigm is described. Two controversies are described: one on the utility of discrete (e.g., 1 to 6) vs. quasi-continuous (e.g., 0 to 100) ratings and the other on the applicability of a clinical screening mammography-reporting scale for ROC analyses. Both of these are important issues and it would be a disservice to the readers of the book if I did not express my position on them.

### 11.3 The ROC counts table

In a positive-directed rating scale with five discrete levels, the ratings could be the ordered labels:

- “1”: definitely non-diseased,
- “2”: probably non-diseased,
- “3”: could be non-diseased or diseased,
- “4”: probably diseased,
- “5”: definitely diseased.

At the conclusion of the ROC study an ROC counts table is constructed. This is the generalization to rating studies of the  $2 \times 2$  decision vs. truth table introduced in Chapter 9, Table 9.1. This type of data representation is sometimes called a frequency table, but frequency usually means a rate of number of events per some unit, so I prefer the clearer term “counts”.

Table 11.1 is a representative counts table for a 5-rating study that summarizes the collected data. It is the starting point for analysis. It lists the number of counts in each ratings bin, listed separately for non-diseased and diseased cases, respectively. The data is from an actual clinical study (Barnes et al., 1989).

In this table:

- $r = 5$  means “rating equal to 5”
- $r = 4$  means “rating equal to 4”
- Etc.

There are  $K_1 = 60$  non-diseased cases and  $K_2 = 50$  diseased cases. Of the 60 non-diseased cases:

- one received the “5” rating,
- two the “4” rating,
- eight the “3” rating,
- 19 the “2” rating and
- 30 the “1” rating.

The distribution of counts is tilted towards the “1” rating end. In contrast, the distribution of the diseased cases is tilted towards the “5” rating end. Of the 50 diseased cases:

- 22 received the “5” rating,
- 12 the “4” rating,
- five the “3” rating,
- six the “2” rating and
- five the “1” rating.

A little thought should convince one that the observed tilting of the counts, towards the “1” end for actually non-diseased cases, and towards the “5” end for actually diseased cases, is reasonable.

The spread appears to be more pronounced for the diseased cases, e.g., five of the 50 cases appeared to be definitely non-diseased to the observer. However, one is forewarned not to jump to conclusions about the spread of the data being larger for diseased than for non-diseased cases based on observed rating alone. While it turns out to be true as will be shown later, the **ratings are merely ordered labels**, and modeling is required, see Chapter 13, that uses only the *ordering information* implicit in the labels, not the *actual values*, to reach quantitative conclusions.

## 11.4 Operating points from counts table

Table 11.2 illustrates how ROC operating points are calculated from the cell counts. In this table:

- $r \geq 5$  means “counting ratings greater than or equal to 5”
- $r \geq 4$  means “counting ratings greater than or equal to 4”
- Etc.

One starts with non-diseased cases that were rated five or more (in this example, since 5 is the highest allowed rating, the “or more” clause is inconsequential)

Table 11.2: Computation of operating points from cell counts.

	$r \geq 5$	$r \geq 4$	$r \geq 3$	$r \geq 2$	$r \geq 1$
FPF	0.0167	0.05	0.1833	0.5	1
TPF	0.4400	0.68	0.7800	0.9	1

and divides by the total number of non-diseased cases,  $K_1 = 60$ . This yields the abscissa of the lowest non-trivial operating point, namely  $FPF_{\geq 5} = 1/60 = 0.017$ . The subscript on FPF is intended to make explicit which ratings are being cumulated. The corresponding ordinate is obtained by dividing the number of diseased cases rated “5” or more and dividing by the total number of diseased cases,  $K_2 = 50$ , yielding  $TPF_{\geq 5} = 22/50 = 0.440$ . Therefore, the coordinates of the lowest operating point are  $(0.017, 0.44)$ . The abscissa of the next higher operating point is obtained by dividing the number of non-diseased cases that were rated “4” or more and dividing by the total number of non-diseased cases, i.e.,  $TPF_{\geq 4} = 3/60 = 0.05$ . Similarly the ordinate of this operating point is obtained by dividing the number of diseased cases that were rated “4” or more and dividing by the total number of diseased cases, i.e.,  $FPF_{\geq 4} = 34/50 = 0.680$ . The procedure, which at each stage cumulates the number of cases equal to or greater (in the sense of increased confidence level for disease presence) than a specified ordered label, is repeated to yield the rest of the operating points listed in Table 11.2. Since they are computed directly from the data, without any assumption, they are called empirical or observed operating points.

After doing this once, it would be nice to have a formula implementing the process, one use of which would be to code the procedure. But first one needs appropriate notation for the bin counts.

Let  $K_{1r}$  denote the number of non-diseased cases rated  $r$ , and  $K_{2r}$  denote the number of diseased cases rated  $r$ . For convenience, define dummy counts  $K_{1(R+1)} = K_{2(R+1)} = 0$ , where  $R$  is the number of ROC bins,  $R = 5$  in the current example. This construct allows inclusion of the origin  $(0,0)$  in the formulae. The range of  $r$  is  $r = 1, 2, \dots, (R + 1)$ . Within each truth-state, the individual bin counts sum to the total number of non-diseased and diseased cases, respectively. The following equations summarize all this:

$$K_1 = \sum_{r=1}^{R+1} K_{1r}$$

$$K_2 = \sum_{r=1}^{R+1} K_{2r}$$

$$K_{1(R+1)} = K_{2(R+1)} = 0$$

$$r = 1, 2, \dots, (R + 1)$$

The operating points are defined by:

$$\left. \begin{aligned} FPF_r &= \frac{1}{K_1} \sum_{s=r}^{R+1} K_{1s} \\ TPF_r &= \frac{1}{K_2} \sum_{s=r}^{R+1} K_{2s} \end{aligned} \right\} \quad (11.1)$$

### 11.4.1 Labeling the points

The labeling  $O_n$  of the points follows the following convention: From Eqn. (11.1), the point corresponding to  $r = 1$  would correspond to the upper right corner (1,1) of the ROC plot, a trivial operating point since it is common to all datasets, and is therefore not shown. The labeling starts with the next lower-left point, labeled  $O_1$ , which corresponds to  $r = 2$ ; the next lower-left point is labeled  $O_2$ , corresponding to  $r = 3$ , etc., and the point labeled  $O_4$  is the lowest non-trivial operating point corresponding to  $r = R = 5$  and finally  $O_R$  corresponding to  $r = R + 1$  is the origin (0,0) of the ROC plot, which is also a trivial operating point, because it is common to all datasets, and is therefore not shown. **To summarize, the operating points are labeled starting with the upper right corner, labeled  $O_1$ , and working down the curve, each time increasing the number by one. The total number of points is  $R - 1$ .** The relation between  $n$  in the label and  $r$  in Eqn. (11.1) is  $n = r - 1$ . An example of the labeling is shown in the next chapter, Fig. 12.1.

### 11.4.2 Examples

In the following examples  $R = 5$  is the number of ROC bins and  $K_{1(R+1)} = K_{2(R+1)} = 0$ . If  $r = 1$  one gets the uppermost “trivial” operating point (1,1):

$$FPF_1 = \frac{1}{K_1} \sum_{s=1}^{R+1} K_{1s} = \frac{60}{60} = 1 \quad TPF_1 = \frac{1}{K_2} \sum_{s=1}^{R+1} K_{2s} = \frac{50}{50} = 1$$

The uppermost non-trivial operating point is obtained for  $r = 2$ , when:

$$FPF_2 = \frac{1}{K_1} \sum_{s=2}^{R+1} K_{1s} = \frac{30}{60} = 0.5 \quad TPF_2 = \frac{1}{K_2} \sum_{s=2}^{R+1} K_{2s} = \frac{45}{50} = 0.9$$

The next lower operating point is obtained for  $r = 3$ :

$$FPF_3 = \frac{1}{K_1} \sum_{s=3}^{R+1} K_{1s} = \frac{11}{60} = 0.183 TPF_3 = \frac{1}{K_2} \sum_{s=3}^{R+1} K_{2s} = \frac{39}{50} = 0.780$$

The next lower operating point is obtained for  $r = 4$ :

$$FPF_4 = \frac{1}{K_1} \sum_{s=4}^{R+1} K_{1s} = \frac{3}{60} = 0.05 TPF_4 = \frac{1}{K_2} \sum_{s=4}^{R+1} K_{2s} = \frac{34}{50} = 0.680$$

The lowest non-trivial operating point is obtained for  $r = 5$ :

$$FPF_5 = \frac{1}{K_1} \sum_{s=5}^{R+1} K_{1s} = \frac{1}{60} = 0.017 TPF_5 = \frac{1}{K_2} \sum_{s=5}^{R+1} K_{2s} = \frac{22}{50} = 0.440$$

The next value  $r = 6$  yields the trivial operating point (0,0):

$$FPF_6 = \frac{1}{K_1} \sum_{s=6}^{R+1} K_{1s} = \frac{0}{60} = 0 TPF_6 = \frac{1}{K_2} \sum_{s=6}^{R+1} K_{2s} = \frac{0}{50} = 0$$

This exercise shows explicitly that an R-rating ROC study can yield, at most,  $R + 1$  distinct non-trivial operating points; i.e., those corresponding to  $r = 2, 3, \dots, R$ .

The modifier “at most” is needed, because if both counts (i.e., non-diseased and diseased) for bin  $r'$  are zeroes, then that operating point merges with the one immediately below-left of it:

$$FPF_{r'} = \frac{1}{K_1} \sum_{s=r'}^{R+1} K_{1s} = \frac{1}{K_1} \sum_{s=r'+1}^{R+1} K_{1s} = FPF_{r'+1} TPF_{r'} = \frac{1}{K_2} \sum_{s=r'}^{R+1} K_{2s} = \frac{1}{K_2} \sum_{s=r'+1}^{R+1} K_{2s} = TPF_{r'+1}$$

Since bin  $r'$  is unpopulated, one can re-label the bins to exclude the unpopulated bin, and now the total number of bins is effectively  $R - 1$ .

Since one is cumulating counts, which cannot be negative, the highest non-trivial operating point resulting from cumulating the 2 through 5 ratings has to be to the upper-right of the next adjacent operating point resulting from cumulating the 3 through 5 ratings. This in turn has to be to the upper-right of the operating point resulting from cumulating the 4 through 5 ratings. This in turn has to be to the upper right of the operating point resulting from the 5 ratings. In other words, as one cumulates ratings bins, the operating point must move

monotonically up and to the right, or more accurately, the point cannot move down or to the left. If a particular bin has zero counts for non-diseased cases, and non-zero counts for diseased cases, the operating point moves vertically up when this bin is cumulated; if it has zero counts for diseased cases, and non-zero counts for non-diseased cases, the operating point moves horizontally to the right when this bin is cumulated.

## 11.5 Automating all this

It is useful to replace the preceding detailed explanation with a simple algorithm, as in the following code (see first seven lines):

```
options(digits = 3)
FPF <- OpPts[1,]
TPF <- OpPts[2,]
df <- data.frame(FPF = FPF, TPF = TPF)
df <- t(df)
print(df)
#>      [,1] [,2] [,3] [,4] [,5]
#> FPF 0.0167 0.05 0.183 0.5     1
#> TPF 0.4400 0.68 0.780 0.9     1
mu <- qnorm(.5)+qnorm(.9);sigma <- 1
Az <- pnorm(mu/sqrt(2))
cat("uppermost point based estimate of mu = ", mu, "\n")
#> uppermost point based estimate of mu = 1.28
cat("corresponding estimate of Az = ", Az, "\n")
#> corresponding estimate of Az = 0.818
```

Notice that the values of the arrays FPF and TPF are identical to those listed in Table 11.2. Regarding the last four lines of code, it was shown in Chapter 9 that in the equal variance binormal model the operating point determines the parameters  $\mu = 1.282$ , Eqn. (10.17), or equivalently  $A_{z;\sigma=1} = 0.818$ , Eqn. (10.23). The last four lines illustrate the application of these formulae using the coordinates (0.5, 0.9) of the uppermost non-trivial operating point, i.e., one is fitting the equal variance model to the uppermost operating point.

Shown next is the equal-variance model fit to the uppermost non-trivial operating point, left plot, and for comparison, the right plot is the unequal variance model fit to all operating points. The unequal variance model is the subject of an upcoming chapter.

```
# equal variance fit to uppermost operating point
p1 <- plotROC (mu, sigma, FPF, TPF)
# the following values are from unequal-variance model fitting
```

```

# to be discussed later
mu <- 2.17; sigma <- 1.65
# this formula to be discussed later
Az <- pnorm(mu/sqrt(1+sigma^2))
cat("binormal unequal variance model estimate of Az = ", Az, "\n")
#> binormal unequal variance model estimate of Az =  0.87
# unequal variance fit to all operating points
p2 <- plotROC (mu, sigma, FPF, TPF)

grid.arrange(p1,p2,ncol=2)

```

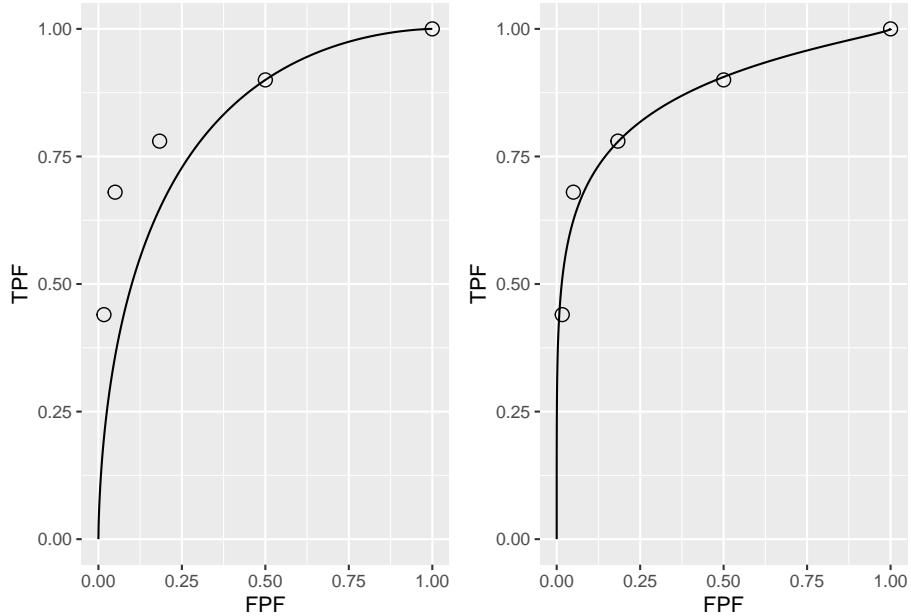


Figure 11.1: (A): The left figure is the predicted ROC curve for  $\mu = 1.282$  superposed on the operating points. (B): The right figure is the same data fitted with a two-parameter model described later.

It should come as no surprise that the uppermost operating point is *exactly* on the predicted curve: after all, this point was used to calculate  $\mu = 2.17$ . The corresponding value of  $\zeta$  can be calculated from Eqn. (3.17), namely:

$$\zeta = \Phi^{-1}(Sp)$$

$$\mu = \zeta + \Phi^{-1}(Se)$$

These are coded below:

```
qnorm(1-0.5)
#> [1] 0
mu=qnorm(0.9)
#> [1] 0.888
```

Either way, one gets the same result:  $\zeta = 0$ . It should be clear that this makes sense: FPF = 0.5 is consistent with half of the (symmetrical) unit-normal non-diseased distribution being above  $\zeta = 0$ . The transformed value  $\zeta$  (zero in this example) is a genuine numerical value. *To reiterate, ratings cannot be treated as genuine numerical values, but thresholds, estimated from an appropriate model, can be treated as genuine numerical values.*

Exercise: calculate  $\zeta$  for each of the remaining operating points. *Notice that  $\zeta$  increases as one moves down the curve.*

- In Fig. 11.1 (A), the ROC curve, as determined by the uppermost operating point, passes exactly through this point but misses the others. If a different operating point were used to estimate  $\mu$  and  $A_{z;\sigma=1}$ , the estimated values would have been different and the new curve would pass exactly through the *new* selected point. No single-point based choice of  $\mu$  would yield a satisfactory visual fit to all the observed operating points. **This is the reason one needs a modified model, with an extra parameter, namely the unequal variance binormal model, to fit radiologist data** (the extra parameter is the ratio of the standard deviations of the two distributions).
- Fig. 11.1 (B) shows the predicted ROC curve by the unequal variance binormal model, to be introduced in Chapter 06. The corresponding parameter values are  $\mu = 2.17$  and  $\sigma = 1.65$ .
- Notice the improved visual quality of the fit. Each observed point is “not engraved in stone”, rather both FPF and TPF are subject to sampling variability. Estimation of confidence intervals for FPF and TPF was addressed, see (10.31) and (10.33). [A detail: the estimated confidence interval in the preceding chapter was for a single operating point; since the multiple operating points are correlated – some of the counts used to calculate them are common to two or more operating points – the method tends to overestimate the confidence interval. A modeling approach to estimating confidence intervals accounts for these correlations and yields tighter confidence intervals.]

## 11.6 Relation between ratings paradigm and the binary paradigm

Table 11.1 and Table 11.2 correspond to  $R = 5$ . In Chapter 9 it was shown that the binary task requires a single fixed threshold parameter  $\zeta$  and a decision or binning rule Eqn. (11.2): assign the case a diseased rating of 2 if  $Z > \zeta$  and a rating of 1 otherwise.

**The R-rating task can be viewed as  $R - 1$  simultaneously conducted binary tasks each with its own fixed threshold  $\zeta_r$ , where  $r = 1, 2, \dots, R - 1$ . It is efficient compared to  $R - 1$  sequentially conducted binary tasks; however, the onus is on the observer to maintain fixed-multiple thresholds through the duration of the study.**

The rating method is a more efficient way of collecting the data compared to running the study repeatedly with appropriate instructions to cause the observer to adopt different fixed thresholds specific to each replication. In the clinical context such repeated studies would be impractical because it would introduce memory effects, wherein the diagnosis of a case would depend on how many times the case had been seen, along with other cases, in previous sessions. A second reason is that it is difficult for a radiologist to change the operating threshold in response to instructions. To my knowledge, repeated use of the binary paradigm has not been used in any clinical ROC study

In order to model the binning, one defines dummy thresholds  $\zeta_0 = -\infty$  and  $\zeta_R = +\infty$ , in which case the thresholds satisfy the ordering requirement  $\zeta_{r-1} \leq \zeta_r$ ,  $r = 1, 2, \dots, R$ . The rating or binning rule is:

$$\left. \begin{aligned} \text{if } (\zeta_{r-1} \leq z < \zeta_r) \Rightarrow \text{rating} = r \\ r = 1, 2, \dots, R \end{aligned} \right\} \quad (11.2)$$

For Table 11.2, the **empirical** thresholds are as follows:

$$\left. \begin{aligned} \zeta_r &= r + 1 \\ r &= 1, 2, \dots, R - 1 \\ \zeta_0 &= -\infty \\ \zeta_R &= \infty \end{aligned} \right\} \quad (11.3)$$

The empirical thresholds are integers, as distinct from the floating point values predicted by Eqn. (11.5). **Either way one gets the same operating points.** This is a subtle and important distinction, which is related to the next section: one has enormous flexibility in the choice of the scale adopted for the decision variable axis.

In Table 11.1 the number of bins is  $R = 5$ . The “simultaneously conducted binary tasks” nature of the rating task can be appreciated from the following

examples. Suppose one selects the threshold for the first binary task to be  $\zeta_4 = 5$ . By definition,  $\zeta_5 = \infty$ ; therefore a case rated 5 satisfies the binning rule  $\zeta_4 \leq 5 < \zeta_5$ , i.e., Eqn. (11.2). The operating point corresponding to  $\zeta_4 = 5$ , obtained by cumulating all cases rated five, yields  $(0.017, 0.440)$ . In the second binary-task, one selects as threshold  $\zeta_3 = 4$ . Therefore, a case rated four satisfies the binning rule  $\zeta_3 \leq 4 < \zeta_4$ . The operating point corresponding to  $\zeta_3 = 4$ , obtained by cumulating all cases rated four or five, yields  $(0.05, 0.680)$ . Similarly, for  $\zeta_2 = 3$ ,  $\zeta_1 = 2$  and  $\zeta_0 = -\infty$ , which yield counts in bins 3, 2 and 1, respectively. The last is a trivial operating point. The non-trivial operating points are generated by thresholds  $\zeta_r$ , where  $r = 1, 2, 3$  and 4. A five-rating study has four associated thresholds and a corresponding number of equivalent binary studies. In general, an  $R$  rating study has  $R - 1$  associated thresholds.

## 11.7 Ratings are not numerical values

The ratings are to be thought of as ordered labels, not as numeric values. Arithmetic operations that are allowed on numeric values, such as averaging, are not allowed on ratings. One could have relabeled the ratings in Table 4.2 as A, B, C, D and E, where  $A < B$  etc. As long as the counts in the body of the table are unaltered, such relabeling would have no effect on the observed operating points and the fitted curve. Of course one cannot average the labels A, B, etc. of different cases. The issue with numeric labels is not fundamentally different. At the root is that the difference in thresholds corresponding to the different operating points are not in relation to the difference between their numeric values. There is a way to estimate the underlying thresholds, if one assumes a specific model, for example the unequal-variance binormal model to be described in Chapter 06. The thresholds so obtained are genuine numeric values and can be averaged. [Not to hold the reader in suspense, the four thresholds corresponding to the data in Table 4.1 are 0.007676989, 0.8962713, 1.515645 and 2.396711; see §6.4.1; these values would be unchanged if, for example, the labels were doubled, with allowed values 2, 4, 6, 8 and 10, or any of an infinite number of rearrangements that preserves their ordering.]

The temptation to regard confidence levels / ratings as numeric values can be particularly strong when one uses a large number of bins to collect the data. One could use of quasi-continuous ratings scale, implemented for example, by having a slider-bar user interface for selecting the rating. The slider bar typically extends from 0 to 100, and the rating could be recorded as a floating-point number, e.g., 63.45. Here too one cannot assume that the difference between a zero-rated case and a 10 rated case is a tenth of the difference between a zero-rated case and a 100 rated case. So averaging the ratings is not allowed. Additionally, one cannot assume that different observers use the labels in the same way. One observer's 4-rating is not equivalent to another observers 4-rating. Working directly with the ratings is a bad idea: valid analytical methods use the rankings of the ratings, not their actual values. The reason for the

emphasis is that there are serious misconceptions about ratings. I am aware of a publication stating, to the effect, that a modality resulted in an increase in average confidence level for diseased cases. Another publication used a specific numerical value of a rating to calculate the operating point for each observer – this assumes all observers use the rating scale in the same way.

## 11.8 A single “clinical” operating point from ratings data

The reason for the quotes in the title to this section is that a single operating point on a laboratory ROC plot, no matter how obtained, has little relevance to how radiologists operate in the clinic. However, some consider it useful to quote an operating point from an ROC study. For a 5-rating ROC study, Table 11.1, it is not possible to unambiguously calculate the operating point of the observer in the binary task of discriminating between non-diseased and diseased cases. One possibility would be to use the “three and above” ratings to define the operating point, but one might just have well have chosen “two and above”. A second possibility is to instruct the radiologist that a “four and above” rating, for example, implies the case would be reported “clinically” as diseased. However, the radiologist can only pretend so far that this study, which has no clinical consequences, is somehow a “clinical” study.

If a single laboratory study based operating point is desired (Nishikawa, 2012), the best strategy, in my opinion, is to obtain the rating via two questions. This method is also illustrated in Table 3.1 of a book on detection theory (Macmillan and Creelman, 1991). The first question is “is the case diseased?” The binary (Yes/No) response to this question allows unambiguous calculation of the operating point, as in Chapter 9. The second question is: “what is your confidence in your previous decision?” and allow three responses, namely Low, Medium and High. The dual-question approach is equivalent to a 6-point rating scale, Fig. 11.2. The answer to the first question, is the patient diseased, allows unambiguous construction of a single “clinical” operating point for disease presence. The answer to the second question, what is your confidence level in that decision, yields multiple operating points.

The ordering of the ratings can be understood as follows. The four, five and six ratings are as expected. If the radiologist states the patient is diseased and the confidence level is high that is clearly the highest end of the scale, i.e., six, and the lower confidence levels, five and four, follow, as shown. If, on the other hand, the radiologist states the patient is non-diseased, and the confidence level is high, then that must be the lowest end of the scale, i.e., “1”. The lower confidence levels in a negative decision must be higher than “1”, namely “2” and “3”, as shown. As expected, the low confidence ratings, namely “3” (non-diseased, low confidence) and “4” (diseased, low confidence) are adjacent to each other. With this method of data-collection, there is no

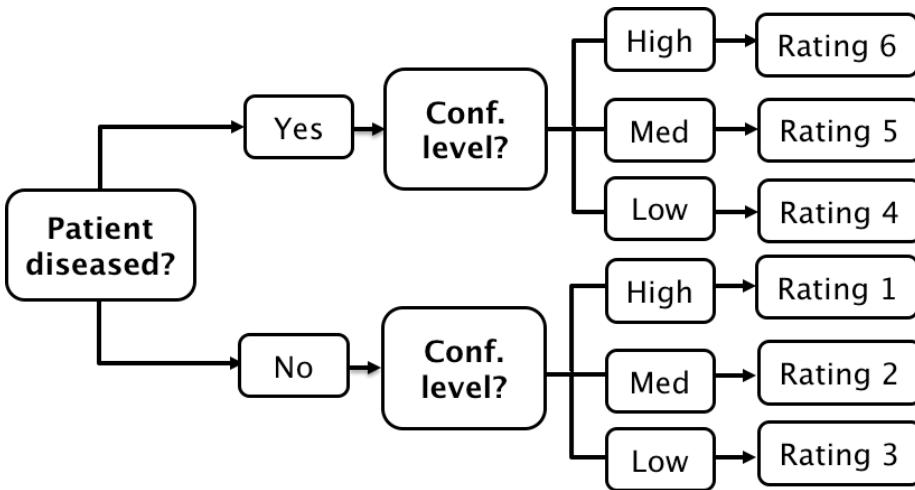


Figure 11.2: A method for acquiring ROC data on an effectively 6-point scale that also yields an unambiguous single operating point for declaring patients diseased. Note the reversal of the final ratings in the last “column” in the lower half of the figure.

confusion as to what rating defines the single desired operating point as this is determined by the binary response to the first question. The 6-point rating scale is also sufficiently fine to not smooth out the ability of the radiologist to maintain distinct different levels. In my experience, using this scale one expects rating noise of about  $\pm \frac{1}{2}$  a rating bin, i.e., the same difficult case, shown on different occasions to the same radiologist (with sufficient time lapse or other intervening cases to minimize memory effects) is expected to elicit a “3” or “4”, with roughly equal probability.

## 11.9 The forced choice paradigm

In each of the four paradigms (ROC, FROC, LROC and ROI) described in TBA Chapter 01, patient images are displayed one patient at a time. A fifth paradigm involves presentation of multiple images to the observer, where one image (or set of images from one patient, i.e., a case) is from a diseased patient, and the rest are from non-diseased patients. The observer’s task is to pick the image, or the case, that is most likely to be from the diseased patient. If the observer is correct, the event is scored as a “one” and otherwise it is scored as a “zero”. The process is repeated with other sets of independent patient images, each time satisfying the condition that one patient is diseased and the rest are non-diseased. The sum of the scores divided by the total number of

scores is the probability of a correct choice, denoted  $P(C)$ . If the total number of cases presented at the same time is denoted  $n$ , then the task is termed n-alternative forced choice or nAFC (Green and Swets, 1966). If only two cases are presented, one diseased and the other non-diseased, then  $n = 2$  and the task is 2AFC. In Fig. 11.3, in the left image a Gaussian nodule is superposed on a square region extracted from a non-diseased mammogram. The right image is a region extracted from a different non-diseased mammogram (one should not use the same background in the two images – the analysis assumes that different, i.e., independent images, are shown). If the observer clicks on the left image, a correct choice is recorded. [In some 2AFC-studies, the backgrounds are simulated non-diseased images. They resemble mammograms; the resemblance depends on the expertise of the observer: expert radiologists can tell that they are not true mammograms. They are actually created by filtering the random white noise with a  $1/f^3$  spatial filter (Burgess, 2011).]

The 2AFC paradigm is popular, because its analysis is straightforward, and there exists a theorem<sup>4</sup> that  $P(C)$ , the probability of a correct choice in the 2AFC task, equals, to within sampling variability, the *true* area under the true (not fitted, not empirical) ROC curve. Another reason for its popularity is possibly the speed at which data can be collected, sometimes only limited by the speed at which disk stored images can be displayed on the monitor. While useful for studies into human visual perception on relatively simple images, and the model observer community has performed many studies using this paradigm (Bochud et al., 1999), I cannot recommend it for clinical studies because *it does not resemble any clinical task*. In the clinic, radiologists never have to choose the diseased patient out of a pair consisting of one diseased and one non-diseased. Additionally, the forced-choice paradigm is wasteful of known-truth images, often a difficult/expensive resource to come by, because better statistics<sup>21</sup> (tighter confidence intervals) are obtained by the ratings ROC method or by utilizing location specific extensions of the ROC paradigm. [I am not aware of the 2AFC method being actually used to assess imaging systems using radiologists to perform real clinical tasks on real images.]

Fig. 11.3: Example of image presentation in a 2AFC study. The left image contains, at its center, a positive contrast Gaussian shape disk superposed on a non-diseased mammogram. The right image does not contain a lesion at its center and the background is from a different non-diseased patient. If the observer clicks on the left image it is recorded as a correct choice, otherwise it is recorded as an incorrect choice. The number of correct choices divided by the number of paired presentations is an estimate of the probability of a correct choice, which can be shown to be identical, apart from sampling variability, to the true area under the ROC curve. This is an example of a signal known exactly location known exactly (SKE-LKE) task widely used by the model observer community.

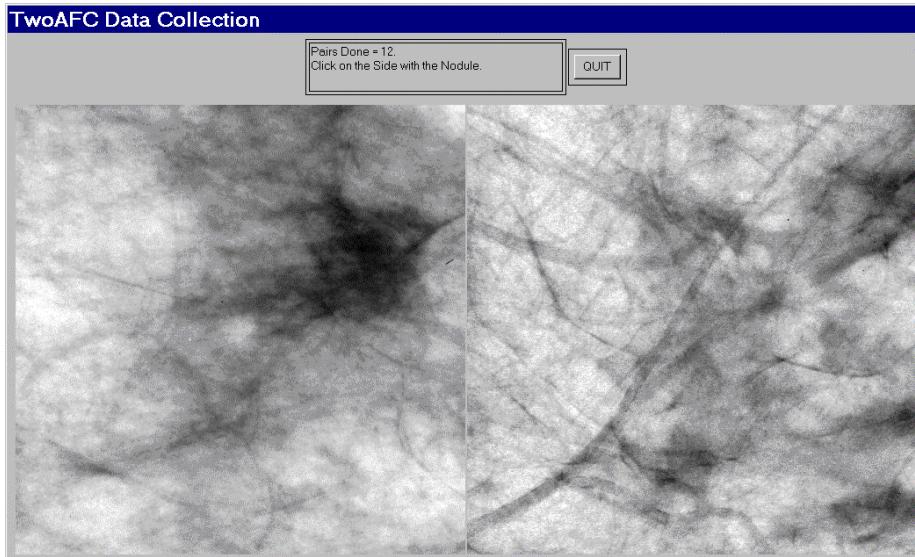


Figure 11.3: Example of image presentation in a 2AFC study.

## 11.10 Observer performance studies as laboratory simulations of clinical tasks

- Observer performance paradigms (ROC, FROC, LROC and ROI) should be regarded as experiments conducted in a laboratory (i.e., controlled) setting that are intended to be representative of the actual clinical task. They should not be confused with performance in a real “live” clinical setting: there is a known “laboratory effect” (Gur et al., 2008). For example, in the just cited study radiologists performed better during live clinical interpretations than they did later, on the same cases, in a laboratory ROC study. This is to be expected because there is more at stake during live interpretations: e.g., the patient’s health and the radiologist’s reputation, than during laboratory ROC studies. The claimed “laboratory effect” has caused some minor controversy. A paper (Soh et al., 2013) titled “Screening mammography: test set data can reasonably describe actual clinical reporting” argues against the laboratory effect.
- Real clinical interpretations happen every day in radiology departments all over the world. On the other hand, in the laboratory, the radiologist is asked to interpret the images “as if in a clinical setting” and render a “diagnosis”. The laboratory decisions have no clinical consequences, e.g., the radiologist will not be sued for mistakes and their laboratory study decisions will have no impact on the clinical management of the pa-

tients. [Usually laboratory ROC studies are conducted on retrospectively acquired images. Patients, whose images were used in an ROC study, have already been imaged in the clinic and decisions have already been made on how to manage them.]

- There is no guarantee that results of the laboratory study are directly applicable to clinical practice. Indeed there is an assumption that the laboratory study correlates with clinical performance. Strict equality is not required, simply that the performance in the laboratory is related monotonically to actual clinical performance. Monotonicity assures preservation of performance orderings, e.g., a radiologist has greater performance than another does or one modality is superior to another, regardless of how they are measured, in the laboratory or in the clinic. The correlation is taken to be an axiomatic truth by researchers, when in fact it is an assumption. To the extent that the participating radiologist brings his/her full clinical expertise to bear on each laboratory image interpretation, i.e., takes the laboratory study seriously, this assumption is likely to be valid.
- This title of this section provoked a strong response from a collaborator. To paraphrase him, "... *I think it is a pity in this book chapter you argue that these studies are simulations. I mean, the reason people perform these studies is because they believe in the results*".
- I also believe in observer performance studies. Distrust of the word "simulation" seems to be peculiar to this field. Simulations are widely used in "hard" sciences, e.g., they are used in astrophysics to determine conditions dating to  $10^{-31}$  seconds after the big bang. Simulations are not to be taken lightly. Conducting clinical studies is very difficult as there are many factors not under the researcher's control. Observer performance studies of the type described in this book are the closest that one can come to the "real thing" as they include key elements of the actual clinical task: the entire imaging system, radiologists (assuming the radiologist take these studies seriously in the sense of bringing their full expertise to bear on each image interpretation) and real clinical images. As such are expected to correlate with real "live" interpretations.

## 11.11 Discrete vs. continuous ratings: the Miller study

- There is controversy about the merits of discrete vs. continuous ratings (Rockette et al., 1992; Wagner et al., 2001). Since the late Prof. Charles E. Metz and the late Dr. Robert F. Wagner have both backed the latter (i.e., continuous or quasi-continuous ratings) new ROC study designs sometimes tend to follow their advice. I recommend a 6-point rating scale

as outlined in Fig. 11.2. This section provides the background for the recommendation.

- A widely cited (22,909 citations at the time of writing) 1954 paper by Miller (Miller, 1956) titled “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information” is relevant. It is a readable paper, freely downloadable in several languages ([www.musanim.com/miller1956/](http://www.musanim.com/miller1956/)). In my judgment, this paper has not received the attention it should have in the ROC community, and for this reason portions from it are reproduced below. [George Armitage Miller, February 3, 1920 – July 22, 2012, was one of the founders of the field of cognitive psychology.]
- Miller’s first objective was to comment on absolute judgments of unidimensional stimuli. Since all (univariate, i.e., single decision per case) ROC models assume a unidimensional decision variable, Miller’s work is highly relevant. He comments on two papers by Pollack (Pollack, 1952, 1953). Pollack asked listeners to identify tones by assigning numerals to them, analogous to a rating task described above. The tones differed in frequency, covering the range 100 to 8000 Hz in equal logarithmic steps. A tone was sounded and the listener responded by giving a numeral (i.e., a rating, with higher values corresponding to higher frequencies). After the listener had made his response, he was told the correct identification of the tone. When only two or three tones were used, the listeners never confused them. With four different tones, confusions were quite rare, but with five or more tones, confusions were frequent. With fourteen different tones, the listeners made many mistakes. Since it is so succinct, the entire content of the first (1952) paper by Pollack is reproduced below:
- “In contrast to the extremely acute sensitivity of a human listener to discriminate small differences in the frequency or intensity between two sounds is his relative inability to identify (and name) sounds presented individually. When the frequency of a single tone is varied in equal-logarithmic steps in the range between 100 cps and 8000 cps (and when the level of the tone is randomly adjusted to reduce loudness cues), the amount of information transferred is about 2.3 bits per stimulus presentation. This is equivalent to perfect identification among only 5 tones. The information transferred, under the conditions of measurement employed, is reasonably invariant under wide variations in stimulus conditions.”
- By “information” is meant (essentially) the number of levels, measured in bits (binary digits), thereby making it independent of the unit of measurement: 1 bit corresponds to a binary rating scale, 2 bits to a four-point rating scale and  $2^{2.3} = 4.9$ , i.e., about 5 ratings bins. Based on Pollack’s original unpublished data, Miller put an upper limit of 2.5 bits (corresponding to about 6 ratings bins) on the amount of information

that is transmitted by listeners who make absolute judgments of auditory pitch. The second paper (@ Pollack, 1953) by Pollack was related to: (1) the frequency range of tones; (2) the utilization of objective reference tones presented with the unknown tone; and (3) the “dimensionality”—the number of independently varying stimulus aspects. Little additional gain in information transmission was associated with the first factor; a moderate gain was associated with the second; and a relatively substantial gain was associated with the third (we return to the dimensionality issue below).

- As an interesting side-note, Miller states:

“Most people are surprised that the number is as small as six. Of course, there is evidence that a musically sophisticated person with absolute pitch can identify accurately any one of 50 or 60 different pitches. Fortunately, I do not have time to discuss these remarkable exceptions. I say it is fortunate because I do not know how to explain their superior performance. So I shall stick to the more pedestrian fact that most of us can identify about one out of only five or six pitches before we begin to get confused.

It is interesting to consider that psychologists have been using seven-point rating scales for a long time, on the intuitive basis that trying to rate into finer categories does not really add much to the usefulness of the ratings. Pollack’s results indicate that, at least for pitches, this intuition is fairly sound.

Next you can ask how reproducible this result is. Does it depend on the spacing of the tones or the various conditions of judgment? Pollack varied these conditions in a number of ways. The range of frequencies can be changed by a factor of about 20 without changing the amount of information transmitted more than a small percentage. Different groupings of the pitches decreased the transmission, but the loss was small. For example, if you can discriminate five high-pitched tones in one series and five low-pitched tones in another series, it is reasonable to expect that you could combine all ten into a single series and still tell them all apart without error. When you try it, however, it does not work. The channel capacity for pitch seems to be about six and that is the best you can do.”

- In contrast to the careful experiments conducted in the psychophysical context to elucidate this issue, I was unable to find a single study, in the medical imaging field, of the number of discrete rating levels that an observer can support. Instead, a recommendation has been made to acquire data on a quasi-continuous scale (Wagner et al., 2001).

- There is no question that for multidimensional data, as observed in the second study by Pollack (Pollack, 1953), the observer can support more than 7 ratings bins. To quote Miller:

“You may have noticed that I have been careful to say that this magical number seven applies to one-dimensional judgments. Everyday experience teaches us that we can identify accurately any one of several hundred faces, any one of several thousand words, any one of several thousand objects, etc. The story certainly would not be complete if we stopped at this point. We must have some understanding of why the one-dimensional variables we judge in the laboratory give results so far out of line with what we do constantly in our behavior outside the laboratory. A possible explanation lies in the number of independently variable attributes of the stimuli that are being judged. Objects, faces, words, and the like differ from one another in many ways, whereas the simple stimuli we have considered thus far differ from one another in only one respect.”

- In the medical imaging context, a trivial way to increase the number of ratings would be to color-code the images: red, green and blue; now one can assign a red image rated 3, a green image rated 2, etc., which would be meaningless unless the color encoded relevant diagnostic information. Another ability, quoted in the publication (Wagner et al., 2001) advocating continuous ratings is the ability to recognize faces, again a multidimensional categorization task, as noted by Miller. Also quoted as an argument for continuous ratings is the ability of computer aided detection schemes that calculate many features for each perceived lesion and combine them into a single probability of malignancy, which is on a highly precise floating point 0 to 1 scale, which can be countered by the fact that radiologists are not computers. Other arguments for greater number of bins: it cannot hurt and one should acquire the rating data at greater precision than the noise, especially if the radiologist is able to maintain the finer distinctions. I worry that radiologists who are willing to go along with greater precision are over-anxious to co-operate with the experimentalist. Expert radiologists will not modify their reading style and one should be suspicious when overzealous radiologists accede to an investigator's request to interpret images in a style that does not resemble the clinic. Radiologists, especially experts, do not like more than about four ratings. I once worked closely with a famous chest radiologist (the late Dr. Robert Fraser) who refused to use more than four ratings.
- Another reason given for using continuous ratings is it reduces instances of data degeneracy. Data is sometimes said to be degenerate if the curve-fitting algorithm, the binormal model and the proper binormal model, cannot fit it (in simple terms, the program crashes). This occurs, for

example, if there are no interior points on the ROC plot. Modifying radiologist behavior to accommodate the limitations of analytical methods seems to be inherently dubious. One could simply randomly add or subtract half an integer from the observed ratings, thereby making the rating scale more granular and reduce instances of degeneracy (this is actually done in some ROC software to overcome degeneracy issues). Another possibility is to use the empirical (trapezoidal) area under the ROC curve, which can always be calculated; there are no degeneracy problems with it. Actually, fitting methods now exist that are robust to data degeneracy, such as discussed in TBA Chapter 18 and Chapter 20, so this reason for acquiring continuous data no longer applies.

- The rating task involves a unidimensional scale and I see no way of getting around the basic channel-limitation noted by Miller and for this reason I recommend a 6 point scale, as in Fig. 11.2.
- On the other side of the controversy (Berbaum et al., 2002), a position that I agree with, it has been argued that given a large number of allowed ratings levels the cooperating observer essentially bins the data into a much smaller number of bins (e.g., 0, 20, 40, 60, 80, 100) and then adds a zero-mean noise term to appear to be “spreading out the ratings”. This ensures that the binormal model does not crash. However, if the intent is to get the observer to spread the ratings, so that the binormal model does not crash, a better approach is to use alternate models that do not crash and are, in fact, very robust with respect to degeneracy of the data. More on this later (see Chapters TBA CBM and RSM).

## 11.12 The BI-RADS ratings scale and ROC studies

It is desirable that the rating scale be relevant to the radiologists’ daily practice. This assures greater consistency – the fitting algorithms assume that the thresholds are held constant for the duration of the ROC study. Depending on the clinical task, a natural rating scale may already exist. For example, in 1992 the American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) to standardize mammography reporting<sup>36</sup>. There are six assessment categories: category 0 indicates need for additional imaging; category 1 is a negative (clearly non-diseased) interpretation; category 2 is a benign finding; category 3 is probably benign, with short-interval follow-up suggested; category 4 is a suspicious abnormality for which biopsy should be considered; category 5 is highly suggestive of malignancy and appropriate action should be taken. The 4th edition of the BI-RADS manual<sup>37</sup> divides category 4 into three subcategories 4A, 4B and 4C and adds category 6 for a proven malignancy. The 3-category may be further subdivided into “probably benign with a recommendation for normal or short-term follow-up” and a 3+ category, “probably benign

Table 11.3: The Barlow et al study: the ordering of the BI-RADS ratings in the first column correlates with cancer-rate in the last column.

	Total number of mammograms	Mammograms without breast cancer (percent)	Mammograms with breast cancer (percent)	Cancers per 1000 screening mammograms
1: Normal	356,030	355,734 (76.2)	296 (12.3)	0.83
2: Benign finding	56,614	56,533 (12.1)	81 (3.4)	1.43
3: Probably benign, recommend normal or short term follow up	8,692	8,627 (1.8)	65 (2.7)	7.48
3+: Probably benign, recommend immediate follow up	3,094	3,049 (0.7)	45 (1.9)	14.54
0: Need additional imaging evaluation	42,823	41,442 (8.9)	1,381 (57.5)	32.25
4: Suspicious finding, biopsy should be considered	2,022	1,687 (0.4)	335 (13.9)	165.68
5: Highly suggestive of malignancy	237	38 (0.0)	199 (8.3)	839.66

with a recommendation for immediate follow-up". Apart from categories 0 and 2, the categories form an ordered set with higher categories representing greater confidence in presence of cancer. How to handle the 0s and the 2s is the subject of some controversy, described next.

## 11.13 The controversy

Two large clinical studies have been reported in which BI-RADS category data were acquired for > 400,00 screening mammograms interpreted by many (124 in the 1st study) radiologists (Barlow et al., 2004; Fenton et al., 2007). The purpose of the first study was to relate radiologist characteristics to actual performance (e.g., does performance depend on reading volume – the number of cases interpreted per year), so it could be regarded as a more elaborate version of (Beam et al., 1996), described in Chapter 9. The purpose of the second study was to determine the effectiveness of computer-aided detection (CAD) in screening mammography.

The reported ROC analyses used the BIRADS assessments labels ordered as follows: 1 < 2 < 3 < 3+ < 0 < 4 < 5. The last column of Table 11.3 shows that with this ordering the numbers of cancer per 1000 patients increases monotonically. The CAD study is discussed later, for now the focus is on the adopted BIRADS scale ordering that is common to both studies and which has raised controversy (the controversy appears to be limited to observer performance study analysts).

The use of the BI-RADS ratings shown in Table 11.3 has been criticized (Jiang and Metz, 2010) in an editorial titled:

### BI-RADS Data Should Not Be Used to Estimate ROC Curves

Since BI-RADS is a clinical rating scheme widely used in mammography, the editorial, if correct, implies that ROC analysis of clinical mammography data is not possible. Since the BI-RADS scale was arrived at after considerable deliberation, inability to perform ROC analysis with it would strike at the root of clinical utility of the ROC method. The purpose of this section is to express the reasons why I have a different take on this controversy.

It is claimed in the editorial that the Barlow et al. study confuses cancer yield with confidence level and that BI-RADS categories 1 and 2 should not be separate entries of the confidence scale, because both indicate no suspicion for cancer.

I agree with the Barlow et al. suggested ordering of the “2s” as more likely to have cancer than the “1s”. A category-2 means the radiologist found something to report, and the location of the finding is part of the clinical report. Even if the radiologist believes the finding is definitely benign, there is a finite probability that a category-2 finding is cancer, as evident in the last column of Table 11.3 ( $1.43 > 0.83$ ). In contrast, there are no findings associated with a category-1 report. A paper (Hartmann et al., 2005) titled:

#### Benign breast disease and the risk of breast cancer

should convince any doubters that benign lesions do have a finite chance of cancer.

The problem with “where to put the 0s” arises only when one tries to analyze clinical BI-RADS data. In a laboratory study, the radiologist would not be given the category-0 option. In analyzing a clinical study it is incumbent on the study designer to justify the choice of the rating scale adopted. Showing that the proposed ordering agrees with the probability of cancer is justification – and in my opinion, given the very large sample size this was accomplished convincingly in the Barlow et al. study.

**Moreover, the last column of Table 11.3 suggests that any other ordering would violate an important principle, namely, optimal ordering is achieved when each case is rated according to its likelihood ratio (defined as the probability of the case being diseased divided by the probability of the case being non-diseased). The likelihood ratio is the “betting odds” of the case being diseased, which is expected to be monotonic with the empirical probability of the case being diseased, i.e., the last column of Table 11.3. Therefore, the ordering adopted in Table 11.3 is equivalent to adopting a likelihood ratio scale and any other ordering would not be monotonic with likelihood ratio.**

The likelihood ratio is described in more detail in the TBA Chapter 20, which describes ROC fitting methods that yield “proper” ROC curves, i.e., ones that

have monotonically decreasing slope as the operating point moves up the curve from (0,0) to (1,1) and therefore do not (inappropriately) cross the chance diagonal. Key to these fitting methods is adoption of a likelihood ratio scale to rank-order cases, instead of the ratings assumed by the unequal variance binormal model. The proper ROC fitting algorithm implemented in PROPROC software reorders confidence levels assumed by the binormal model, TBA Chapter 20, paragraph following Fig. 20.4. This is analogous to the reordering of the clinical ratings based on cancer rates assumed in Table 11.3. It is illogical to allow reordering of ratings in “blind” software but question the same when done in a principled way by a researcher. As expected, the modeled ROC curves in the Barlow publication, their Fig. 4, show no evidence of improper behavior. This is in contrast to a clinical study (about fifty thousands patients spread over 33 hospitals with each mammogram interpreted by two radiologists) using a non-BIRADS 7-point rating scale which yielded markedly improper ROC curves (Pisano et al., 2005) for the film modality when using ROC ratings (not BIRADS). This suggests that use of a non-clinical ratings scale for clinical studies, without independent confirmation of the ordering implied by the scale, is problematical.

The reader might be interested as to reason for the 0-ratings being more predictive of cancer than a 3+ rating, Table 11.3. In the clinic the zero rating implies, in effect, “defer decision, incomplete information, additional imaging necessary”. A zero rating could be due to technical problems with the images: e.g., improper positioning (e.g., missing breast tissue close to the chest wall) or incorrect imaging technique (improper selection of kilovoltage and/or tube charge), making it impossible to properly interpret the images. Since the images are part of the permanent patient record, there are both healthcare and legal reasons why the images need to be optimal. Incorrect technical factors are expected to occur randomly and therefore not predictive of cancer. However, if there is a suspicious finding and the image quality is sub-optimal, the radiologist may be unable to commit to a decision, they may seek additional imaging, perhaps better compression or a slightly different view angle to resolve the ambiguity. Such zero ratings are expected with suspicious findings, and therefore are expected to be predictive of cancer.

As an aside, the second paper (Fenton et al., 2007) using the ordering shown in Table 11.3 questioned the utility of CAD for breast cancer screening (this was ca. 2007). This paper was met with flurry of correspondence disputing the methodology (summarized above). The finding regarding utility of CAD has been validated by more recent studies, again with very large case and reader samples, showing that usage of CAD can actually be detrimental to patient outcome (Philpotts, 2009) and a call (Fenton, 2015) for ending insurance reimbursement for CAD.

## 11.14 Discussion

In this chapter the widely used ratings paradigm was described and illustrated with a sample dataset. The calculation of ROC operating points from this table was detailed. A formal notation was introduced to describe the counts in this table and the construction of operating points and an R example was given. I do not wish to leave the impression that the ratings paradigm is used only in medical imaging. In fact the historical reference (Macmillan and Creelman, 1991) to the two-question six-point scale in Fig. 11.2, namely Table 3.1 in the book by MacMillan and Creelman, was for a rating study on performance in recognizing odors. The early users of the ROC ratings paradigm were mostly experimental psychologists and psychophysicists interested in studying perception of signals, some in the auditory domain, and some in other sensory domains.

While it is possible to use the equal variance binormal model to obtain a measure of performance, the results depend upon the choice of operating point, and evidence was presented for the generally observed fact that most ROC ratings datasets are inconsistent with the equal variance binormal model. This indicates the need for an extended model, to be discussed in TBA Chapter 06.

The rating paradigm is a more efficient way of collecting the data compared to repeating the binary paradigm with instructions to cause the observer to adopt different fixed thresholds specific to each repetition. The rating paradigm is also more efficient than the 2AFC paradigm; more importantly, it is more clinically realistic.

Two controversial but important issues were addressed: the reason for my recommendation for adopting a discrete 6-point rating scale, and correct usage of clinical BIRADS ratings in ROC studies. When a clinical scale exists, the empirical disease occurrence rate associated with each rating should be used to order the ratings. Ignoring an existing clinical scale would be a disservice to the radiology community.

The next step is to describe a model for ratings data. Before doing that, it is necessary to introduce an empirical performance measure, namely the area under the empirical or trapezoidal ROC, which does not require any modeling.

## 11.15 References

# Chapter 12

## Empirical AUC

### 12.1 TBA How much finished

80%

### 12.2 Introduction

The ROC plot, introduced in Chapter 03, is defined as the plot of sensitivity (y-axis) vs. 1-specificity (x-axis). Equivalently, it is the plot of TPF (y-axis) vs. FPF (x-axis). An equal variance binormal model was introduced which allows an ROC plot to be fitted to a single observed operating point. In Chapter 04, the more commonly used ratings paradigm was introduced.

One of the reasons for fitting observed counts data, such as in Table 4.1 in Chapter 04, to a parametric model, is to derive analytical expressions for the separation parameter  $\mu$  of the model or the area AUC under the curve. Other figures of merit, such as the TPF at a specified FPF, or the partial area to the left of a specified FPF, can also be calculated from this model. Each figure of merit can serve as the basis for comparing two readers to determine which one is better. They have the advantage of being single values, as opposed to a pair of sensitivity-specificity values, thereby making it easier to unambiguously compare performances. Additionally, they often yield physical insight into the task, e.g., the separation parameter is the perceptual signal to noise corresponding to the diagnostic task.

It was shown, TBA Fig. 4.1 (A - B), that the equal variance binormal model did not describe a clinical dataset and that an unequal variance binormal model yielded a better visual fit. This turns out to be an almost universal finding. Before getting into the complexity of the unequal variance binormal model curve

Table 12.1: On the need for two indices to label cases in an ROC study.

N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11
D1	D2	D3	D4	D5	D6	D7				

fitting, it is appropriate to introduce a simpler empirical approach, which is very popular with some researchers. The New Oxford American Dictionary definition of “empirical” is: “based on, concerned with, or verifiable by observation or experience rather than theory or pure logic”. The method is also termed “non-parametric” as it does not involve any parametric assumptions (specifically normality assumptions). Notation is introduced for labeling individual cases that is used in subsequent chapters. An important theorem relating the empirical area under the ROC to a formal statistic, known as the Wilcoxon, is described. The importance of the theorem derives from its applications to non-parametric analysis of ROC data.

## 12.3 The empirical ROC plot

The empirical ROC plot is constructed by connecting adjacent observed operating points, including the trivial ones at (0,0) and (1,1), with straight lines. The trapezoidal area under this plot is a non-parametric figure of merit that is threshold independent. Since no parametric assumptions are involved, some prefer it to parametric methods, such as the one to be described in the next chapter. [In the context of AUC, the terms empirical, trapezoidal, or non-parametric all mean the same thing.]

### 12.3.1 Notation for cases

As in §3.5, cases are indexed by  $k_t t$  where  $t$  indicates the truth-status at the case (i.e., patient) level, with  $t = 1$  for non diseased cases and  $t = 2$  for diseased cases. Index  $k_1$  ranges from one to  $K_1$  for non-diseased cases and  $k_2$  ranges from one to  $K_2$  for diseased cases, where  $K_1$  and  $K_2$  are the total number of non-diseased and diseased cases, respectively. In Table 5.1, each case is represented as a shaded box, lighter shading for non-diseased cases and darker shading for diseased cases. There are 11 non-diseased cases, labeled N1 – N11, in the upper row of boxes and there are seven diseased cases, labeled D1 – D7, in the lower row of boxes.

TBA In 12.1 the upper row shows 11 non-diseased cases, labeled N1 – N11, while the lower row shows seven diseased cases, labeled D1 – D7. To address any case one needs two indices: the row number  $t$  and the column number  $k_t t$ . Since in general the column number depends on the value of  $t$ , one needs two

indices to specify the column index. To address a case one needs two indices; the first index is the row number  $t$  and the second index is the column number  $k_t$ . Since the total number of columns depends on the row number, the column index has to be  $t$ -dependent, i.e.,  $k_t$ , denoting the column index  $k_t$  of a case with truth index  $t$ . Alternative notation in more commonly usage uses a single index  $k$  to label the cases. It reserves the first  $K_1$  positions for non-diseased cases and the rest for diseased cases: e.g.,  $k = 3$  corresponds to the third non-diseased case,  $k = K_1 + 5$  corresponds to the fifth diseased case, etc. Because it extends more easily to more complex data structures, e.g., FROC, I prefer the two-index notation.

### 12.3.2 An empirical operating point

Let  $z_{k_t}$  represent the z-sample of case  $k_t$ . For a given reporting threshold  $\zeta$ , and assuming a positive-directed rating scale (i.e., higher values correspond to greater confidence in presence of disease), empirical false positive fraction  $FPF(\zeta)$  and empirical true positive fraction  $TPF(\zeta)$  are defined by:

$$\left. \begin{aligned} FPF(\zeta) &= \frac{1}{K_1} \sum_{k_1=1}^{K_1} I(z_{k_1} \geq \zeta) \\ TPF(\zeta) &= \frac{1}{K_2} \sum_{k_2=1}^{K_2} I(z_{k_2} \geq \zeta) \end{aligned} \right\} \quad (12.1)$$

Here  $I(x)$  is the indicator function that equals one if  $x$  is true and is zero otherwise.

In Eqn. (12.1) the indicator functions act as counters, effectively counting instances where the z-sample of a case equals or exceeds  $\zeta$ , and division by the appropriate denominator yields the desired left hand sides of these equations. The operating point  $O(\zeta)$  corresponding to threshold  $\zeta$  is defined by:

$$O(\zeta) = (FPF(\zeta), TPF(\zeta)) \quad (12.2)$$

The essential difference between Eqn. (12.1) and Eqn. (10.18) is that the former is non-parametric while the latter is parametric. In TBA Chapter 03 analytical (or parametric, i.e., model parameter dependent) operating points were obtained. In contrast, here one uses the observed ratings to calculate the empirical operating point.

## 12.4 Empirical operating points from ratings data

Consider a ratings ROC study with  $R$  bins. Describing an R-rating empirical ROC plot requires  $R - 1$  ordered empirical thresholds, see Eqn. (11.3).

The operating point  $O(\zeta_r)$  is given by:

$$O(\zeta_r) = (FPF(\zeta_r), TPF(\zeta_r)) \quad (12.3)$$

Its coordinates are defined by:

$$\left. \begin{aligned} FPF_r &\equiv FPF(\zeta_r) = \frac{1}{K_1} \sum_{k_1=1}^{K_1} I(z_{k_11} \geq \zeta_r) \\ TPF_r &\equiv TPF(\zeta_r) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} I(z_{k_22} \geq \zeta_r) \end{aligned} \right\} \quad (12.4)$$

For example,

$$\left. \begin{aligned} FPF_4 &\equiv FPF(\zeta_4) = \frac{1}{K_1} \sum_{k_1=1}^{K_1} I(z_{k_11} \geq \zeta_4) \\ TPF_4 &\equiv TPF(\zeta_4) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} I(z_{k_22} \geq \zeta_4) \\ O_4 &\equiv (FPF_4, TPF_4) = (0.017, 0.44) \end{aligned} \right\} \quad (12.5)$$

In Table 11.1 a sample clinical ratings data set was introduced. Shown below is a partial code listing of mainEmpRocPlot.R showing implementation of Eqn. (5.7). Except for the last statement, the plotting part of the code is suppressed.

```
K1 <- 60
K2 <- 50
FPF <- c(0, cumsum(rev(c(30, 19, 8, 2, 1))) / K1)
TPF <- c(0, cumsum(rev(c(5, 6, 5, 12, 22))) / K2)

ROCOp <- data.frame(FPF = FPF, TPF = TPF)
ROCplot <- ggplot(
```

```

data = ROCOp,
mapping = aes(x = FPF, y = TPF)) +
geom_line(size = 1) +
geom_point(size = 4) +
theme_bw() +
theme(panel.grid.major = element_blank(),
      panel.grid.minor = element_blank(),
      panel.border = element_rect(color = "black"),
      axis.text = element_text(size = 15),
      axis.title = element_text(size = 20)) +
scale_x_continuous(
  expand = c(0, 0),
  breaks = c(0.25, 0.5, 0.75, 1)) +
scale_y_continuous(
  expand = c(0, 0), breaks = c(0.25, 0.5, 0.75, 1)) +
coord_cartesian(ylim = c(0,1), x = c(0,1)) +
annotation_custom(
  grob = textGrob(bquote(italic("0"))),
  gp = gpar(fontsize = 22)),
  xmin = -0.03, xmax = -0.03,
  ymin = -0.03, ymax = -0.03) +
annotation_custom(
  grob = textGrob(bquote(italic(0[4]))),
  gp = gpar(fontsize = 22)),
  xmin = 0.06, xmax = 0.06,
  ymin = 0.40, ymax = 0.40) +
annotation_custom(
  grob = textGrob(bquote(italic(0[3]))),
  gp = gpar(fontsize = 22)),
  xmin = 0.10, xmax = 0.10,
  ymin = 0.64, ymax = 0.64) +
annotation_custom(
  grob = textGrob(bquote(italic(0[2]))),
  gp = gpar(fontsize = 22)),
  xmin = 0.16, xmax = 0.16,
  ymin = 0.83, ymax = 0.83) +
annotation_custom(
  grob = textGrob(bquote(italic(0[1]))),
  gp = gpar(fontsize = 22)),
  xmin = 0.49, xmax = 0.49,
  ymin = 0.94, ymax = 0.94)

p <- ggplotGrob(ROCPlot)
p$layout$clip[p$layout$name == "panel"] <- "off"
grid.draw(p)

```

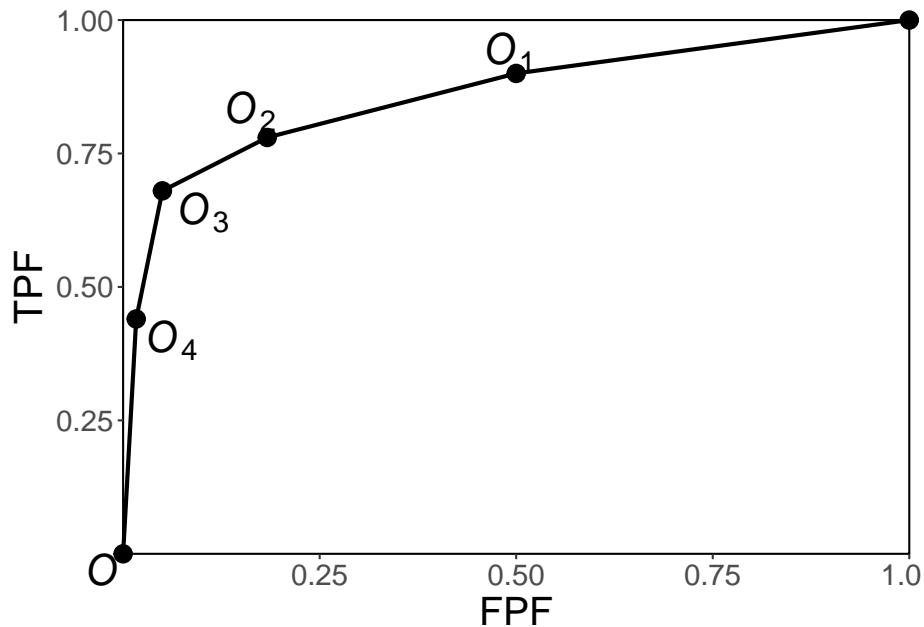


Figure 12.1: Empirical ROC plot for the data in Table 4.1. By convention the operating points are numbered starting with the uppermost non-trivial one and working down the plot and the trivial operating points  $(0,0)$  and  $(1,1)$  are not shown.

The function `cumsum()` is used to calculate the cumulative sum. The `rev()` function reverses the order of the array supplied as its argument. The reader should use the debugging techniques (basically copy and paste parts of the code to the Console window and hit enter) to understand how this code implements Eqn. (12.4).

Fig. 12.1 is the empirical ROC plot. It illustrates the convention used to label the operating points introduced in TBA §4.3 is, i.e.,  $O_1$  is the uppermost non-trivial point, and the subscripts increment by unity as one moves down the plot. By convention, not shown are the trivial operating points  $O_0 \equiv (FPF_0, TPF_0) = (1, 1)$  and  $O_R \equiv (FPF_R, TPF_R) = (0, 0)$ , where  $R = 5$ .

## 12.5 AUC under the empirical ROC plot

Fig. 12.2 shows the empirical plot for the data in Table 4.1. The area under the curve (AUC) is the shaded area. By dropping imaginary vertical lines from the non-trivial operating points onto the x-axis, the shaded area is seen to be the sum of one triangular shaped area and four trapezoids. One may be tempted to

write equations to calculate the total area using elementary algebra, but that would be unproductive. There is a theorem (see below) that the empirical area is exactly equal to a particular statistic known as the Mann-Whitney-Wilcoxon statistic (Wilcoxon, 1945; Mann and Whitney, 1947), which, in this book, is abbreviated to the Wilcoxon statistic. Calculating this statistic is much simpler than calculating and summing the areas of the triangle and trapezoids, or doing planimetry.

```
RocDataTable = array(dim = c(2,4))
RocDataTable[1,] <- c(30,19,8,3)
RocDataTable[2,] <- c(5,11,12,22)

ret <- RocOperatingPointsFromRatingsTable(
  RocDataTable[1,],
  RocDataTable[2,] )
FPF <- ret$FPF
TPF <- ret$TPF

ROC_Points <- data.frame(FPF = FPF, TPF = TPF)
# add the trivial points
ROC_Points <- rbind(
  c(0, 0),
  ROC_Points, c(1, 1))

shade <- data.frame(
  FPF = c(ROC_Points$FPF, 1),
  TPF = c(ROC_Points$TPF, 0))

p <- ggplot(ROC_Points,
            aes(x = FPF, y = TPF) ) +
  geom_polygon(data = shade, fill = 'grey') +
  geom_line(size = 1) +
  geom_point(size = 4) +
  theme_bw() +
  theme(
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank()) +
  labs(x = expression(FPF)) +
  labs(y = expression(TPF)) +
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0)) +
  coord_cartesian(ylim = c(0,1), x = c(0,1))
print(p)
```

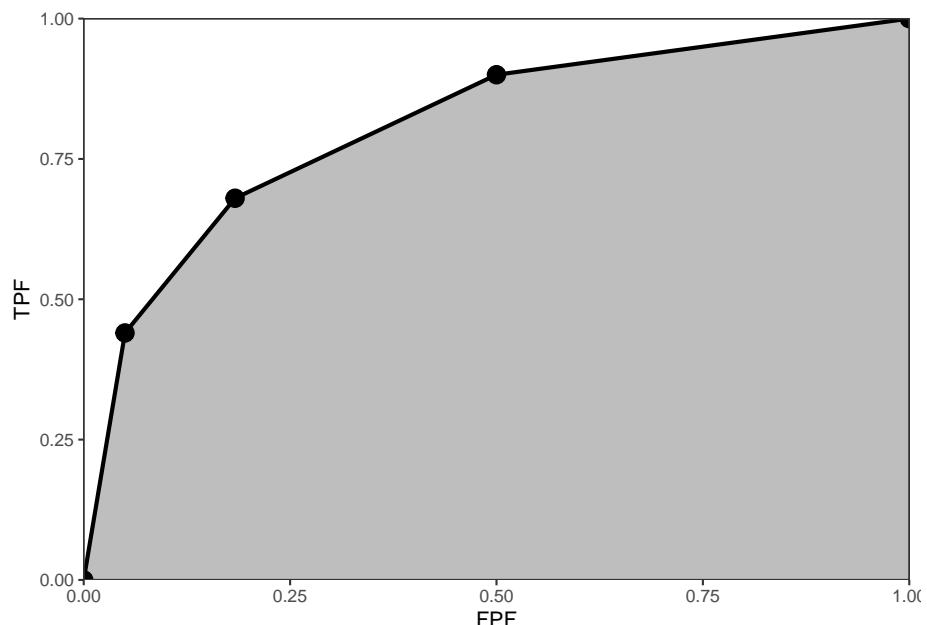


Figure 12.2: The empirical ROC plot corresponding to Table 4.1; the shaded area is the area AUC under this plot, a widely used figure of merit in non-parametric ROC analysis.

## 12.6 The Wilcoxon statistic

A statistic is any value calculated from observed data. The Wilcoxon statistic is defined in terms of the ratings, by:

$$W = \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \psi(z_{k_1 1}, z_{k_2 2}) \quad (12.6)$$

The function  $\psi(x, y)$  is defined by:

$$\begin{aligned} \psi(x, y) &= 1 & x < y \\ \psi(x, y) &= 0.5 & x = y \\ \psi(x, y) &= 0 & x > y \end{aligned} \quad (12.7)$$

The function  $\psi(x, y)$  is sometimes called the kernel function. It is unity if the diseased case is rated higher, 0.5 if the two are rated the same and zero otherwise. Each evaluation of the kernel function results from a comparison of a case from the non-diseased set with one from the diseased set. In Eqn. (12.6) the two summations and division by the total number of comparisons yields the observed, i.e., empirical, probability that diseased cases are rated higher than non-diseased ones. Since it is a probability, it can range from zero to one. However, if the observer has any discrimination ability at all, one expects diseased cases to be rated equal or greater than non-diseased ones, so in practice one expects  $0.5 \leq W \leq 1$ . The limit 0.5 corresponds to a guessing observer, whose operating point lies on the chance diagonal of the ROC plot.

## 12.7 Bamber's Equivalence theorem

The Wilcoxon statistic  $W$  equals the area  $AUC$  under the empirical ROC plot:

$$W = AUC \quad (12.8)$$

Numerical illustration: While hardly a proof, as an illustration of the theorem it is helpful to calculate the sum on the right hand side of Eqn. (12.6) and compare it to direct integration of the area under the empirical ROC curve (i.e., adding the area of a triangle and several trapezoids). The function is called `trapz(x,y)`, see below. It takes two array arguments,  $x$  and  $y$ , where in the current case  $x$  is  $FPF$  and  $y$  is  $TPF$ . One has to be careful to include the end-points as otherwise the area will be underestimated. The Wilcoxon  $W$  and the numerical estimate of the empirical area  $AUC$  are implemented in the following code.

```

trapz = function(x, y)
{ #### computes the integral of y with respect to x using trapezoidal integration.
  idx = 2:length(x)
  return (as.double( (x[idx] - x[idx-1]) %*% (y[idx] + y[idx-1])) / 2)
}

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

RocOperatingPoints <- function( K1, K2 ) {

  nOpPts <- length(K1) - 1 # number of op points
  FPF <- array(0, dim = nOpPts)
  TPF <- array(0, dim = nOpPts)

  for (r in (nOpPts+1):2) {
    FPF[r-1] <- sum(K1[r:(nOpPts+1)]) / sum(K1)
    TPF[r-1] <- sum(K2[r:(nOpPts+1)]) / sum(K2)
  }
  FPF <- rev(FPF)
  TPF <- rev(TPF)

  return( list(
    FPF = FPF,
    TPF = TPF
  ) )
}

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

zk1 <- rep(1:length(RocCountsTable[1,]),RocCountsTable[1,])#convert frequency table to
zk2 <- rep(1:length(RocCountsTable[2,]),RocCountsTable[2,])#do:

```

```
w <- Wilcoxon (zk1, zk2)
cat("The wilcoxon statistic is = ", w, "\n")
#> The wilcoxon statistic is = 0.8606667
ret <- RocOperatingPoints(RocCountsTable[1,], RocCountsTable[2,])
FPF <- ret$FPF; FPF <- c(0, FPF, 1)
TPF <- ret$TPF; TPF <- c(0, TPF, 1)
AUC <- trapz(FPF, TPF) # trapezoidal integration
cat("direct integration yields AUC = ", AUC, "\n")
#> direct integration yields AUC = 0.8606667
```

Note the equality of the two estimates.

The following proof is adapted from (Bamber, 1975) and while it may appear to be restricted to discrete ratings, the result is in fact quite general, i.e., it is applicable even if the ratings are acquired on a continuous scale. The reason is that in an R-rating ROC study the observed z-samples or ratings take on integer values, 1 through R. If R is large enough, ordering information present in the continuous data is not lost upon binning. In the following it is helpful to keep in mind that one is dealing with discrete distributions of the ratings, described by probability mass functions as opposed to probability density functions, e.g.,  $P(Z_2 = \zeta_i)$  is not zero, as would be the case for continuous ratings. The proof is illustrated with Fig. 12.3.

The abscissa of the operating point  $i$  is  $P(Z_1 \geq \zeta_i)$  and the corresponding ordinate is  $P(Z_2 \geq \zeta_i)$ . Here  $Z_1$  is a random sample from a non-diseased case and  $Z_2$  is a random sample from a diseased case. The shaded trapezoid defined by drawing horizontal lines from operating points  $i$  (upper) and  $i+1$  (lower) to the right edge of the ROC plot, Fig. 12.3, has height:

$$P(Z_2 \geq \zeta_i) - P(Z_2 \geq \zeta_{i+1}) = P(Z_2 = \zeta_i) \quad (12.9)$$

The validity of this equation can perhaps be more easily seen when the first term is written in the form:

$$P(Z_2 \geq \zeta_i) = P(Z_2 = \zeta_i) + P(Z_2 \geq \zeta_{i+1}) \quad (12.10)$$

The lengths of the top and bottom edges of the trapezoid are, respectively:

$$1 - P(Z_1 \geq \zeta_i) = P(Z_1 < \zeta_i) \quad (12.11)$$

and

$$1 - P(Z_1 \geq \zeta_{i+1}) = P(Z_1 < \zeta_{i+1}) \quad (12.12)$$

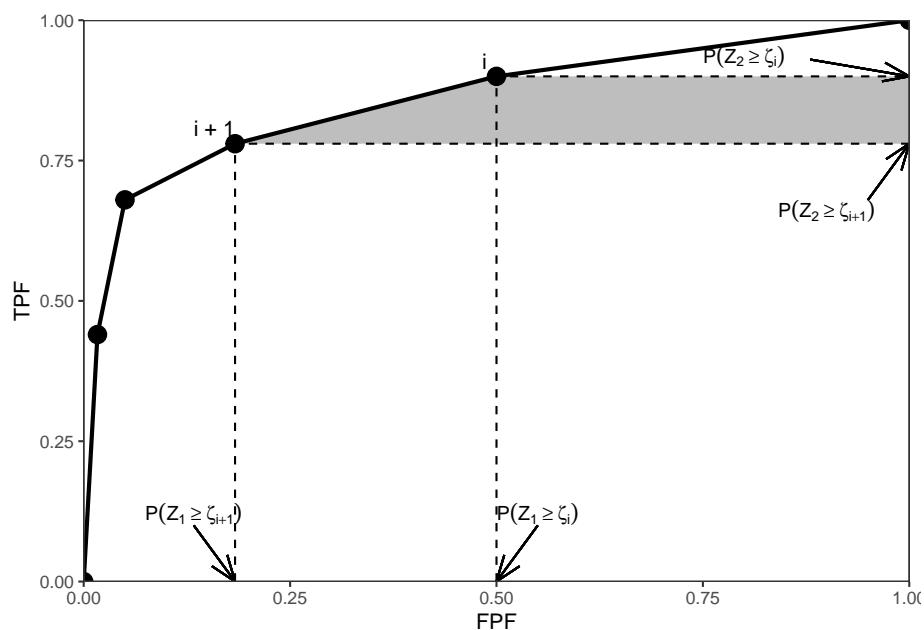


Figure 12.3: Illustration of the derivation of Bamber's equivalence theorem. Shows an empirical ROC plot for  $R = 5$ ; the shaded area is due to points labeled  $i$  and  $i + 1$ .

The area  $A_i$  of the shaded trapezoid in Fig. 12.3 is (the steps are shown explicitly):

$$\left. \begin{aligned} A_i &= \frac{1}{2}P(Z_2 = \zeta_i)[P(Z_1 < \zeta_i) + P(Z_1 < \zeta_{i+1})] \\ A_i &= P(Z_2 = \zeta_i)\left[\frac{1}{2}P(Z_1 < \zeta_i) + \frac{1}{2}(P(Z_1 = \zeta_i) + P(Z_1 < \zeta_i))\right] \\ A_i &= P(Z_2 = \zeta_i)\left[\frac{1}{2}P(Z_1 = \zeta_i) + P(Z_1 < \zeta_i)\right] \end{aligned} \right\} \quad (12.13)$$

Summing over all values of  $i$ , one gets for the total area under the empirical ROC plot:

$$\left. \begin{aligned} AUC &= \sum_{i=0}^{R-1} A_i \\ &= \frac{1}{2} \sum_{i=0}^{R-1} P(Z_2 = \zeta_i) P(Z_1 = \zeta_i) + \sum_{i=0}^{R-1} P(Z_2 = \zeta_i) P(Z_1 < \zeta_i) \end{aligned} \right\} \quad (12.14)$$

It is shown in the Appendix that the term  $A_0$  corresponds to the triangle at the upper right corner of Fig. 12.3, and the term  $A_4$  corresponds to the horizontal trapezoid defined by the lowest non-trivial operating point.

Eqn. (12.14) can be restated as:

$$AUC = \frac{1}{2}P(Z_1 = Z_2) + P(Z_1 < Z_2) \quad (12.15)$$

The Wilcoxon statistic was defined in Eqn. (12.6). It can be seen that the comparisons implied by the summations and the weighting implied by the kernel function are estimating the two probabilities in the expression for in Eqn. (12.15). Therefore,  $AUC = W$ .

## 12.8 Importance of Bamber's theorem

The equivalence theorem is the starting point for all non-parametric methods of analyzing ROC plots, e.g., (Hanley and Hajian-Tilaki, 1997; DeLong et al., 1988). Prior to Bamber's work one knew how to plot an empirical operating characteristic and how to calculate the Wilcoxon statistic, but their equality had not been analytically proven. This was Bamber's essential contribution. In the absence of this theorem, the Wilcoxon statistic would be "just another statistic" in the context of ROC analysis. The theorem is so important that a major paper

appeared in Radiology (Hanley and McNeil, 1982) devoted to the equivalence. The title of this paper was “The meaning and use of the area under a receiver operating characteristic (ROC) curve”. The equivalence theorem literally gives meaning to the empirical area under the ROC.

## 12.9 Discussion / Summary

In this chapter, a simple method for estimating the area under the ROC plot has been described. The empirical AUC is a non-parametric measure of performance. Its simplicity and clear physical interpretation as the AUC under the empirical ROC (not fitted, not true) has spurred much theoretical development. These include the De Long et al method for estimating the variance of AUC of a single ROC empirical curve, and comparing pairs of ROC empirical curves<sup>5</sup>. Bamber’s theorem, namely the equivalence between the empirical AUC and the Wilcoxon statistic has been derived and demonstrated.

Since the empirical AUC always yields a number, the researcher could be unaware about unusual behavior of the empirical ROC curve, so it is always a good idea to plot the data and look for evidence of large extrapolations. An example would be data points clustered at low FPF values, which imply a large AUC contribution, unsupported by intermediate operating points, from the line connecting the uppermost non-trivial operating point to (1,1).

## 12.10 Appendix 5.A: Details of Wilcoxon theorem

### 12.10.1 Upper triangle

For  $i = 0$ , Eqn. (12.13) implies (since the lowest empirical threshold is unity, the lowest allowed rating, and there are no cases rated less than one):

$$\left. \begin{aligned} A_0 &= P(Z_2 = 1) \left[ \frac{1}{2} P(Z_1 = 1) + P(Z_1 < 1) \right] \\ A_0 &= \frac{1}{2} P(Z_1 = 1) P(Z_2 = 1) \end{aligned} \right\} \quad (12.16)$$

The base of the triangle is:

$$1 - P(Z_1 \geq 2) = P(Z_1 < 2) = P(Z_1 = 1) \quad (12.17)$$

The height of the triangle is:

$$1 - P(Z_2 \geq 2) = P(Z_2 < 2) = P(Z_2 = 1) \quad (12.18)$$

Q.E.D.

### 12.10.2 Lowest trapezoid

For  $i = 4$ , Eqn. (12.13) implies:

$$\left. \begin{aligned} A_4 &= P(Z_2 = 5) \left[ \frac{1}{2}P(Z_1 = 5) + P(Z_1 < 5) \right] \\ A_4 &= \frac{1}{2}P(Z_2 = 5) [P(Z_1 = 5) + 2P(Z_1 < 5)] \\ A_4 &= \frac{1}{2}P(Z_2 = 5) [P(Z_1 = 5) + P(Z_1 < 5) + P(Z_1 < 5)] \\ A_4 &= \frac{1}{2}P(Z_2 = 5) [1 + P(Z_1 < 5)] \end{aligned} \right\} \quad (12.19)$$

The upper side of the trapezoid is

$$1 - P(Z_1 \geq 5) = P(Z_1 < 5) \quad (12.20)$$

The lower side is unity. The average of the two sides is:

$$\frac{1 + P(Z_1 < 5)}{2} \quad (12.21)$$

The height is:

$$P(Z_2 \geq 5) = P(Z_2 = 5) \quad (12.22)$$

Multiplication of the last two expressions yields  $A_4$ .

## 12.11 References



# Chapter 13

## Binormal model

### 13.1 TBA How much finished

70%

### 13.2 TBA Introduction

The equal variance binormal model was described in TBA Chapter 02. The ratings method of acquiring ROC data and calculation of operating points was discussed in TBA Chapter 04. It was shown, TBA Fig. 11.1, that for a clinical dataset the unequal-variance binormal model visually fitted the data better than the equal-variance binormal model, although how the unequal variance fit was obtained was not discussed. This chapter deals with details of the unequal-variance binormal model, often abbreviated to **binormal model**, establishes necessary notation, and derives expressions for sensitivity, specificity and the area under the predicted ROC curve).

The binormal model describes univariate datasets, in which there is *one ROC rating per case*, as in a single observer interpreting cases, one at a time, in a single modality. By convention the qualifier “univariate” is often omitted. In TBA Chapter 21 a bivariate model will be described where each case yields two ratings, as in a single observer interpreting cases in two modalities, or the homologous problem of two observers interpreting cases in a single modality.

The main aim of this chapter is to demystify statistical curve fitting. With the passing of Dorfman, Metz and Swensson, parametric modeling is being neglected. Researchers are instead focusing on non-parametric analysis using the empirical AUC. While useful and practical, empirical AUC yields almost no insight into what is limiting performance. Taking the mystery out of curve fitting

will allow the reader to appreciate later chapters that describe more complex fitting methods, which yield important insights into factors limiting performance.

Here is the organization of this chapter. It starts with a description of the binormal model and how it accommodates data binning. An important point, on which there is much confusion, on the invariance of the binormal model to arbitrary monotone transformations of the ratings is explicated with an example. Expressions for sensitivity and specificity are derived. Two notations used to characterize the binormal model are explained. Expressions for the pdfs of the binormal model are derived. A simple linear fitting method is illustrated: this used to be the only recourse a researcher had before Dorfman and Alf's seminal publication (Dorfman and Alf, 1969). The maximum likelihood method for estimating parameters of the binormal model is detailed. Validation of the fitting method is described, i.e., how can one be confident that the fitting method, which makes normality and other assumptions, is valid for a dataset arising from an unknown distribution. The Appendix has a detailed derivation, originally published in a terse paper (Thompson and Zucchini, 1989) on the partial-area under the ROC curve. The partial-area is defined by the area under the binormal ROC curve from  $FPF = 0$  to  $FPF = c$ , where  $0 \leq c \leq 1$ . As a special case  $c = 1$  yields the total area under the binormal ROC.

### 13.3 Binormal model

#### 13.3.1 The basic model

The unequal-variance binormal model (henceforth abbreviated to binormal model; when I mean equal variances, it will be made explicit) is defined by (capital letters indicate random variables and their lower-case counterparts are realized values):

$$Z_{k_t t} \sim N(\mu_t, \sigma_t^2); t = 1, 2 \quad (13.1)$$

where

$$\left. \begin{array}{l} \mu_1 = 0 \\ \mu_2 = \mu \\ \sigma_1^2 = 1 \\ \sigma_2^2 = \sigma^2 \end{array} \right\} \quad (13.2)$$

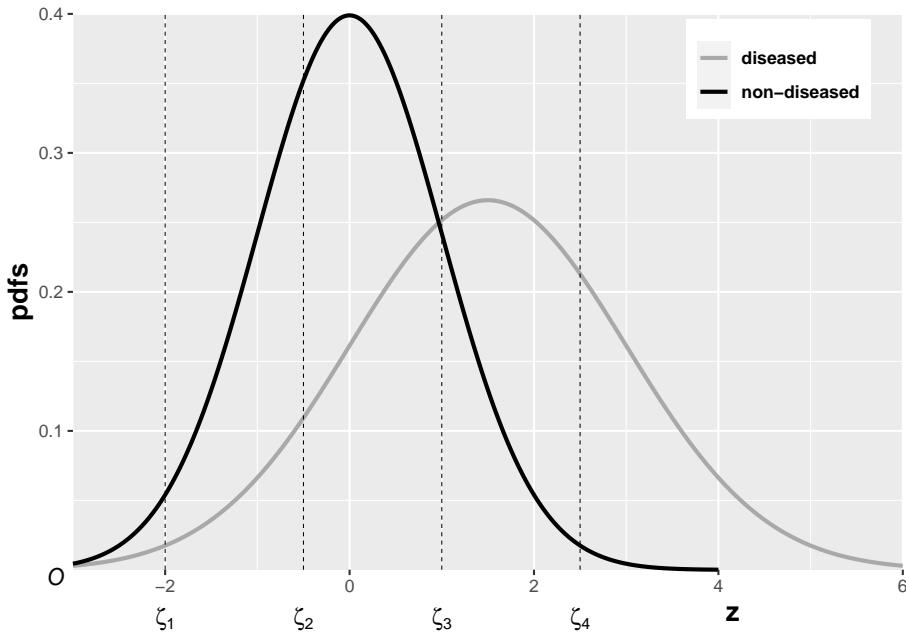
Eqn. (13.1) states that the z-samples for non-diseased cases are distributed as a  $N(0, 1)$  distribution, i.e., the unit normal distribution, while the z-samples for

diseased cases are distributed as a  $N(\mu, \sigma^2)$  distribution, i.e., a normal distribution with mean  $\mu$  and variance  $\sigma^2$ . This is a 2-parameter model of the z-samples, not counting additional threshold parameters needed for data binning.<sup>1</sup>

### 13.3.2 Additional parameters for binned data

In an R-rating ROC study the observed ratings  $r$  take on integer values, 1 through  $R$ , it being understood that higher ratings correspond to greater confidence for disease. Defining dummy cutoffs  $\zeta_0 = -\infty$  and  $\zeta_R = +\infty$ , the binning rule for a case with realized z-sample  $z$  is (Chapter 11, Eqn. (11.2)):

$$\text{if } (\zeta_{r-1} \leq z \leq \zeta_r) \Rightarrow \text{rating} = r \quad (13.3)$$



<sup>1</sup>A more complicated version of this model allows the mean of the non-diseased distribution to be non-zero and its variance different from unity. The 4-parameter model is no more general than the 2-parameter model. The reason is that one is free to transform the decision variable, and associated thresholds, by applying arbitrary monotonic increasing function transformation, which do not change the ordering of the ratings and hence do not change the ROC curve. So if the mean of the noise distribution were non-zero, subtracting this value from all Z-samples would shift the effective mean of the non-diseased distribution to zero (the shifted Z-values are monotonically related to the original values) and the mean of the shifted diseased distribution becomes  $\mu_2 - \mu_1$ . Next, one scales or divides (division by a positive number is also a monotonic transformation) all the Z-samples by  $\sigma_1$ , resulting in the scaled non-diseased distribution having unit variance, and the scaled diseased distribution has mean  $\frac{\mu_2 - \mu_1}{\sigma_1}$  and variance  $(\frac{\sigma_2}{\sigma_1})^2$ . Therefore, if one starts with 4 parameters then one can, by simple shifting and scaling operations, reduce the model to 2 parameters, as in Eqn. (13.1). [The author has seen a publication on Bayesian ROC estimation using the four-parameter model.]

In the unequal-variance binormal model, the variance  $\sigma^2$  of the z-samples for diseased cases is allowed to be different from unity. Most ROC datasets are consistent with  $\sigma > 1$ . The above figure, generated with  $\mu = 1.5, \sigma = 1.5, \zeta_1 = -2, \zeta_2 = -0.5, \zeta_3 = 1, \zeta_4 = 2.5$ , illustrates how realized z-samples are converted to ratings, i.e., application of the binning rule (13.3). For example, a case with z-sample equal to -2.5 would be rated “1”, and one with z-sample equal to -1 would be rated “2”, cases with z-samples greater than 2.5 would be rated “5”, etc.

### 13.3.3 Sensitivity and specificity

Let  $Z_t$  denote the random z-sample for truth state  $t$  ( $t = 1$  for non-diseased and  $t = 2$  for diseased cases). Since the distribution of z-samples from disease-free cases is  $N(0, 1)$ , the expression for specificity, Chapter “Modeling Binary Paradigm”, Eqn. 3.13, applies. It is reproduced below:

$$Sp(\zeta) = P(Z_1 < \zeta) = \Phi(\zeta) \quad (13.4)$$

To obtain an expression for sensitivity, consider that for truth state  $t = 2$ , the random variable  $\frac{Z_2 - \mu}{\sigma}$  is distributed as  $N(0, 1)$ :

$$\frac{Z_2 - \mu}{\sigma} \sim N(0, 1)$$

Sensitivity is  $P(Z_2 > \zeta)$ , which implies, because  $\sigma$  is positive (subtract  $\mu$  from both sides of the “greater than” symbol and divide by  $\sigma$ ):

$$Se(\zeta|\mu, \sigma) = P(Z_2 > \zeta) = P\left(\frac{Z_2 - \mu}{\sigma} > \frac{\zeta - \mu}{\sigma}\right) \quad (13.5)$$

The right-hand-side can be rewritten as follows:

$$Se(\zeta|\mu, \sigma) = 1 - P\left(\frac{Z_2 - \mu}{\sigma} \leq \frac{\zeta - \mu}{\sigma}\right) = 1 - \Phi\left(\frac{\zeta - \mu}{\sigma}\right) = \Phi\left(\frac{\mu - \zeta}{\sigma}\right)$$

Summarizing, the formulae for the specificity and sensitivity for the binormal model are:

$$Sp(\zeta) = \Phi(\zeta) \quad Se(\zeta|\mu, \sigma) = \Phi\left(\frac{\mu - \zeta}{\sigma}\right) \quad (13.6)$$

The coordinates of the operating point defined by  $\zeta$  are given by:

$$FPF(\zeta) = 1 - Sp(\zeta) = 1 - \Phi(\zeta) = \Phi(-\zeta) \quad (13.7)$$

$$TPF(\zeta|\mu, \sigma) = \Phi\left(\frac{\mu - \zeta}{\sigma}\right) \quad (13.8)$$

These expressions allow calculation of the operating point for any  $\zeta$ . An equation for a curve is usually expressed as  $y = f(x)$ . An expression of this form for the ROC curve, i.e., the y-coordinate (TPF) expressed as a function of the x-coordinate (FPF), follows upon inversion of the expression for FPF, Eqn. (13.7):

$$\zeta = -\Phi^{-1}(FPF) \quad (13.9)$$

Substitution of Eqn. (13.9) in Eqn. (13.8) yields:

$$TPF = \Phi\left(\frac{\mu + \Phi^{-1}(FPF)}{\sigma}\right) \quad (13.10)$$

This equation gives the dependence of TPF on FPF, i.e., the equation for the ROC curve. It will be put into standard notation next.

### 13.3.4 Binormal model in conventional notation

The following notation is widely used in the literature:

$$a = \frac{\mu}{\sigma}; b = \frac{1}{\sigma} \quad (13.11)$$

The reason for the  $(a, b)$  instead of the  $(\mu, \sigma)$  notation is that Dorfman and Alf assumed, in their seminal paper (Dorfman and Alf, 1969), that the diseased distribution (signal distribution in signal detection theory) had unit variance, and the non-diseased distribution (noise) had standard deviation  $b$  ( $b > 0$ ) or variance  $b^2$ , and that the separation of the two distributions was  $a$ , see figure below. In this example:  $a = 1.11$  and  $b = 0.556$ , corresponding to  $\mu = 2$  and  $\sigma = 1.8$ . Dorfman and Alf's fundamental contribution, namely estimating these parameters from ratings data, to be described below, led to the widespread usage of the  $(a, b)$  parameters estimated by their software (RSCORE), and its newer variants (e.g., RSCORE-II, ROCFIT and ROCKIT).

By dividing the z-samples by  $b$ , the variance of the distribution labeled "Noise" becomes unity, its mean stays at zero, and the variance of the distribution labeled "Signal" becomes  $1/b$ , and its mean becomes  $a/b$ , as shown below. It illustrates that the inverses of Eqn. (13.11) are:

$$\mu = \frac{a}{b}; \sigma = \frac{1}{b} \quad (13.12)$$

Eqns. (13.11) and (13.12) allow conversion from one notation to another.

```
grid.arrange(p1,p2,ncol=2)
```

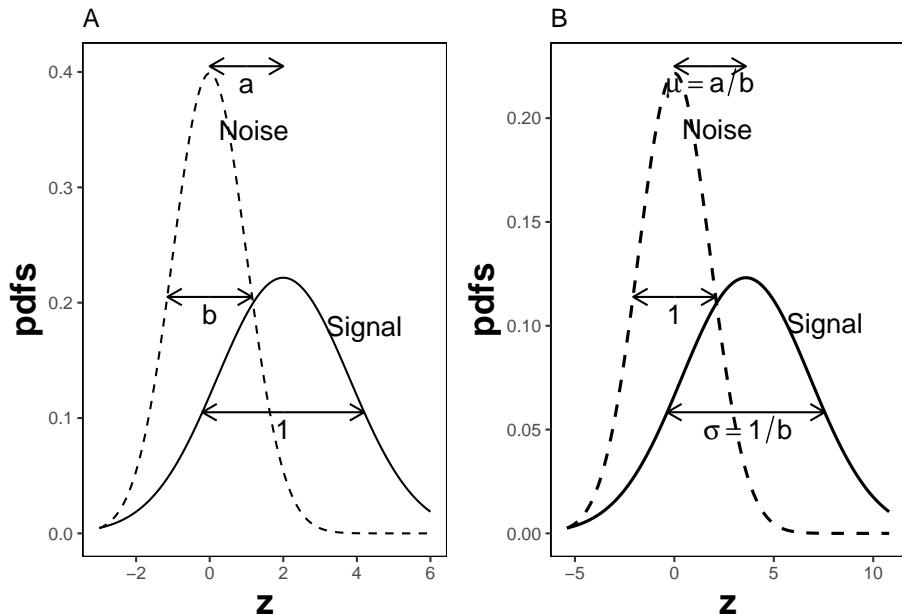


Figure 13.1: Plot A shows the definitions of the  $(a, b)$  parameters of the binormal model. In plot B the x-axis has been rescaled so that the noise distribution has unit variance, thereby illustrations between  $(a, b)$  and the  $(\mu, \sigma)$  parameters.

### 13.4 Binormal ROC curve

Using the  $(a, b)$  notation, Eqn. (13.10) for the ROC curve reduces to:

$$TPF = \Phi(a + b\Phi^{-1}(FPF)) \quad (13.13)$$

Since  $\Phi^{-1}(FPF)$  is an increasing function of its argument  $FPF$ , and  $b > 0$ , the argument of the  $\Phi$  function is an increasing function of  $FPF$ . Since  $\Phi$  is a monotonically increasing function of its argument,  $TPF$  is a monotonically

increasing function of  $FPF$ . This is true regardless of the sign of  $a$ . If  $FPF = 0$ , then  $\Phi^{-1}(0) = -\infty$  and  $TPF = 0$ . If  $FPF = 1$ , then  $\Phi^{-1}(1) = +\infty$  and  $TPF = 1$ . Regardless of the value of  $a$ , as long as  $b \geq 0$ , the ROC curve starts at  $(0,0)$  and increases monotonically ending at  $(1,1)$ .

From Eqn. (13.7) and Eqn. (13.8), the expressions for  $FPF$  and  $TPF$  in terms of model parameters  $(a, b)$  are:

$$\left. \begin{aligned} FPF(\zeta) &= \Phi(-\zeta) \\ TPF &= \Phi(a - b\zeta) \end{aligned} \right\} \quad (13.14)$$

## 13.5 Scalar threshold-independent measure

Sensitivity-specificity is a dual (two-valued) measure of performance. Using a dual measure it is difficult to unambiguously compare two systems since one cannot separate the effect of reporting threshold from the measures. For example, if sensitivity is higher for one system but specificity is higher for another, this could be due to different thresholds. Sensitivity and specificity depend on the threshold. As the threshold changes, sensitivity and specificity are both affected in opposite directions. Desirable is a scalar measure of performance that takes this variation into account and does not depend on any specific threshold.

Generally accepted measures are the partial-area  $A_{z;c}$  under the ROC, Eqn. (13.15), the full-area  $A_z$  under the ROC, Eqn. (13.18), and the  $d'$  index Eqn. (13.19).

Before deriving analytical expressions for these measures let us further examine the premise that sensitivity-specificity is undesirable because it is a 2D measure. A trivial way to convert it to a scalar measure is to sum the two values: high sensitivity and high specificity are both desirable, so a high value of their sum is certainly also desirable. In fact this is the basis for the Youden index, defined as sensitivity plus specificity minus one (Youden, 1950). (Subtracting one makes the Youden index range from 0 to 1.) However, this index varies with the position of the operating point on the ROC curve. (The operating point at which it is maximum is often thought of as the optimal operating point on the ROC curve.)

To emphasize, we desire a scalar measure that is threshold independent.

### 13.5.1 Partial AUC

While this is a scalar measure, it does depend on choice of operating point. It is included here as it yields, as a special case, a scalar measure that does not depend on choice of operating point. The details are in Section 13.12.1, which

derives the formula for the partial-area under the unequal-variance binormal model. The final result is:

$$A_{z;c} = \int_{z_2=-\infty}^{\Phi^{-1}(c)} \int_{z_1=-\infty}^{\frac{a}{\sqrt{1+b^2}}} \phi(z_1, z_2; \rho) dz_1 dz_2 \quad (13.15)$$

The threshold  $\zeta_1$  corresponding to  $FPF = c$  is given by:

$$\zeta_1 = -\Phi^{-1}(c) \quad (13.16)$$

$A_{z;c}$  is the area under the partial ROC curve extending from  $FPF = 0$  to  $FPF = c$  and  $\phi(z_1, z_2; \rho)$  is the standard bivariate normal distribution, where the correlation coefficient  $\rho$  of the distribution is defined by:

$$\rho = -\frac{b}{\sqrt{1+b^2}} \quad (13.17)$$

The bivariate 2D integral can be evaluated numerically. The following code illustrates calculation of the partial-area measure using the function `pmvnorm` in R package `mvtnorm`. The following parameter values were used:  $a = 2$ ,  $b = 1$  and  $\zeta_1 = 1.5$ . (The parameter  $b$  was deliberately chosen equal to one so that we do not have to worry about improper ROC curves.)

```

1 a <- 2;b <- 1;zeta1 <- 1.5
2 A_z <- pnorm(a/sqrt(1+b^2))
3 opPtx <- pnorm(-zeta1)
4 opPty <- pnorm(a - b * zeta1)
5 rho <- -b/sqrt(1+b^2)
6 Lower1 <- -Inf
7 Upper1 <- qnorm(opPtx)
8 Lower2 <- -Inf
9 Upper2 <- a/sqrt(1+b^2)
10 sigma <- rbind(c(1, rho), c(rho, 1))
11 A_zc <- as.numeric(pmvnorm(
12   c(Lower1, Lower2),
13   c(Upper1, Upper2),
14   sigma = sigma))

```

The partial-area measure is  $A_{z;c} = 0.0352195$ . The corresponding full-area measure is  $A_z = 0.9213504$ .  $A_{z;c}$  is small because the reporting threshold is high. However,  $A_{z;c}$  should not be confused with true performance of the observer, as shown in Section 13.6.

### 13.5.2 Full AUC

A special case of this formula is the area under the full ROC curve, shown below using both parameterizations of the binormal model:

$$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right) \quad (13.18)$$

The binormal fitted AUC increases as  $a$  increases or as  $b$  decreases. Equivalently, it increases as  $\mu$  increases or as  $\sigma$  decreases. In the example just given, the full AUC is  $A_z = 0.9213504$ .

### 13.5.3 The $d'$ measure

The  $d'$  parameter is defined as the separation of two unit-variance normal distributions yielding the same AUC as that predicted by the  $(a, b)$  parameter binormal model. It is defined by:

$$d' = \sqrt{2}\Phi^{-1}(A_z) \quad (13.19)$$

The  $d'$  index corresponding to the above binormal parameters is 2. The transformation from an index that ranges from 0.5 to 1 to one that ranges from 0 to infinity can be viewed as desirable. The  $d'$  index can be regarded as a perceptual signal-to-noise-ratio.

## 13.6 Partial AUC vs. true performance

- A *partial-area observer* such as in Section 13.5.1 rates cases as follows: for the sub-set of cases defined by  $z \geq \zeta_1$  the observer reports *explicit* ratings exactly equal to the observed z-samples (or some monotonic transformation of the z-samples). For the remaining cases the observer assigns a *fixed value rating that is smaller than  $\zeta_1$*  (the exact value does not matter; these cases are said to be assigned *implicit* ratings).
- In contrast, the *full-area observer* reports explicit ratings for all cases.

*To measure true performance of the partial-area observer one must, of course, include all cases.* The ROC curve extends continuously from the origin to the solid dot *plus the area under the dotted line* extending from the solid dot to (1,1). True performance, the area under the continuous section plus that under the straight line extension, is denoted  $A_{z;c,TRUE}$  and is defined by:

$$A_{z;c,\text{TRUE}} = A_{z;c} + \frac{(1 - FPF)(1 + TPF)}{2} \quad (13.20)$$

In other words one adds to  $A_{z;c}$  the area of the trapezoid with bases each equal to  $(1 - FPF)$  and opposing sides equal to  $TPF$  and unity.

Since the partial-area observer does not preserve ordering information, *true performance of a partial-area observer is smaller than performance  $A_z$  of a full-area observer.*

$$A_{z;c,\text{TRUE}} \leq A_z \quad (13.21)$$

True performance is illustrated with the following simulation 2AFC study. The Wilcoxon function, defined next, can be thought of as the mathematical equivalent of a 2AFC study, conducted with all possible pairings of non-diseased and diseased cases. For each pairing, if the z-sample of the diseased case exceeds that of the non-diseased case one adds unity to a zero-initialized counter; if it is smaller one does nothing; if they are equal one adds 0.5; and finally one divides by the number of comparisons.

```

1  Wilcoxon <- function (zk1, zk2)
2  {
3      K1 = length(zk1)
4      K2 = length(zk2)
5      W <- 0
6      for (k1 in 1:K1) {
7          W <- W + sum(zk1[k1] < zk2)
8          W <- W + 0.5 * sum(zk1[k1] == zk2)
9      }
10     W <- W/K1/K2
11     return (W)
12 }
```

The following code saves 10,000 pairs of ratings in two arrays:  $z[1,]$  and  $z[2,]$ . The first array corresponds to non-diseased cases and the second to diseased cases. Note the usage, at lines 3-4, of the  $a, b$  values to define the two distributions. The array  $zc$ , initially a copy of  $z$ , is selectively binned by setting, lines 6-7, all ratings less than  $\zeta_1$  to -100. The ordering information for these z-samples is lost.

```

1  nPairs <- 10000
2  z <- array(dim = c(2, nPairs))
3  z[1,] <- rnorm(nPairs, sd = b)
4  z[2,] <- rnorm(nPairs, mean = a, sd = 1)
5  zc <- z
```

```

6  zc[1,z[1,] <- -100 # ratings of partial area observer
7  zc[2,z[2,] <- -100 # do:

```

The following code prints the predicted and observed full areas under the ROCs followed by the predicted and observed true performances. With this many cases sampling variability is small and the predicted and observed values are close.

```

#> A_z predicted =  0.9213504
#> A_z observed =  0.9226259
#> A_z{c;true} predicted =  0.8244498
#> A_z{c;true} observed =  0.8260062

```

Note that:

- $A_{z;c,\text{TRUE}} < A_z$ , because ordering information is lost for all cases with z-samples less than  $\zeta_1$ .
- $A_{z;c,\text{TRUE}} >> A_{z;c}$ , because of the large contribution from the area under the straight line, left poanel Fig. 13.2.

## 13.7 Illustrative plots

In the ROC plots below the partial-area observer curve is shown as a continuous line extending from the origin to the limiting point *plus* a dotted line extending from the limiting point to (1,1). The continuous section is determined by cumulating cases with z-samples  $z \geq \zeta_1$  while the (1,1) point is determined by cumulating all cases.

The ROC curve for both types of observers is shown in the left panel of 13.2 for the following parameters:  $a = 2$ ,  $b = 1$  and  $\zeta_1 = 1.5$ ;  $\zeta_1$  corresponds to  $c \equiv FPF = \Phi(-\zeta_1) = 0.0668072$  and  $TPF = \Phi(a - b\zeta_1) = 0.6914625$ . In other words the limiting point coordinates are (0.067, 0.691), shown in the plot by the solid dot. Partial AUC  $A_{z;c}$  equals 0.0352195. The full-area ROC curve, shown by the complete solid curve, extends from (0,0) to (1,1), the area under which is  $A_z = 0.9213504$ .

As FPF increases true-performance increases. the right panel of Fig. 13.2 shows the variation of true performance  $A_{z;c,\text{TRUE}}$  with FPF. The curve starts from (0, 0.5) and ends at (1.000, 0.921). For low values of FPF the curve is very steep while for  $FPF > 0.25$  the curve levels out, approaching the maximum value defined by  $A_z = 0.9213504$ . True performance is maximized at  $\zeta_1 = -\infty$ .

Fig. 13.3, left panel, corresponding to  $a = 1$ ,  $b = 0.2$  and  $\zeta_1 = 1.5$ , shows an improper ROC curve. The dashed line is well above the continuous curve

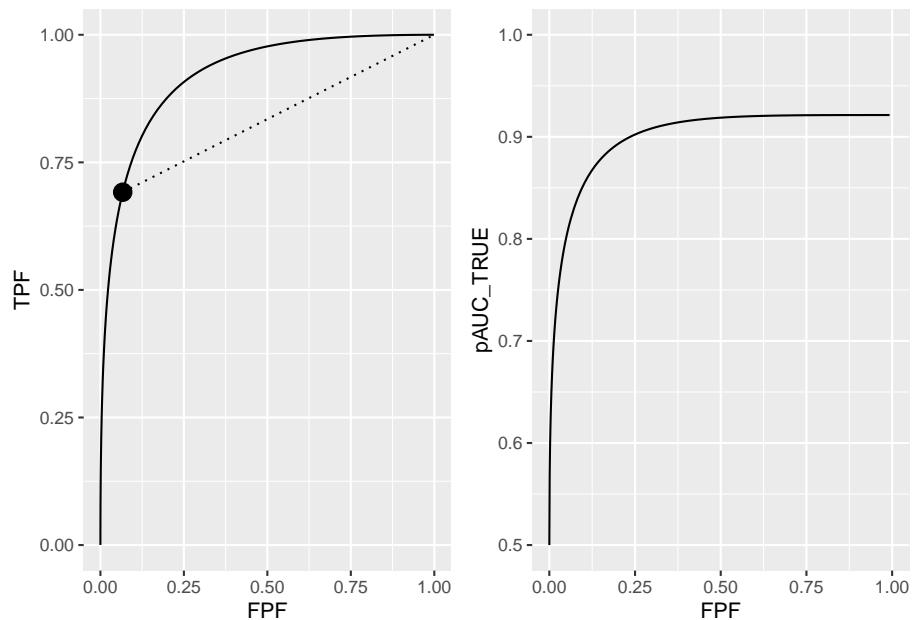


Figure 13.2: Left panel: binormal ROC curve corresponding to  $a = 2$  and  $b = 1$ . The dot is the operating point corresponding to  $\zeta_1 = 1.5$ . The continuous curve extending from the origin to  $(1,1)$  represents the full ROC. Note that in the region above the dot the continuous curve is above the dotted line, meaning true performance of an observer who only rates a sub-set of cases is less than performance of an observer who rates all cases. Right panel: variation of true performance with FPF; at  $FPF = 0$  the plot starts at ordinate equal to 0.5 and levels out at  $FPF = 1$  at  $AUC = A_z = 0.921$ .

and true performance is maximized at a finite value of  $\zeta_1$ , corresponding to  $FPF = 0.153$ , see right panel. This is an invalid conclusion since an improper ROC curve is a fitting artifact of the binormal model easily avoided by using modern curve-fitting methods (eg., PROPROC, CBM or RSM). However, since the wAFROC has an operating characteristic with an improper-like feature but which is not a fitting artifact, this example serves a purpose, elaborated on in Chapter 39, where it is shown that by maximizing the area under the wAFROC one can find the optimal threshold of an algorithmic observer.

```
#> true performance max occurs at FPF = 0.1525
```

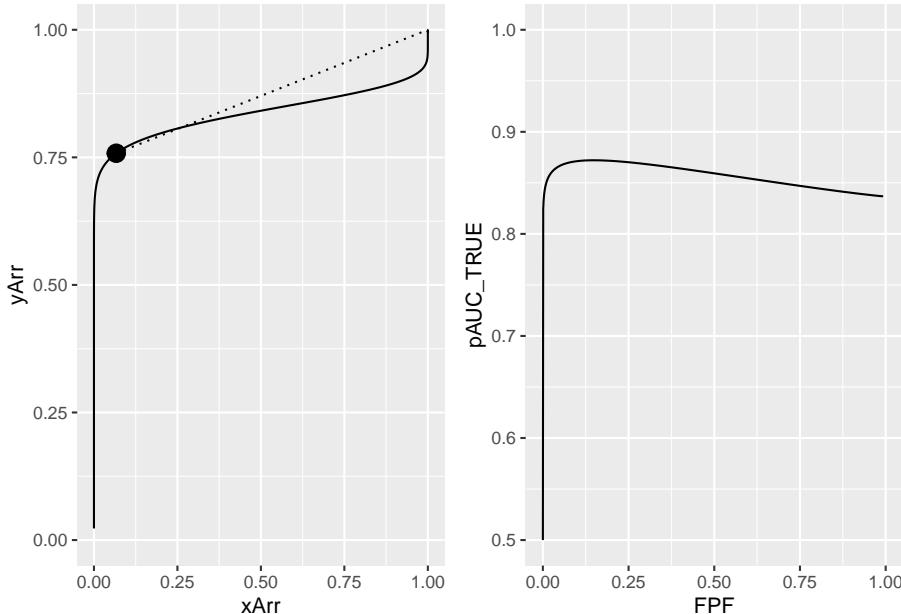


Figure 13.3: The left panel shows the visibly improper ROC curve for  $a = 1$  and  $b = 0.2$ . The solid line is below the dotted line. The right panel shows the variation of true performance  $pAUC\_TRUE$  with  $FPF$ . True performance is maximized at  $FPF = 0.153$ . Since improper ROC fits are fitting artifacts, this example does not negate the previous finding that true performance for a proper ROC curve is maximized by setting the threshold to report all cases, i.e.,  $FPF = 1$ .

## 13.8 Geometrical argument

Defining geometrical features of a proper ROC are:

- As one moves up the curve the slope decreases monotonically;
- At each point the slope is greater than that of the straight line connecting the point to (1,1);
- The curve ends at (1,1).

The geometry ensures that true performance for a proper ROC is maximized at  $\zeta_1 = -\infty$ , i.e., at  $\text{PPF} = 1$ , as in Fig. 13.2, right panel.

### 13.9 Optimal operating point on ROC

We have seen that optimal ROC AUC is achieved by setting  $\zeta_1 = -\infty$ , i.e., by reporting all cases as diseased. Of course, from clinical considerations, this is nonsense. Consider screening mammography, where typically for every 1000 cases only 5 are malignant. Recalling everybody would incur huge costs from having to rule out cancer in 995 actually non-diseased patients. Of course the 5 malignant cancers would be confirmed at the follow-up diagnostic mammography examination. But one can clearly see that the benefit of correctly detecting the 5 malignancies is far outweighed by the 995 unnecessary recalls. And if one is going to recall everybody, why perform the initial screening mammography exam?

So what is going on? The problem is that AUC measures classification performance in a 2AFC task. A screening examination is not a 2AFC task: the radiologist is not presented two cases, one non-diseased and one diseased, and asked to pick the diseased patient. Rather, the radiologist is shown images of a single patient, and the object is to maximize the detection rate while minimizing false positives.

To address this optimization task one needs to know the costs and benefits of the four decision outcomes in the binary paradigm: true and false positives, and true and false negatives. This has been addressed in (Metz, 1978). Here is the reasoning. Let

- $C_0$  denote the overhead cost of performing the imaging examination,
- $C_{\text{TP}}$  denote the cost of a true positive decision (a benefit can be expressed as a negative cost),
- $C_{\text{FN}}$  denote the cost of a false negative decision,
- $C_{\text{FP}}$  denote the cost of a false positive decision, and
- $C_{\text{TN}}$  denote the cost (or negative benefit) of a true negative decision.

It is shown (Metz, 1978) that the average cost of the examination is:

$$\bar{C} = C_0 + C_{\text{TP}}P(\text{TP}) + C_{\text{TN}}P(\text{TN}) + C_{\text{FP}}P(\text{FP}) + C_{\text{FN}}P(\text{FN}) \quad (13.22)$$

In this equation  $P(\text{TP})$  is the probability of a TP-event, etc. These probabilities are related to disease prevalence  $P(+)$  and the operating point by:

$$\left. \begin{array}{l} P(\text{TP}) = P(+) \text{TPF} \\ P(\text{TN}) = (1 - P(+))(1 - \text{FPF}) \\ P(\text{FP}) = (1 - P(+))\text{FPF} \\ P(\text{FN}) = P(+)(1 - \text{TPF}) \end{array} \right\} \quad (13.23)$$

With these substitutions one gets for the average cost:

$$\bar{C} = C_0 + C_{\text{TN}}P(-) + C_{\text{FN}}P(+) + (C_{\text{TP}} - C_{\text{FN}})P(+) \text{TPF} + (C_{\text{FP}} - C_{\text{TN}})P(-) \text{FPF} \quad (13.24)$$

Equating the derivative of the average cost to zero, to minimize the average cost, one gets:

$$\frac{d(\text{TPF})}{d(\text{FPF})} = \frac{C_{\text{FP}} - C_{\text{TN}}}{C_{\text{FN}} - C_{\text{TP}}} \frac{P(-)}{P(+)} \quad (13.25)$$

This defines the slope  $\frac{d(\text{TPF})}{d(\text{FPF})}$  of the ROC at the optimal operating point, i.e., the point that minimizes the average cost of the examination. Note that  $P(-) = 1 - P(+)$ .

- If disease prevalence is high, then the optimal operating point is where the slope of the ROC is low, which is near the upper-right corner. With mostly diseased cases it makes sense to set the operating point at high sensitivity and low specificity. Conversely, with low prevalence, one should set the operating point at low sensitivity and high specificity.
- For a given disease prevalence, if the cost of a FP decision is high (or if the benefit of a TN is high - recall that a benefit is the same as a negative cost), then the optimal operating point is where the slope of the ROC is high, which is near the lower-left corner. One sets the operating point at low sensitivity and high specificity.
- For a given disease prevalence, if the cost of a FN decision is high (or if the benefit of a TP is high), then the optimal operating point is where the slope of the ROC is low, which is near the upper-right corner. One sets the operating point at high sensitivity and low specificity.

The costs and benefits are often difficult to quantify. If one assumes that the right hand side of Eqn. (13.25) equals unity (e.g., the four costs / benefits are equal and disease-prevalence is 50%) then the optimal operating point is defined by that point on the ROC curve where the slope is unity, which is the point

of nearest approach of the curve to the upper-left corner. This corresponds to maximizing the Youden index (Youden, 1950), defined as the sum of sensitivity and specificity minus one. This is demonstrated in the following code.

```
a <- 2;b <- 1
z <- seq(-3,5.5,0.05)
FPF <- pnorm(-z)
TPF <- pnorm(a - b*z)
Youden <- TPF + (1 - FPF) - 1
curve <- data.frame(FPF = FPF, TPF = TPF, YOU = Youden)
dist <- sqrt(FPF^2 + (1 - TPF)^2)
p1 <- ggplot2::ggplot(curve, aes(x = FPF, y = TPF)) +
  geom_line() +
  scale_x_continuous(limits = c(0,1)) + scale_y_continuous(limits = c(0,1))
p2 <- ggplot2::ggplot(curve, aes(x = FPF, y = YOU)) +
  geom_line() +
  scale_x_continuous(limits = c(0,1)) + scale_y_continuous(limits = c(0,1))
indxDist <- which(dist == min(dist))
indxYoud <- which(Youden == max(Youden))
if (indxDist != indxYoud) stop("The two indices are different") else {
  cat("Op Pt corresponding to max Youden and min distance is: \n",
      FPF[indxDist],
      "\nTPF = ",
      TPF[indxDist])
}
#> Op Pt corresponding to max Youden and min distance is:
#> FPF =  0.1586553
#> TPF =  0.8413447
```

## 13.10 Discussion

The binormal model is historically very important and the contribution by Dorfman and Alf (Dorfman and Alf, 1969) was seminal. Prior to their work, there was no valid way of estimating AUC from observed ratings counts. Their work and a key paper (Lusted, 1971) accelerated research using ROC methods. The number of publications using their algorithm, and the more modern versions developed by Metz and colleagues, is probably well in excess of 500. Because of its key role, I have endeavored to take out some of the mystery about how the binormal model parameters are estimated. In particular, a common misunderstanding that the binormal model assumptions are violated by real datasets, when in fact it is quite robust to apparent deviations from normality, is addressed.

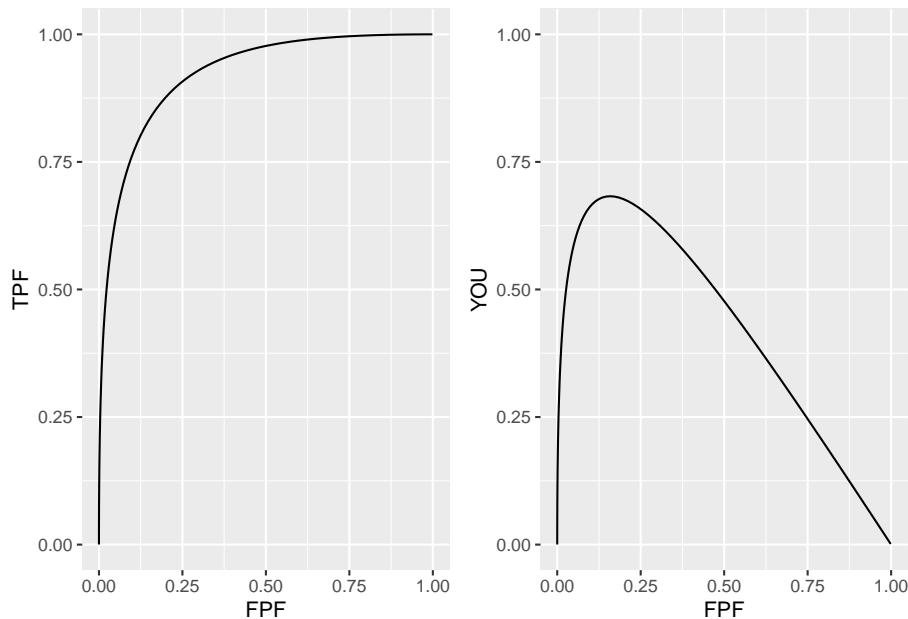


Figure 13.4: Left panel: binormal ROC curve corresponding to  $a = 2$  and  $b = 1$ . Right panel: variation of Youden index with FPF; the plot shows a maximum at  $\text{FPF} = 0.1586553$ ; this corresponds to the nearest approach of the ROC curve to the upper-left corner.

A good understanding of this chapter should enable the reader to better understand alternative ROC models, discussed later.

It has been stated that the  $b$ -parameter of the binormal model is generally observed to be less than one, consistent with the diseased distribution being wider than the non-diseased one. The ROC literature is largely silent on the reason for this finding. One reason, namely location uncertainty, is presented in Chapter “Predictions of the RSM”, where RSM stands for Radiological Search Model. Basically, if the location of the lesion is unknown, then  $z$ -samples from diseased cases can be of two types, samples from the correct lesion location, or samples from other non-lesion locations. The resulting mixture distribution will then appear to have larger variance than the corresponding samples from non-diseased cases. This type of mixing need not be restricted to location uncertainty. Even if location is known, if the lesions are non-homogenous (e.g., they contain a range of contrasts) then a similar mixture-distribution induced broadening is expected. The contaminated binormal model (CBM) - see Chapter TBA - also predicts that the diseased distribution is wider than the non-diseased one.

The fact that the  $b$ -parameter is less than unity implies that the predicted ROC curve is improper, meaning its slope is not monotone decreasing as the operating point moves up the curve. The result is that a portion of the curve, near  $(1,1)$  that crosses the chance-diagonal and hooks upward approaching  $(1,1)$  with infinite slope. Ways of fitting proper ROC curves are described in Chapter “Other proper ROC models”. Usually the hook is not readily visible, which has been used as an excuse to ignore the problem. For example, in Fig. 6.4, one would have to “zoom-in” on the upper right corner to see it, but the reader should make no mistake about it, the hook is there as .

A recent example is Fig. 1 in the publication resulting from the Digital Mammographic Imaging Screening Trial (DMIST) clinical trial (Pisano et al., 2005) involving 49,528 asymptomatic women from 33 clinical sites and involving 153 radiologists, where each of the film modality ROC plots crosses the chance diagonal and hooks upwards to  $(1,1)$ , which as is known, results anytime  $b < 1$ .

The unphysical nature of the hook (predicting worse than chance-level performance for supposedly expert readers) is not the only reason for seeking alternate ROC models. The binormal model is susceptible to degeneracy problems. If the dataset does not provide any interior operating points (i.e., all observed points lie on the axes defined by  $FPP = 0$  or  $TPF = 1$ ) then the model fits these points with  $b = 0$ . The resulting straight-line segment fits do not make physical sense. These problems are addressed by the contaminated binormal model<sup>16</sup> to be discussed in Chapter “Other proper ROC models”. The first paper in the series has particularly readable accounts of data degeneracy.

To this day the binormal model is widely used to fit ROC datasets. In spite of its limitations, the binormal model has been very useful in bringing a level of quantification to this field that did not exist prior to (Dorfman and Alf, 1969).

## 13.11 Appendix I: Density functions

According to Eqn. (13.1) the probability that a z-sample is smaller than a specified threshold  $\zeta$ , i.e., the CDF function, is:

$$P(Z \leq \zeta | Z \sim N(0, 1)) = 1 - FPF(\zeta) = \Phi(\zeta)$$

$$P(Z \leq \zeta | Z \sim N(\mu, \sigma^2)) = 1 - TPF(\zeta) = \Phi\left(\frac{\zeta - \mu}{\sigma}\right)$$

Since the *pdf* is the derivative of the corresponding CDF function, it follows that (the subscripts N and D denote non-diseased and diseased cases, respectively):

$$pdf_N(\zeta) = \frac{\partial \Phi(\zeta)}{\partial \zeta} = \phi(\zeta) \equiv \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\zeta^2}{2}\right)$$

$$pdf_D(\zeta) = \frac{\partial \Phi\left(\frac{\zeta - \mu}{\sigma}\right)}{\partial \zeta} = \frac{1}{\sigma} \phi\left(\frac{\zeta - \mu}{\sigma}\right) \equiv \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\zeta - \mu)^2}{2\sigma^2}\right)$$

The second equation can be written in  $(a, b)$  notation as:

$$pdf_D(\zeta) = b\phi(b\zeta - a) = \frac{b}{\sqrt{2\pi}} \exp\left(-\frac{(b\zeta - a)^2}{2}\right)$$

## 13.12 Appendix II: Area under binormal ROC

### 13.12.1 General case (partial-area)

This section is based on (Thompson and Zucchini, 1989). In what follows, FPF is abbreviates to x and TPF to y. Then the equation for the ROC curve is (13.13):

$$y = \Phi(a + b\Phi^{-1}(x)) \quad (13.26)$$

The partial-area under the ROC curve from  $x = 0$  to  $x = c$ , where  $0 \leq c \leq 1$ , is given by:

$$A_{z;c} = \int_0^c y dx = \int_0^c dx \Phi(a + b\Phi^{-1}(x)) \quad (13.27)$$

Define change of variable:

$$x = \Phi(x_1) \quad (13.28)$$

which implies:

$$\left. \begin{aligned} x_1 &= \Phi^{-1}(x) \\ dx &= dx_1 \phi(x_1) \end{aligned} \right\} \quad (13.29)$$

This yields:

$$\left. \begin{aligned} A_{z;c} &= \int_0^c dx \Phi(a + bx_1) \\ &= \int_{-\infty}^{\Phi^{-1}(c)} dx_1 \phi(x_1) \Phi(a + bx_1) \end{aligned} \right\} \quad (13.30)$$

The right hand side of Eqn. (13.30) can be expressed as an integral over the bivariate normal distribution as follows. From the definition of the  $\Phi$  function the above integral can be written as the following double integral:

$$A_{z;c} = \int_{x_1=-\infty}^{\Phi^{-1}(c)} dx_1 \phi(x_1) \int_{x_2=-\infty}^{a+bx_1} \phi(x_2) dx_2 \quad (13.31)$$

Change variables from  $(x_1, x_2)$  to  $(z_1, z_2)$  as follows:

$$\left. \begin{aligned} z_2 &= x_1 \\ z_1 &= (x_2 - bx_1) f \end{aligned} \right\} \quad (13.32)$$

Here  $f$  is a quantity to be determined, which will allow us to complete the transformation to the desired bivariate integral. The second equation above can be written as:

$$x_2 = \frac{z_1}{f} + bx_1 = \frac{z_1}{f} + bz_2 \quad (13.33)$$

The Jacobian (Stein and Barcellos, 1992) of the transformation is

$$J = \begin{pmatrix} 0 & 1 \\ \frac{1}{f} & b \end{pmatrix} \quad (13.34)$$

The magnitude of the determinant of  $J$  is  $1/f$ .

From a theorem in calculus (Stein and Barcellos, 1992), the double integral over  $(x_1, x_2)$  can be expressed in terms of a double integral over  $(z_1, z_2)$  as follows:

$$A_{z;c} = \frac{1}{f} \int_{z_2=-\infty}^{\Phi^{-1}(c)} dz_2 \phi(z_2) \int_{z_1=-\infty}^{z_1^{UL}} \phi\left(\frac{z_1}{f} + bz_2\right) dz_1 \quad (13.35)$$

The upper limit of the inner integral can be calculated as follows. Using the second equation in Eqn. (13.32):

$$z_1^{UL} = (x_2^{UL} - bx_1) f = (a + bx_1 - bx_1) f = af \quad (13.36)$$

Eqn. (13.35) simplifies to:

$$A_{z;c} = \frac{1}{f} \int_{z_2=-\infty}^{\Phi^{-1}(c)} dz_2 \phi(z_2) \int_{z_1=-\infty}^{af} \phi\left(\frac{z_1}{f} + bz_2\right) dz_1 \quad (13.37)$$

Perform a change of variable from  $f$  to a correlation-like quantity  $\rho$  defined by:

$$f = \sqrt{1 - \rho^2} \quad (13.38)$$

Define  $\rho$  in terms of the b-parameter as follows:

$$b\sqrt{1 - \rho^2} = -\rho \quad (13.39)$$

This implies that  $\rho$  is given by:

$$\rho = -\frac{b}{\sqrt{1 + b^2}} \quad (13.40)$$

The argument of the right-most  $\phi$  function in Eqn. (13.37) simplifies as follows:

$$\frac{z_1}{f} + bz_2 = \frac{z_1 + bz_2\sqrt{1 - \rho^2}}{\sqrt{1 - \rho^2}} = \frac{z_1 - \rho z_2}{\sqrt{1 - \rho^2}} \quad (13.41)$$

The expression for the partial-area under the ROC reduces to:

$$A_{z;c} = \frac{1}{\sqrt{1 - \rho^2}} \int_{z_2=-\infty}^{\Phi^{-1}(c)} dz_2 \phi(z_2) \int_{z_1=-\infty}^{a\sqrt{1-\rho^2}} \phi\left(\frac{z_1 - \rho z_2}{\sqrt{1 - \rho^2}}\right) dz_1 \quad (13.42)$$

Eqn. (13.39) implies:

$$1 - \rho^2 = \frac{1}{1 + b^2} \quad (13.43)$$

Therefore,

$$A_{z;c} = \frac{1}{\sqrt{1 - \rho^2}} \int_{z_2=-\infty}^{\Phi^{-1}(c)} dz_2 \phi(z_2) \int_{z_1=-\infty}^{\frac{a}{\sqrt{1+b^2}}} \phi\left(\frac{z_1 - \rho z_2}{\sqrt{1 - \rho^2}}\right) dz_1 \quad (13.44)$$

The standard bivariate normal distribution with correlation coefficient  $\rho$  is defined by:

$$\phi(z_1, z_2; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{z_1^2 - 2\rho z_1 z_2 + z_2^2}{2(1-\rho^2)}\right) \quad (13.45)$$

The standard normal distribution is defined by:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \quad (13.46)$$

It can be shown using these definitions that:

$$\phi(z_1, z_2; \rho) = \frac{1}{\sqrt{1-\rho^2}} \phi(z_2) \phi\left(\frac{z_1 - \rho z_2}{\sqrt{1-\rho^2}}\right) \quad (13.47)$$

Using this form the expression for the partial-area is:

$$A_{z;c} = \int_{z_2=-\infty}^{\Phi^{-1}(c)} \int_{z_1=-\infty}^{\frac{a}{\sqrt{1+b^2}}} \phi(z_1, z_2; \rho) dz_1 dz_2 \quad (13.48)$$

### 13.12.2 Special case (total area)

Since  $c$  is the upper limit of FPF, setting  $c = 1$  yields the total area under the binormal ROC curve <sup>2</sup>:

---

<sup>2</sup>Since the integral over  $z_2$  is over the entire range it integrates out to unity leaving the one-dimensional density function  $\phi(z_1)$  inside the integral. The last step follows from the definition of the  $\Phi$  function.

$$\left. \begin{aligned} A_z &= \int_{z_2=-\infty}^{\infty} \int_{z_1=-\infty}^{\frac{a}{\sqrt{1+b^2}}} \phi(z_1, z_2; \rho) dz_1 dz_2 \\ &= \int_{z_1=-\infty}^{\frac{a}{\sqrt{1+b^2}}} \phi(z_1) dz_1 \\ &= \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \end{aligned} \right\} \quad (13.49)$$

An equivalent forms for the total area under the unequal variance binormal ROC curve is:

$$\left. \begin{aligned} A_z &= \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \\ &= \Phi\left(\frac{\frac{a}{b}}{\sqrt{1+\frac{1}{b^2}}}\right) \\ &= \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right) \end{aligned} \right\} \quad (13.50)$$

### 13.13 Appendix III: Invariance property of pdfs

The binormal model is not as restrictive as might appear at first sight. Any monotone increasing transformation  $Y = f(Z)$  applied to the observed  $z$ -samples, and the associated thresholds, will yield the same observed data, e.g., Table 11.1. This is because such a transformation leaves the ordering of the ratings unaltered and hence results in the same operating points. While the distributions for  $Y$  will not be binormal (i.e., two independent normal distributions), one can safely “pretend” that one is still dealing with an underlying binormal model. An alternative way of stating this is that any pair of distributions is allowed as long as they are reducible to a binormal model form by a monotonic increasing transformation of  $Y$ : e.g.,  $Z = f^{-1}$ . [If  $f$  is a monotone increasing function of its argument, so is  $f^{-1}$ .] For this reason, the term “pair of latent underlying normal distributions” is sometimes used to describe the binormal model. The robustness of the binormal model has been investigated (Hanley, 1988; Dorfman et al., 1997). The referenced paper by Dorfman et al has an excellent discussion of the robustness of the binormal model.

The robustness of the binormal model, i.e., the flexibility allowed by the infinite choices of monotonic increasing functions, application of each of which leaves the ordering of the data unaltered, is widely misunderstood. The non-Gaussian appearance of histograms of ratings in ROC studies can lead one to incorrect

conclusions that the binormal model is inapplicable to these datasets. To quote a reviewer of one of my recent papers:

I have had multiple encounters with statisticians who do not understand this difference.... They show me histograms of data, and tell me that the data is obviously not normal, therefore the binormal model should not be used.

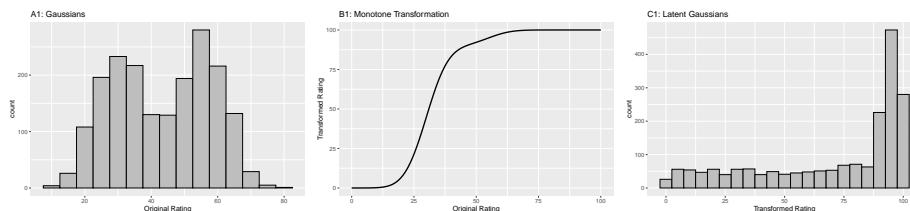
The reviewer is correct. The misconception is illustrated next.

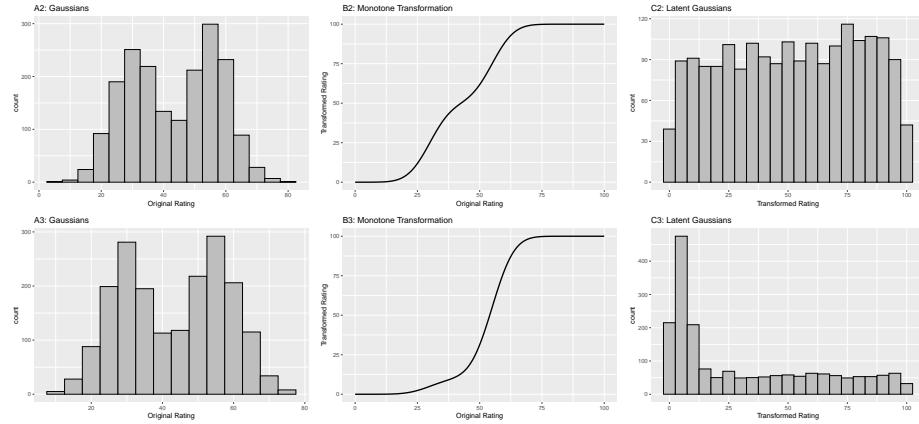
```
# shows that monotone transformations have no effect on
# AUC even though the pdfs look non-gaussian
# common misconception about ROC analysis
fArray <- c(0.1,0.5,0.9)
seedArray <- c(10,11,12)
for (row in 1:3) {
  f <- fArray[row]
  seed <- seedArray[row]
  set.seed(seed)
  # numbers of cases simulated
  K1 <- 900
  K2 <- 1000
  mu1 <- 30
  sigma1 <- 7
  mu2 <- 55
  sigma2 <- 7
  # Simulate true gaussian ratings using above parameter values
  z1 <- rnorm(K1,mean = mu1,sd = sigma1)
  z1[z1>100] <- 100;z1[z1<0] <- 0 # constrain to 0 to 100
  z2 <- rnorm(K2,mean = mu2,sd = sigma2)
  z2[z2>100] <- 100;z2[z2<0] <- 0 # constrain to 0 to 100
  # calculate AUC for true Gaussian ratings
  AUC1 <- TrapezoidalArea(z1, z2)
  Gaussians <- c(z1, z2)
  # display histograms of true Gaussian ratings, A1, A2 or A3
  x <- data.frame(x=Gaussians) # line 27
  x <-
    ggplot(data = x, mapping = aes(x = x)) +
    geom_histogram(binwidth = 5, color = "black", fill="grey") +
    xlab(label = "Original Rating") +
    ggtitle(label = paste0("A", row, ":", "Gaussians"))
  print(x)
  z <- seq(0.0, 100, 0.1)
  # transform the latent Gaussians to true Gaussians
  transformation <-
```

```

data.frame(
  x = z,
  z = Y(z,mu1,mu2,sigma1,sigma2,f))
# display transformation functions, B1, B2 or B3
x <-
  ggplot(mapping = aes(x = x, y = z)) +
  geom_line(data = transformation, size = 1) +
  xlab(label = "Original Rating") +
  ylab(label = "Transformed Rating") +
  ggtitle(label = paste0("B", row, ":", "Monotone Transformation"))
print(x)
y <- Y(c(z1, z2),mu1,mu2,sigma1,sigma2,f)
y1 <- y[1:K1];y2 <- y[(K1+1):(K1+K2)]
# calculate AUC for transformed ratings
AUC2 <- TrapezoidalArea( y1, y2)
# display histograms of latent Gaussian ratings, C1, C2 or C3
x <- data.frame(x=y)
x <- ggplot(data = x, mapping = aes(x = x)) +
  geom_histogram(binwidth = 5, color = "black", fill="grey") +
  xlab(label = "Transformed Rating") +
  ggtitle(label = paste0("C", row, ":", "Latent Gaussians"))
print(x)
# print AUCs, note they are identical (for each row)
options(digits = 9)
cat("row =", row, ", seed =", seed, ", f =", f,
    "\nAUC of actual Gaussians =", AUC1,
    "\nAUC of latent Gaussians =", AUC2, "\n")
}
#> row = 1 , seed = 10 , f = 0.1
#> AUC of actual Gaussians = 0.99308
#> AUC of latent Gaussians = 0.99308
#> row = 2 , seed = 11 , f = 0.5
#> AUC of actual Gaussians = 0.993668889
#> AUC of latent Gaussians = 0.993668889
#> row = 3 , seed = 12 , f = 0.9
#> AUC of actual Gaussians = 0.995041111
#> AUC of latent Gaussians = 0.995041111

```





**Figure captions (A1 - C3):** Illustrating the invariance of ROC analysis to arbitrary monotone transformations of the ratings. Each row contains 3 plots: labeled 1, 2 and 3. Each column contains 3 plots labeled A, B and C. So, for example, plot C2 refers to the second row and third column. The for-loop generates the plot one row at a time. Each of the latent Gaussian plots C1, C2 and C3 appears not binormal. However, using the inverse of the monotone transformations shown B1, B2 and B3, they can be transformed to the binormal model histograms A1, A2 and A3. Plot A1 shows the histogram of simulated ratings from a binormal model. Two peaks, one at 30 and the other at 55 are evident (by design, all ratings in this figure are in the range 0 to 100). Plot B1 shows the monotone transformation for  $f = 0.1$ . Plot C1 shows the histogram of the transformed rating. The choice of  $f$  leads to a transformed rating histogram that is peaked near the high end of the rating scale. For A1 and C1 the corresponding AUCs are identical (0.993080000). Plot A2 is for a different seed value, plot B2 is the transformation for  $f = 0.5$  and now the transformed histogram is almost flat, plot C2. For plots A2 and C2 the corresponding AUCs are identical (0.993668889). Plot A3 is for a different seed value, B3 is the transformation for  $f = 0.9$  and the transformed histogram C3 is peaked near the low end of the transformed rating scale. For plots A3 and (C3) the corresponding AUCs are identical (0.995041111).

The idea is to simulate continuous ratings data in the range 0 to 100 from a binormal model.  $K_1 = 900$  non-diseased cases are sampled from a Gaussian centered at  $\mu_1 = 30$  and standard deviation  $\sigma_1 = 7$ .  $K_2 = 1000$  diseased cases are sampled from a Gaussian centered at  $\mu_2 = 55$  and standard deviation  $\sigma_2 = 7$ . The variable  $f$ , which is in the range (0,1), controls the shape of the transformed distribution. If  $f$  is small, the transformed distribution will be peaked towards 0 and if  $f$  is unity, it will be peaked at 100. If  $f$  equals 0.5, the transformed distribution is flat. Insight into the reason for this transformation is in (Press et al., 2007), Chapter 7: it has to do with transformations of random variables. The transformation function,  $Y(Z)$ , implements:

$$Y(Z) = \left[ (1-f) \Phi\left(\frac{Z-\mu_1}{\sigma_1}\right) + f \Phi\left(\frac{Z-\mu_2}{\sigma_2}\right) \right] 100 \quad (13.51)$$

The multiplication by 100 ensures that the transformed variable is in the range 0 to 100 (if not, it is code-constrained to be). The code realizes the random samples, calculates the empirical AUC, displays the histogram of the true binormal samples, plots the transformation function, calculates the empirical AUC using the transformed samples, and plots the histogram of the transformed samples (the latent binormal).

- B1 shows the transformation for  $f = 0.1$ . The steep initial rise of the curve has the effect of flattening the histogram of the transformed ratings at the low end of the rating scale, C1. Conversely, the flat nature of the curve near upper end of the rating range has the effect of causing the histogram of the transformed variable to peak in that range.
- B2 shows the transformation for  $f = 0.5$ . This time the latent rating histogram, C2, is almost flat over the entire range, definitely not visually binormal.
- B3 shows the transformation for  $f = 0.9$ . This time the transformed rating histogram, C3, is peaked at the low end of the transformed rating scale.
- The output lists the values of the seed variable and the value of the shape parameter  $f$ . *For each value of seed and the shape parameter, the AUCs of the actual Gaussians and the transformed variables are identical.*
- The values of the parameters were chosen to best illustrate the true binormal nature of the plots A2 and A3. This has the effect of making the AUCs close to unity.

The histograms in C1, C2 and C3 appear to be non-Gaussian. The corresponding non-diseased and diseased ratings will fail tests of normality. [Showing this is left as an exercise for the reader.] Nevertheless, they are latent Gaussians in the sense that the inverses of the transformations shown in B1, B2 and B3 will yield histograms that are strictly binormal, i.e., A1, A2 and A3. By appropriate changes to the monotone transformation function, the histograms shown in C1, C2 and C3 can be made to resemble a wide variety of shapes, for example, quasi-bimodal (don't confuse bimodal with binormal) histograms.]

**Visual examination of the shape of the histograms of ratings, or standard tests for normality, yield little, if any, insight into whether the underlying binormal model assumptions are being violated.**

## 13.14 Appendix IV: Fitting an ROC curve

### 13.14.1 JAVA fitted ROC curve

This section, described in the physical book, has been abbreviated to a relevant website.

### 13.14.2 Simplistic straight line fit to the ROC curve

To be described next is a method for fitting data such as in Table 11.1 to the binormal model, i.e., determining the parameters  $(a, b)$  and the thresholds  $\zeta_r$ ,  $r = 1, 2, \dots, R - 1$ , to best fit, in some to-be-defined sense, the observed cell counts. The most common method uses an algorithm called maximum likelihood. But before getting to that, I describe the least-square method, which is conceptually simpler, but not really applicable, as will be explained shortly.

#### 13.14.2.1 Least-squares estimation

By applying the function  $\Phi^{-1}$  to both sides of Eqn. (13.10), one gets (the “inverse” function cancels the “forward” function on the right hand side):

$$\Phi^{-1}(TPF) = a + b\Phi^{-1}(FPF)$$

This suggests that a plot of  $y = \Phi^{-1}(TPF)$  vs.  $x = \Phi^{-1}(FPF)$  is expected to follow a straight line with slope  $b$  and intercept  $a$ . Fitting a straight line to such data is generally performed by the method of least-squares, a capability present in most software packages and spreadsheets. Alternatively, one can simply visually draw the best straight line that fits the points, memorably referred to (Press et al., 2007) as “chi-by-eye”. This was the way parameters of the binormal model were estimated prior to Dorfman and Alf’s work (Dorfman and Alf, 1969). The least-squares method is a quantitative way of accomplishing the same aim. If  $(x_t, y_t)$  are the data points, one constructs  $S$ , the sum of the squared deviations of the observed ordinates from the predicted values (since  $R$  is the number of ratings bins, the summation runs over the  $R - 1$  operating points):

$$S = \sum_{i=1}^{R-1} (y_i - (a + bx_i))^2$$

The idea is to minimize  $S$  with respect to the parameters  $(a, b)$ . One approach is to differentiate this with respect to  $a$  and  $b$  and equate each resulting derivative expression to zero. This yields two equations in two unknowns, which are solved

for  $a$  and  $b$ . If the reader has never done this before, one should go through these steps at least once, but it would be smarter in future to use software that does all this. In R the least-squares fitting function is `lm(y~x)`, which in its simplest form fits a linear model `lm(y~x)` using the method of least-squares (in case you are wondering `lm` stands for linear model, a whole branch of statistics in itself; in this example one is using its simplest capability).

```
# ML estimates of a and b (from Eng JAVA program)
# a <- 1.3204; b <- 0.6075
# # these are not used in program; just here for comparison

FPF <- c(0.017, 0.050, 0.183, 0.5)
# this is from Table 6.11, last two rows
TPF <- c(0.440, 0.680, 0.780, 0.900)
# ...do...

PhiInvFPF <- qnorm(FPF)
# apply the PHI_INV function
PhiInvTPF <- qnorm(TPF)
# ... do ...

fit <- lm(PhiInvTPF~PhiInvFPF)
print(fit)
#>
#> Call:
#> lm(formula = PhiInvTPF ~ PhiInvFPF)
#>
#> Coefficients:
#> (Intercept)  PhiInvFPF
#>     1.328844    0.630746
```

Fig. 13.5 shows operating points from Table 11.1, transformed by the  $\Phi^{-1}$  function; the slope of the line is the least-squares estimate of the  $b$  parameter and the intercept is the corresponding  $a$  parameter of the binormal model.

The last line contains the least squares estimated values,  $a = 1.3288$  and  $b = 0.6307$ . The corresponding maximum likelihood estimates of these parameters, as yielded by the Eng web code, Appendix B, are listed in line 4 of the main program:  $a = 1.3204$  and  $b = 0.6075$ . The estimates appear to be close, particularly the estimate of  $a$ , but there are a few things wrong with the least-squares approach. First, the method of least squares assumes that the data points are independent. Because of the manner in which they are constructed, namely by cumulating points, the independence assumption is not valid for ROC operating points. Cumulating the 4 and 5 responses constrains the resulting operating point to be above and to the right of the point obtained by cumulating the 5 responses only, so the data points are definitely not independent. Similarly,

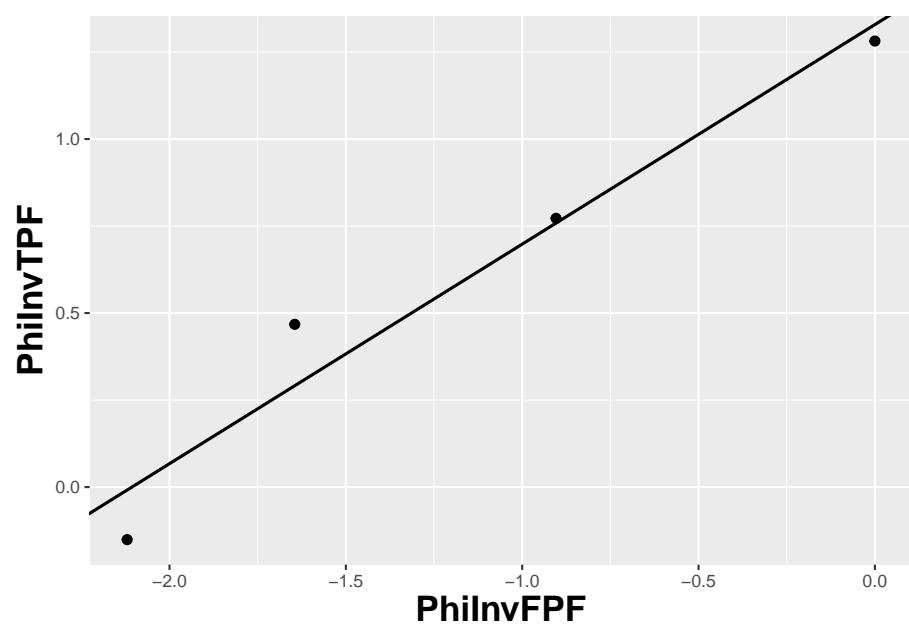


Figure 13.5: The straight line fit method of estimating parameters of the fitting model.

cumulating the 3, 4 and 5 responses constrains the resulting operating point to be above and to the right of the point obtained by cumulating the 4 and 5 responses, and so on. The second problem is the linear least-squares method assumes there is no error in measuring  $x$ ; the only source of error that is accounted for is in the  $y$ -coordinate. In fact, both coordinates of an ROC operating point are subject to sampling error. Third, disregard of error in the  $x$ -direction is further implicit in the estimates of the thresholds, which according to Eqn. (6.2.19), is given by:

$$\zeta_r = -\Phi^{-1}(FPF_r)$$

These are “rigid” estimates that assume no error in the FPF values. As was shown in Chapter 9, 95% confidence intervals apply to these estimates.

A historical note: prior to computers and easy access to statistical functions the analyst had to use a special plotting paper, termed “double probability paper”, that converted probabilities into  $x$  and  $y$  distances using the inverse function.

### 13.14.3 Maximum likelihood estimation (MLE)

The approach taken by Dorfman and Alf was to maximize the likelihood function instead of  $S$ . The likelihood function is the probability of the observed data given a set of parameter values, i.e.,

$$L \equiv P(\text{data} | \text{parameters})$$

Generally “data” is suppressed, so likelihood is a function of the parameters; but “data” is always implicit. With reference to Fig. 6.1, the probability of a non-diseased case yielding a count in the 2nd bin equals the area under the curve labeled “Noise” bounded by the vertical lines at  $\zeta_1$  and  $\zeta_2$ . In general, the probability of a non-diseased case yielding a count in the  $r^{\text{th}}$  bin equals the area under the curve labeled “Noise” bounded by the vertical lines at  $\zeta_{r-1}$  and  $\zeta_r$ . Since the area to the left of a threshold is the CDF corresponding to that threshold, the required probability is  $\Phi(\zeta_r) - \Phi(\zeta_{r-1})$ ; we are simply subtracting two expressions for specificity, Eqn. (6.2.5).

$$\text{count in non-diseased bin } r = \Phi(\zeta_r) - \Phi(\zeta_{r-1})$$

Similarly, the probability of a diseased case yielding a count in the  $r^{\text{th}}$  bin equals the area under the curve labeled “Signal” bounded by the vertical lines at  $\zeta_{r-1}$  and  $\zeta_r$ . The area under the diseased distribution to the left of threshold  $\zeta_r$  is the  $1 - TPF$  at that threshold:

$$1 - \Phi\left(\frac{\mu - \zeta_r}{\sigma}\right) = \Phi\left(\frac{\zeta_r - \mu}{\sigma}\right)$$

The area between the two thresholds is:

$$\begin{aligned} P(\text{count in diseased bin } r) &= \Phi\left(\frac{\zeta_r - \mu}{\sigma}\right) - \Phi\left(\frac{\zeta_{r-1} - \mu}{\sigma}\right) \\ &= \Phi(b\zeta_r - a) - \Phi(b\zeta_{r-1} - a) \end{aligned}$$

Let  $K_{1r}$  denote the number of non-diseased cases in the  $r$ th bin, and  $K_{2r}$  denotes the number of diseased cases in the  $r$ th bin. Consider the number of counts  $K_{1r}$  in non-diseased case bin  $r$ . Since the probability of each count is  $\Phi(\zeta_{r+1}) - \Phi(\zeta_r)$ , the probability of the observed number of counts, assuming the counts are independent, is  $(\Phi(\zeta_{r+1}) - \Phi(\zeta_r))^{K_{1r}}$ . Similarly, the probability of observing counts in diseased case bin  $r$  is  $(\Phi(b\zeta_{r+1} - a) - \Phi(b\zeta_r - a))^{K_{2r}}$ , subject to the same independence assumption. The probability of simultaneously observing  $K_{1r}$  counts in non-diseased case bin  $r$  and  $K_{2r}$  counts in diseased case bin  $r$  is the product of these individual probabilities (again, an independence assumption is being used):

$$(\Phi(\zeta_{r+1}) - \Phi(\zeta_r))^{K_{1r}} (\Phi(b\zeta_{r+1} - a) - \Phi(b\zeta_r - a))^{K_{2r}}$$

Similar expressions apply for all integer values of  $r$  ranging from  $1, 2, \dots, R$ . Therefore the probability of observing the entire data set is the product of expressions like Eqn. (6.4.5), over all values of  $r$ :

$$\prod_{r=1}^R \left[ (\Phi(\zeta_{r+1}) - \Phi(\zeta_r))^{K_{1r}} (\Phi(b\zeta_{r+1} - a) - \Phi(b\zeta_r - a))^{K_{2r}} \right] \quad (13.52)$$

We are almost there. A specific combination of  $K_{11}, K_{12}, \dots, K_{1R}$  counts from  $K_1$  non-diseased cases and counts  $K_{21}, K_{22}, \dots, K_{2R}$  from  $K_2$  diseased cases can occur the following number of times (given by the multinomial factor shown below):

$$\frac{K_1!}{\prod_{r=1}^R K_{1r}!} \frac{K_2!}{\prod_{r=1}^R K_{2r}!} \quad (13.53)$$

The likelihood function is the product of Eqn. (13.52) and Eqn. (13.53):

$$\begin{aligned} L(a, b, \vec{\zeta}) &= \left( \frac{K_1!}{\prod_{r=1}^R K_{1r}!} \frac{K_2!}{\prod_{r=1}^R K_{2r}!} \right) \times \\ &\quad \prod_{r=1}^R \left[ (\Phi(\zeta_{r+1}) - \Phi(\zeta_r))^{K_{1r}} (\Phi(b\zeta_{r+1} - a) - \Phi(b\zeta_r - a))^{K_{2r}} \right] \end{aligned} \quad (13.54)$$

The left hand side of Eqn. (13.54) shows explicitly the dependence of the likelihood function on the parameters of the model, namely  $a, b, \vec{\zeta}$ , where the vector of thresholds  $\vec{\zeta}$  is a compact notation for the set of thresholds  $\zeta_1, \zeta_2, \dots, \zeta_R$ , (note that since  $\zeta_0 = -\infty$ , and  $\zeta_R = +\infty$ , only  $R - 1$  free threshold parameters are involved, and the total number of free parameters in the model is  $R + 1$ ). For example, for a 5-rating ROC study, the total number of free parameters is 6, i.e.,  $a, b$  and 4 thresholds  $\zeta_1, \zeta_2, \zeta_3, \zeta_4$ .

Eqn. (13.54) is forbidding but here comes a simplification. The difference of probabilities such as  $\Phi(\zeta_r) - \Phi(\zeta_{r-1})$  is guaranteed to be positive and less than one [the  $\Phi$  function is a probability, i.e., in the range 0 to 1, and since  $\zeta_r$  is greater than  $\zeta_{r-1}$ , the difference is positive and less than one]. When the difference is raised to the power of  $K_{1r}$  (a non-negative integer) a very small number can result. Multiplication of all these small numbers may result in an even smaller number, which may be too small to be represented as a floating-point value, especially as the number of counts increases. To prevent this we resort to a trick. Instead of maximizing the likelihood function  $L(a, b, \vec{\zeta})$  we choose to maximize the logarithm of the likelihood function (the base of the logarithm is immaterial). The logarithm of the likelihood function is:

$$LL(a, b, \vec{\zeta}) = \log(L(a, b, \vec{\zeta})) \quad (13.55)$$

Since the logarithm is a monotonically increasing function of its argument, maximizing the logarithm of the likelihood function is equivalent to maximizing the likelihood function. Taking the logarithm converts the product symbols in Eqn. (6.4.8) to summations, so instead of multiplying small numbers one is adding them, thereby avoiding underflow errors. Another simplification is that one can ignore the logarithm of the multinomial factor involving the factorials, because these do not depend on the parameters of the model. Putting all this together, we get the following expression for the logarithm of the likelihood function:

$$\begin{aligned} LL(a, b, \vec{\zeta}) &\propto \sum_{r=1}^R K_{1r} \log(\Phi(\zeta_{r+1}) - \Phi(\zeta_r)) \\ &+ \sum_{r=1}^R K_{2r} \log(\Phi(b\zeta_{r+1} - a) - \Phi(b\zeta_r - a)) \end{aligned} \quad (13.56)$$

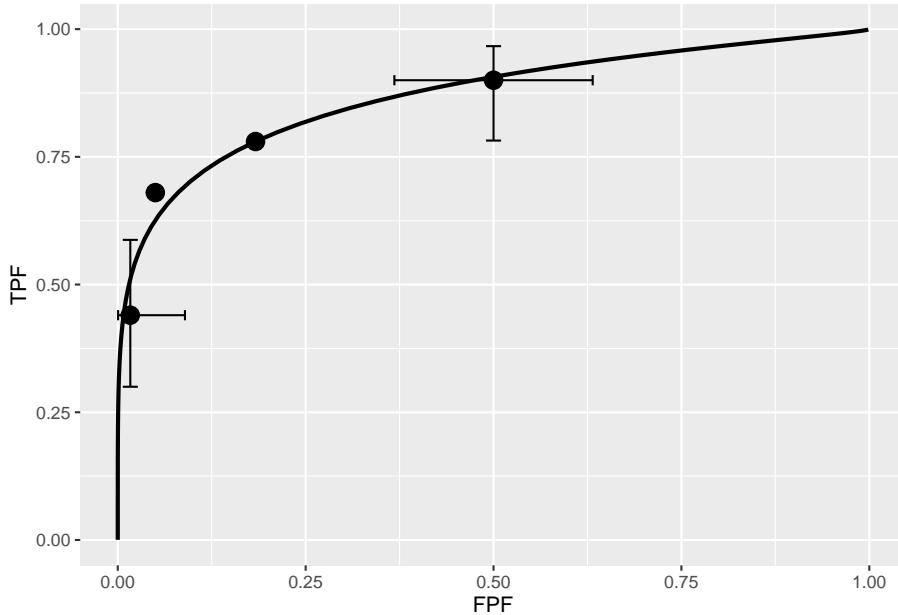
The left hand side of Eqn. (13.56) is a function of the model parameters  $a, b, \vec{\zeta}$  and the observed data, the latter being the counts contained in the vectors  $\vec{K}_1$  and  $\vec{K}_2$ , where the vector notation is used as a compact form for the counts  $K_{11}, K_{12}, \dots, K_{1R}$  and  $K_{21}, K_{22}, \dots, K_{2R}$ , respectively. The right hand side of Eqn. (13.56) is monotonically related to the probability of observing the data given the model parameters  $a, b, \vec{\zeta}$ . If the choice of model parameters is poor, then the probability of observing the data will be small and log likelihood will

be small. With a better choice of model parameters the probability and log likelihood will increase. With optimal choice of model parameters the probability and log likelihood will be maximized, and the corresponding optimal values of the model parameters are called maximum likelihood estimates (MLEs). These are the estimates produced by the programs RSCORE and ROCFIT.

#### 13.14.4 Code implementing MLE

```
# ML estimates of a and b (from Eng JAVA program)
# a <- 1.3204; b <- 0.6075
# these are not used in program; just there for comparison

K1t <- c(30, 19, 8, 2, 1)
K2t <- c(5, 6, 5, 12, 22)
dataset <- Df2RJafrocDataset(K1t, K2t, InputIsCountsTable = TRUE)
retFit <- FitBinormalRoc(dataset)
retFit[1:5]
#> $a
#> [1] 1.32045261
#>
#> $b
#> [1] 0.607492932
#>
#> $zetas
#>      zetaFwd1      zetaFwd2      zetaFwd3      zetaFwd4
#> 0.00768054675 0.89627306763 1.51564784976 2.39672209865
#>
#> $AUC
#> [1] 0.870452157
#>
#> $StdAUC
#>           [,1]
#> [1,] 0.0379042262
print(retFit$fittedPlot)
```



Note the usage of the **RJafroc** package (Chakraborty et al., 2020b). Specifically, the function **FitBinormalRoc**. The ratings table is converted to an **RJafroc** dataset object, followed by application of the fitting function. The results, contained in **retFit** should be compared to those obtained from the website implementation of ROCFIT.

## 13.15 Appendix V: Validating fitting model

The above ROC curve is a good visual fit to the observed operating points. Quantification of the validity of the fitting model is accomplished by calculating the Pearson goodness-of-fit test (Pearson, 1900), also known as the chi-square test, which uses the statistic defined by (Larsen and Marx, 2001):

$$C^2 = \sum_{t=1}^2 \sum_{r=1}^R \frac{(K_{tr} - \langle K_{tr} \rangle)^2}{\langle K_{tr} \rangle} K_{tr} \geq 5 \quad (13.57)$$

The expected values are given by:

$$\begin{aligned} \langle K_{1r} \rangle &= K_1 (\Phi(\zeta_{r+1}) - \Phi(\zeta_r)) \\ \langle K_{2r} \rangle &= K_2 (\Phi(a\zeta_{r+1} - b) - \Phi(a\zeta_r - b)) \end{aligned} \quad (13.58)$$

These expressions should make sense: the difference between the two CDF functions is the probability of a count in the specified bin, and multiplication by the total number of relevant cases should yield the expected counts (a non-integer).

It can be shown that under the null hypothesis that the assumed probability distribution functions for the counts equals the true probability distributions, i.e., the model is valid, the statistic  $C^2$  is distributed as:

$$C^2 \sim \chi_{df}^2 \quad (13.59)$$

Here  $C^2 \sim \chi_{df}^2$  is the chi-square distribution with degrees of freedom  $df$  defined by:

$$df = (R - 1) + (R - 1) - (2 + R - 1) = (R - 3) \quad (13.60)$$

The right hand side of the above equation has been written in an expansive form to illustrate the general rule: for  $R$  non-diseased cells in the ratings table, the degree of freedom is  $R - 1$ : this is because when all but one cells are specified, the last is determined, because they must sum to  $K_1$ . Similarly, the degree of freedom for the diseased cells is also  $R - 1$ . Last, we need to subtract the number of free parameters in the model, which is  $(2 + R - 1)$ , i.e., the  $a, b$  parameters and the  $R - 1$  thresholds. It is evident that if  $R = 3$  then  $df = 0$ . In this situation, there are only two non-trivial operating points and the straight-line fit shown will pass through both of them. With two basic parameters, fitting two points is trivial, and goodness of fit cannot be calculated.

Under the null hypothesis (i.e., model is valid)  $C^2$  is distributed as  $\chi_{df}^2$ . Therefore, one computes the probability that this statistic is larger than the observed value, called the *p-value*. If this probability is very small, that means that the deviations of the observed values of the cell counts from the expected values are so large that it is unlikely that the model is correct. The degree of unlikeliness is quantified by the p-value. Poor fits lead to small p values.

At the 5% significance level, one concludes that the fit is not good if  $p < 0.05$ . In practice one occasionally accepts smaller values of  $p$ ,  $p > 0.001$  before completely abandoning a model. It is known that adoption of a stricter criterion, e.g.,  $p > 0.05$ , can occasionally lead to rejection of a retrospectively valid model (Press et al., 2007).

### 13.15.1 Estimating the covariance matrix

TBA See book chapter 6.4.3. This is implemented in `RJafroc`.

### 13.15.2 Estimating the variance of Az

TBA See book chapter 6.4.4. This is implemented in `RJafroc`.

### **13.16 References**



# Chapter 14

## Sources of AUC variability

### 14.1 TBA How much finished

60%

### 14.2 Introduction

In previous chapters the area AUC under the ROC plot was introduced as the preferred way of summarizing performance in the ROC task, as compared to a pair of sensitivity and specificity values. It can be estimated either non-parametrically, as in Chapter 12, or parametrically, as in Chapter 13, and even better ways of estimating it are described in TBA Chapter 18 and Chapter 20.

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

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

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

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

### 14.3 Three sources of variability

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

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

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

$$A_{z\{c\}} \sim N(A_{z\{\bullet\}}, \sigma_{cs+wr}^2) \quad (14.1)$$

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

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

$$A_{z;j} \sim N(A_{z;\bullet}, \sigma_{br+wr}^2) \quad (14.2)$$

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

3. Consider a fixed reader, e.g.,  $j = 1$ , interpreting a fixed case-sample  $\{1\}$ . Within-reader variability is due to variability of the ratings for the same case: the same reader interpreting the same case on different occasions will give different ratings to it, causing fluctuations in the measured AUC. This assumes that memory effects are minimized, for example, by sufficient time between successive interpretations as otherwise, if a case is

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

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

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

## 14.4 Dependence of AUC on the case sample

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

$$\text{AUC} \rightarrow \text{AUC}_{\{c\}} \quad (14.3)$$

Note that  $\{c\}$  is not an individual *case* index, rather it is a *case-set* index, i.e., different integer values of  $c$  denote different sets, or samples, or groups, or

collections of cases. [The dependence of AUC on the case sample index is not explicitly shown in the literature.]

What does the dependence of AUC on the *c* index mean? Different case samples differ in their *difficulty* levels. A difficult case set contains a greater fraction of difficult cases than is usual. A difficult diseased case is one where disease is difficult to detect. For example, the lesions could be partly obscured by overlapping normal structures in the patient anatomy; i.e., the lesion does not “stick out”. Alternatively, variants of normal anatomy could mimic a lesion, like a blood vessel viewed end on in a chest radiograph, causing the radiologist to miss the real lesion(s) and mistake these blood vessels for lesions. An easy diseased case is one where the disease is easy to detect. For example, the lesion is projected over smooth background tissue, because of which it “sticks out”, or is more conspicuous<sup>2</sup>. How does difficulty level affect non-diseased cases? A difficult non-diseased case is one where variants of normal anatomy mimic actual lesions and could cause the radiologist to falsely diagnose the patient as diseased. Conversely, an easy non-diseased case is like a textbook illustration of normal anatomy. Every structure in it is clearly visualized and accounted for by the radiologist’s knowledge of the patient’s non-diseased anatomy, and the radiologist is confident that any abnormal structure, *if present*, would be readily seen. The radiologist is unlikely to falsely diagnose the patient as diseased. Difficult cases tend to be rated in the middle of the rating scale, while easy ones tend to be rated at the ends of the rating scale.

#### 14.4.1 Case sampling variability of AUC

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

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

20% in screening mammography<sup>3</sup>. This is where follow-up imaging can help determine true disease status at the initial screen. A false negative mistake is unlikely to be repeated at the next screen. After a year, the tumor may have grown, and is more likely to be detected. Having detected the tumor in the most recent screen, radiologists can go back and retrospectively view it in the initial screen, at which it was missed during the “live” interpretation. If one knows where to look, the cancer is easier to see. The previous screen images would be an example of a difficult diseased case. In unfortunate instances, the patient may die from the previously undetected cancer, which would establish the truth status at the initial screen, too late to do the patient any good. The process of determining actual truth is often referred to as defining the “gold standard”, the *ground truth*: or simply *truthing*.

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

There has to be a better way of estimating case-sampling variability. With a parametric model, the maximum likelihood procedure provides a means of estimating variability of each of the estimated parameters, which can be used to estimate the variability of  $A_z$ , as in Chapter 13. The estimate corresponds to case-sampling variability (including an inseparable within-reader variability). If unsure about this point, the reader should run some of the examples in Chapter 13 with increased numbers of cases. The variability is seen to decrease.

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

## 14.5 DeLong method

If the figure-of-merit is the empirical AUC, then a procedure developed by DeLong et al<sup>4</sup> (henceforth abbreviated to DeLong) is applicable that is based on earlier work by (Noether, 1967) and (Bamber, 1975). The author will not go into details of this procedure but limit to showing that it “works”. However, before one can show that it “works”, one needs to know the true value of the variance of empirical AUC. Even if data were simulated using the binormal model, one cannot use the binormal model based estimate of variance as it is an estimate, not to be confused with a true value. Estimates are realizations of random numbers and are themselves subject to variability, which decreases with increasing case-set size. Instead, a “brute-force” (i.e., simulated population sampling) approach is adopted to determine the true value of the variance of AUC. The simulator provides a means of repeatedly generating case-sets interpreted by the same radiologist, and by sampling it enough time, e.g.,  $C = 10,000$  times, each time calculating AUC, one determines the population mean and standard deviation.

The standard deviation determined this way is compared to that yielded by the DeLong method to check if the latter actually works.

```

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

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

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

```

```
#> empirical AUC 0.8626923
#> standard deviation DeLong = 0.03804135
```

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

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

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

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

```

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

```

1. An important observation is that as sample-size increases, case-sampling variability decreases: 0.0417 for the smaller sample size vs. 0.01309 for the larger sample size, and the dependence is as the inverse square root of the numbers of cases, as expected from the central limit theorem.
2. With the smaller sample size ( $K_1/K_2 = 50/52$ ; the back-slash notation, not to be confused with division, is a convenient way of summarizing the case-sample size) the estimated standard deviation (0.038) is within 10% of that estimated by population sampling (0.042). With the larger sample size, ( $K_1/K_2 = 500/520$ ) the two are practically identical (0.01300203 vs. 0.01356696 – the latter value is for seed = 2).
3. Notice also that the one sample empirical AUC for the smaller case-size is 0.863, which is less than two standard deviations from the population mean 0.819. The “two standard deviations” comes from rounding up 1.96: as in Eqn. (10.32), where  $z_{\alpha/2}$  was defined as the upper  $1 - \alpha/2$  quantile of the unit normal distribution and  $z_{0.025} = 1.96$ .
4. To reiterate, with clinical data the DeLong procedure estimates case sampling plus within reader variability. With simulated data as in this example, there is no within-reader variability as the simulator yields identical values for fixed seed.

This demonstration should convince the reader that one does have recourse other than the “brute force” method, at least when the figure of merit is the empirical area under the ROC. That should come as a relief, as population sampling is impractical in the clinical context. It should also impress the reader, as the DeLong method is able to use information present in a *single dataset* to tease out its variability. [This is not magic: the MLE estimate is also able to tease out variability based on a parametric fit to a single dataset and examination of the sharpness of the peak of the log-likelihood function, Chapter 13, as are the resampling methods described next.]

Next, two resampling-based methods of estimating case-sampling variance of AUC are introduced. The word “resampling” means that the dataset itself is regarded as containing information regarding its variability, which can be extracted by sampling from the original data (hence the word “resampling”). These are general and powerful techniques, applicable to any scalar statistic, not just the empirical AUC, which one might be able to use in other contexts.

Table 14.1: Representative counts table.

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

## 14.6 Bootstrap method

The simplest resampling method, at least at the conceptual level, is the bootstrap. *The bootstrap method is based on the assumption that one can regard the observed sample as defining the population from which it was sampled.* Since by definition a population cannot be exhausted, the idea is to resample, *with replacement*, from the observed sample. Each resampling step realizes a particular bootstrap sample set denoted  $\{b\}$ , where  $b = 1, 2, \dots, B$ . The curly brackets emphasize that different integer values of  $b$  denote different *sets of cases*, not individual cases. [In contrast, the notation  $(k)$  will be used to denote *removing* a specific case,  $k$ , as in the jackknife procedure to be described shortly. The index  $b$  should not be confused with the index  $c$ , the case sampling index; the latter denotes repeated sampling from the population, which is impractical in real life; the bootstrap index denotes repeated sampling from the dataset, which is quite feasible.] The procedure is repeated  $B$  times, typically  $B$  can be as small as 200, but to be safe I generally use about 1000 - 2000 bootstraps. The following example uses Table 11.1 from Chapter 11.

For convenience, let us denote cases as follows. The 30 non-diseased cases that received the 1 rating are denoted  $k_{1,1}, k_{2,1}, \dots, k_{30,1}$ . The second index denotes the truth state of the cases. Likewise, the 19 non-diseased cases that received the 2 rating are denoted  $k_{31,1}, k_{32,1}, \dots, k_{49,1}$  and so on for the remaining non-diseased cases. The 5 diseased cases that received the 1 rating are denoted  $k_{1,2}, k_{2,2}, \dots, k_{5,2}$ , the 6 diseased cases that received the 2 rating are denoted  $k_{6,2}, k_{7,2}, \dots, k_{11,2}$ , and so on. Let us figuratively “put” all non-diseased cases (think of each case as an index card, with the case notation and rating recorded on it) into one hat (the non-diseased hat) and all the diseased cases into another hat (the diseased hat). Next, one randomly picks one case (card) from the non-diseased hat, records its rating, and puts the case back in the hat, so that it is free to be possibly picked again. This is repeated 60 times for the non-diseased hat resulting in 60 ratings from non-diseased cases. A similar procedure is performed using the diseased hat, resulting in 50 ratings from diseased cases. The author has just described, in painful detail (one might say) the realization of the 1st bootstrap sample, denoted  $\{b = 1\}$ . This is used to construct the 1st bootstrap counts table, Table 14.1.

So what happened? Consider the 35 non-diseased cases with a 1 rating. If each non-diseased case rated 1 in Table 11.1 were picked one time, the total would have been 30, but it is 35. Therefore, some of the original non-diseased cases

rated 1 must have been picked multiple times, but one must also make allowance as there is no guarantee that a specific case was picked at all. Still focusing on the 35 non-diseased cases with a 1 rating in the first bootstrap sample, the picked labels, reordered after the fact, with respect to the first index, might be:

$$k_{2,1}, k_{2,1}, k_{4,1}, k_{4,1}, k_{4,1}, k_{6,1}, k_{7,1}, k_{7,1}, k_{9,1}, \dots, k_{28,1}, k_{28,1}, k_{30,1}, k_{30,1} \quad (14.4)$$

In this example, case  $k_{1,1}$  was not picked, case  $k_{2,1}$  was picked twice, case  $k_{3,1}$  was not picked, case  $k_{4,1}$  was picked three times, case  $k_{5,1}$  was not picked, case  $k_{6,1}$  was picked once, etc. The total number of cases in Eqn. (14.4) is 35, and similarly for the other cells in Table 14.1. Next, one estimates AUC for this table. Using the Eng website referred to earlier, one gets  $AUC = 0.843$ . [It is OK to use a parametric FOM since the bootstrap is a general procedure applicable, in principle, to any FOM, not just the empirical AUC, unlike the DeLong method, which is restricted to empirical AUC.] The corresponding value for the original data, Table 11.1, was  $AUC = 0.870$ . The first bootstrapped dataset yielded a smaller value than the original dataset because one happened to have picked an unusually difficult bootstrap sample.

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

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

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

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

$$\text{Var}(AUC) = \frac{1}{B-1} \sum_{b=1}^B (AUC_{\{b\}} - AUC_{\{\bullet\}})^2 \quad (14.5)$$

The right hand side is the traditional definition of (unbiased) variance. The dot represents the average over the *replaced index*. Of course, running the website code 200 times and recording the outputs is not a productive use of time. The following code implements two methods for estimating AUC, the empirical AUC, described in Chapter 12 and the binormal model estimate of AUC, described in Chapter 13.

#### 14.6.1 Demonstration of the bootstrap method

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

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

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

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

  # ready to bootstrap
  # first put the counts data into a linear array
  # convert counts table to array
  z1 <- rep(1:length(RocTable[1,]),
            RocTable[1,])
  z2 <- rep(1:length(RocTable[2,]),
            RocTable[2,])#do:
  AUC <- array(dim = B)#to save the bs AUC values
  for (b in 1 : B){
```

```

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

```

Since the bootstrap method is applicable to any scalar figure of merit, two options are provided in the code. If `parametricFOM` is set to `TRUE`, then the binormal model estimate is used, and if set to `FALSE`, the empirical AUC is used. The first set of results are obtained with `parametricFOM` set to `TRUE`.

```

parametricFOM <- TRUE
B <- 200;seed <- 1;set.seed(seed)
RocTable = array(dim = c(2,5))

```

```
RocTable[1,] <- c(30,19,8,2,1)
RocTable[2,] <- c(5,6,5,12,22)

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

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

This shows that the AUC of the original data (i.e., before performing any bootstrapping) is 0.870, the mean AUC of the  $B = 200$  bootstrapped datasets is 0.867, and the standard deviation of the 200 bootstraps is 0.0438. If one runs the website calculator referenced in the previous chapter on the dataset shown in Table 11.1, one finds that the MLE of the standard deviation of the AUC of the fitted ROC curve is 0.0378. The standard deviation is itself a statistic and there is sampling variability associated with it, i.e., there exists such a beast as a standard deviation of a standard deviation; the bootstrap estimate is not too far from the MLE estimate. By setting `seed` to different values, one gets an idea of the variability of the estimate of the standard deviation of AUC. For example, with `seed = 2`, one gets:

```
#> Bootstrap variance estimation:
#> parametricFOM = TRUE
#> seed = 2
#> B = 200
#> OrigAUC = 0.8704519
#> meanAUCboot = 0.8673155
#> stdAUCboot = 0.03815402
```

Note that both the mean of the bootstrap samples and the standard deviation have changed, but both are close to the MLE values. Examined next is the

dependence of the estimates on  $B$ , the number of bootstraps. With `seed = 1` and  $B = 2000$  one gets:

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

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

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

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

## 14.7 Jackknife method

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

In this method, the first case is removed, or jackknifed, from the set of cases and the MLE (or empirical estimation) is conducted on the resulting dataset, which

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

$$AUC_{(k)} \quad k = 1, 2, \dots, K \quad (14.6)$$

The corresponding jackknife pseudovalues  $Y_k$  are defined by:

$$Y_k = K \times AUC - (K - 1) \times AUC_{(k)} \quad (14.7)$$

Here AUC denotes the estimate using the entire dataset, i.e., not removing any cases. The jackknife pseudovalues will turn out to be of central importance in TBA Chapter 09. The *jackknife AUC values*, defined by Eqn. (14.6), should not be confused with jackknife derived psuedovalues, defined by Eqn. (14.7).

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

$$\text{Var}_{\text{jack}} = \frac{(K-1)^2}{K} \frac{1}{K-1} \sum_{k=1}^K (AUC_{(k)} - AUC_{(\bullet)})^2 \quad (14.8)$$

Since variance of  $K$  scalars is defined by:

$$\text{Var}(x) = \frac{1}{K-1} \sum_{k=1}^K (x_k - x_{\bullet})^2 \quad (14.9)$$

It follows that:

$$\text{Var}_{\text{jack}}(\text{AUC}) = \frac{(K-1)^2}{K} \text{Var}(\text{AUC}) \quad (14.10)$$

In Eqn. (14.8) I have deliberately not simplified the right hand side by canceling out  $K - 1$ . The purpose is to show, Eqn. (14.10), that the usual expression for the variance (of the jackknife FOM values) needs to be multiplied by a **variance inflation factor**  $\frac{(K-1)^2}{K}$ , which is approximately equal to  $K$ , in order to obtain the correct jackknife estimate of variance of AUC. This factor was not necessary when one used the bootstrap method. That is because the bootstrap samples are more representative of the actual spread in the data. The jackknife samples are more restricted than the bootstrap samples, so the spread of the data is smaller; hence the need for the variance inflation factor (Efron and Tibshirani, 1993).

```

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

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

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

  AUC_jack <- array(dim = sum(K))
  Y_k <- array(dim = sum(K))
  z_jk <- array(dim = sum(K))
  # ready to jackknife
  for ( k in 1 : sum(K)) {
    RocTable_jk <- array(dim = c(2,length(RocTable[1,])))
    if ( k <= K[ 1 ]){
      z1_jk <- z1[ -k ]
      z2_jk <- z2
    } else{
      z1_jk <- z1
      z2_jk <- z2[ -(k - K[ 1 ]) ]
    }
    #convert array to frequency table
    RocTable_jk[1,1:length(table(z1_jk))] <-
      table(z1_jk)
    RocTable_jk[2,1:length(table(z2_jk))] <-
      table(z2_jk)
    #replace NAs with zeroes
    RocTable_jk[is.na(RocTable_jk)] <- 0
    # AUC_jack for observed data
    if (parametricFOM) {
      temp <- RocfitR(RocTable_jk)
    } else {
      temp <- WilcoxonCountsTable(RocTable_jk)
    }
    AUC_jack[k] <- temp$AUC
    Y_k[k] <- sum(K)*OrigAUC - (sum(K)-1)*AUC_jack[k]
  }
}

```

```

    if (AUC_jack[k] == -1)
      stop("RocfitR did not converge in jackknife loop")
  }
  meanAUCjack <- mean(AUC_jack)
  #Efron and Stein's paper, include jackknife inflation factor
  Var_jack <- var(AUC_jack) * ( sum(K) - 1)^2 / sum(K)
  stdAUCjack <- sqrt(Var_jack)
  return(list(
    OrigAUC = OrigAUC,
    meanAUCjack = meanAUCjack,
    stdAUCjack = stdAUCjack
  ))
}

```

Since the jackknife method is applicable to any scalar figure of merit, two options are provided in the code. If `parametricFOM` is set to `TRUE`, then the binormal model estimate is used, and if set to `FALSE`, the empirical AUC is used. The first set of results are obtained with `parametricFOM` set to `TRUE`. Notice that the code does not use a `set.seed()` statement, as no random number generator is needed in the jackknife method. Systematically removing and replacing each case in sequence, one at a time, is not random sampling, which should further explain the need for the variance inflation factor in Eqn. (14.10).

```

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

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

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

```

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

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

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

## 14.8 Calibrated simulator

### 14.8.1 The need for a calibrated simulator

The population sampling method used previously, 14.5, to compare the DeLong method to a known standard used arbitrarily set simulator values, i.e.,  $\mu = 1.5$  and  $\sigma = 1.3$ . One does not know if these values actually represent real clinical data. In this section a simple method of implementing population sampling using a *calibrated simulator* is described. A calibrated simulator is one whose parameters are chosen to match those of an actual clinical dataset. This way one has some assurance that the simulator is realistic and therefore its verdict on a proposed method or analysis (in our case method of estimating AUC variability) is likely to be correct.

### 14.8.2 Implementation of a simple calibrated simulator

The simple simulator described here is limited to a single reader single modality dataset. A more complex simulator describing multiple readers in multiple modalities is described in a later chapter (TBA). Consider a clinical dataset, such as in Table 11.1. Analyzed by MLE, this yields binormal model parameters,  $a$ ,  $b$  and the thresholds  $\zeta_1, \zeta_2, \zeta_3, \zeta_4$ . After conversion to  $\mu = a/b$  and  $\sigma = 1/b$  and new zetas  $\zeta = \zeta/b$ , the values are (in the same order): 2.173597, 1.646099, 0.01263423, 1.475351, 2.494901, 3.945221 (see code output below):

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

```
#> [1] 1.646099
ab[3:6]/ab[2] # this is zeta in mu-sigma notation
#> [1] 0.01263423 1.47535099 2.49490121 3.94522113
```

[The reason for dividing  $\zeta$  by  $b$  is that when re-scaling the decision variable axis by  $b$  one must also re-scale the cutoffs.] The values  $\mu, \sigma, \zeta$  define the calibrated simulator, in the sense that the parameter values are calibrated to match the dataset in Table 11.1.

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

```
1 doCalSimulator <- function(P, parametricFOM, RocCountsTable) {
2   K <- c(sum(RocCountsTable[1,]), 
3          sum(RocCountsTable[2,]))
4   # perform the initial calibration
5   ret <- RocfitR(RocCountsTable) # AUC for observed data
6   a <- ret$a
7   b <- ret$b
8   zetas <- ret$zeta
9   mu <- a/b
10  sigma <- 1/b
11  zetas <- zetas/b # need to also scale zetas
12  # AUC for observed data
13  if (parametricFOM) {
14    OrigAUC <- RocfitR(RocCountsTable)$AUC
15  } else {
16    OrigAUC <- WilcoxonCountsTable(RocCountsTable)$AUC
17  }
18  # perform the population sampling
19  AUC <- array(dim = P)
20  for (p in 1 : P){
21    while (1) {
22
23      RocCountsTableSimPop <-
24        SimulateRocCountsTable(K, mu, sigma, zetas)
25      if (parametricFOM) {
26
27        # AUC for fitted curve
28        temp <- RocfitR(RocCountsTableSimPop)
29        # a return of -1 means RocFitR did not converge
30        if (temp[1] != -1) {
31          AUC[p] <- temp$AUC
32          break
33        }
```

```

34     } else {
35         AUC[p] <- (WilcoxonCountsTable(RocCountsTableSimPop))$AUC
36         break
37     }
38 }
39 }
40 AUC <- AUC[!is.na(AUC)]
41 meanAUC <- mean(AUC)
42 stdAUC <- sqrt(var(AUC))
43 return(list(
44     mu = mu, # these define the calibration simulator
45     sigma = sigma, #do:
46     zetas = zetas, #do:
47     OrigAUC = OrigAUC,
48     meanAUC = meanAUC,
49     stdAUC = stdAUC
50 ))
51 }

```

In the function `doCalSimulator(P, parametricFOM, RocCountsTable)`, `P` is the desired number of population samples, `parametricFOM` is a logical, if set to TRUE the binormal model is used to calculate *fitted* AUC and otherwise the Wilcoxon statistic is used to calculate *empirical* AUC, and `RocCountsTable` contains the ROC data, such as Table 11.1, to which the simulator is to be calibrated to. Lines 2-3 construct the K-vector, containing  $K_1, K_2$ . Line 5 performs the maximum likelihood fit, using function `RocfitR(RocCountsTable)`. The returned variable contains  $a, b, \zeta$  as a `list`, which are extracted at lines 6-8. Lines 9-11 converts these to the mu-sigma notation. In essence, lines 5 - 11 calibrates the simulator and the calibrated values of the simulator are contained in  $\mu, \sigma, \zeta$ . Lines 13-17 calculates `OrigAUC`, the AUC of the original data, using parametric `RocfitR` or the Wilcoxon statistic, as appropriate, depending on the value of `parametricFOM`. After defining a length `P` array, at line 19, to hold the sampled AUC values, lines 20-39 begins and ends a `for` loop to conduct the `P` population samples. Each pass through the `for` loop yields  $K_1$  samples from the non-diseased distribution and  $K_2$  samples from the diseased distribution, returned in the variable `RocCountsTableSimPop`, which is similar in structure to a counts table like Table 11.1. Within the `for` loop there is an endless `while` loop, needed because `RocfitR` can sometimes fail to converge, signaled by the first member of the returned `list` being minus 1, in which case another iteration of the `while` loop is performed (see line 30) and otherwise the `break` statement (line 32) causes program execution to proceed to the next iteration of the `for` loop. After entering the `while` loop, lines 22-23, a new ROC counts table is generated. The returned `list` is saved to `temp` at line 28, and if `temp[1] != -1` (i.e., `RocfitR` did converge) the AUC value is saved to `AUC[p]`, line 31. Upon exiting the code one has `P` values of AUC in the array `AUC`.

#### 14.8.2.1 Parametric AUC results

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

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

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

The simulations were repeated with `seed = 2`. This time the mean and standard deviation of the 2000 AUC values, are equal to 0.8681855 and 0.0405516, respectively. The respective values corresponding to the two `seed` values are quite close to each other (to within a percent).

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

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

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

#### 14.8.2.2 Non-parametric AUC results

The next simulation is with `seed = 1` and `P = 2000`, but this time `parametricFOM` is set to `FALSE`. The calibration proceeds as before, using `RocfitR` to determine the parameters of the simulation model, calibrating the

simulator requires a parametric fit, but this empirical AUC is used to obtain the 2000 AUC samples. The mean and standard deviation of the AUC values, are 0.8497634 and 0.0367476, respectively. Note that these are smaller than the corresponding parametric estimates. The empirical AUC is expected to be smaller than the corresponding parametric AUC as joining adjacent points with straight lines will underestimate the area under the smooth ROC curve. Repeating with `seed = 2`, the mean and standard deviation of the AUC values, are 0.8503732 and 0.0369091, respectively, which are close to the `seed = 1` values.

## 14.9 Discussion

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

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

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

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

TBA: what to do with removed sections?

## **14.10 References**

# **Significance Testing**



# Chapter 15

# Hypothesis Testing

## 15.1 TBA How much finished

60%

## 15.2 Introduction

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

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

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

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

### 15.3 Single-modality single-reader ROC study

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

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

The following code chunk encodes the `Wilcoxon` function:

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

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

```

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

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

```

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

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

```

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

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

```

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

## 15.4 Type-I errors

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

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

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

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

```

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

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

```

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

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

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

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

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

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

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

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

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

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

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

## 15.5 One vs. two sided tests

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

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

```

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

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

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

df <- data.frame(p2tailed = p2tailed,
                  p1tailedGT = p1tailedGT,
                  p1tailedLT = p1tailedLT)
print(df)
#>   p2tailed p1tailedGT p1tailedLT
#> 1 0.2943993 0.8528004 0.8528004

```

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

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

```

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

```

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

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

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

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

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

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

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

$$\Pr(Z \leq 1.042 | NH) = \Phi(1.042) = 1 - 0.1487 = 0.8513$$

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

## 15.6 Statistical power

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

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

Truth	Fail to reject NH	Reject NH
NH is True	$1 - \alpha$	$\alpha$ (FPF)
NH is False	$\beta$ (FNF)	Power = $1 - \beta$

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

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

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

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

```

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

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

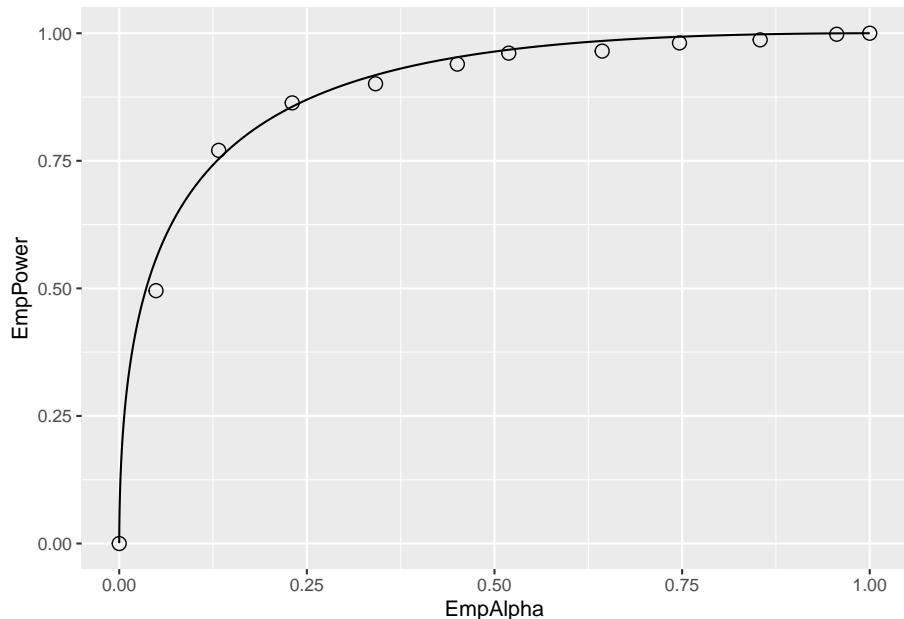
T <- 2000 # Line 16
mu <- c(muNH, muAH) # Line 17
alphaArr <- seq(0.05, 0.95, length.out = 10)
EmpAlpha <- array(dim = length(alphaArr))
EmpPower <- array(dim = length(alphaArr))
for (a in 1:length(alphaArr)) { # Line 20
  alpha <- alphaArr[a]
  reject <- array(0, dim = c(2, T))
  for (h in 1:2) {
 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

```

```

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

```



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

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

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

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

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

### 15.6.1 Factors affecting statistical power

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

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

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

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

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

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

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

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

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

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

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

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

So doubling the numbers of cases (both non-diseased and diseased) results in statistical power increasing from 0.509 to 0.844. Increasing the numbers of cases decreases  $\sigma_{\text{AUC}}$ , the standard deviation of the empirical AUC. The new value of  $\sigma_{\text{AUC}}$  is 0.02947, which should be compared to the value 0.04177 for  $K_1 = 50$ ,  $K_2 = 52$ . Recall that  $\sigma_{\text{AUC}}$  enters the denominator of the Z-statistic, so decreasing it will increase the probability of rejecting the null hypothesis.

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

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

Decreasing  $\alpha$  results in decreased statistical power.

## 15.7 Comments

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

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

## 15.8 Why alpha is chosen as 5%

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

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

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

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

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

## 15.9 Discussion

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

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

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

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

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

### **15.10 References**

# Chapter 16

## DBM method background

### 16.1 TBA How much finished

80%

### 16.2 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., 1992a). The other method, due to Obuchowski and Rockette (Obuchowski and Rockette, 1995a), is the subject of Chapter 10 (TBA). Both methods have been substantially improved by Hillis (Hillis et al., 2008a; Hillis, 2007b, 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.

### 16.2.1 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., 1977b) 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.]

### 16.2.2 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.*

### 16.2.3 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 ( $2 \times 5 \times 200$ ). 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.

#### 16.2.4 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.

### 16.3 Random and fixed factors

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

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

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

tors contribute variability, but the fixed factors do not, rather they contribute constant shifts in performance. The terms *fixed* and *random* factors are used in this specific sense, and are derived, in turn, from ANOVA methods in statistics. With two or more treatments, there are shifts in performance of treatments relative to each other, that one seeks to assess the significance of, against a background of noise contributed by the random factors. If the shifts are sufficiently large compared to the noise, then one can state, with some certainty, that they are real. Quantifying the last statement uses the methods of hypothesis testing introduced in Chapter 15.

## 16.4 Reader and case populations

Consider a sample of  $J$  readers. Conceptually there is a reader-population, modeled as a normal distribution  $\theta_j \sim N(\theta_{\{1\}}, \sigma_{br+wr}^2)$ , describing the variation of skill-level of readers. Here  $\theta$  is a generic FOM. Each reader  $j$  is characterized by a different value of  $\theta_j$ ,  $j = 1, 2, \dots, J$  and one can conceptually think of a bell-shaped curve with variance  $\sigma_{br+wr}^2$  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  $\theta_{\{c\}} \sim N(\theta_{\{\bullet\}}, \sigma_{cs+wr}^2)$  with variance  $\sigma_{cs+wr}^2$  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  $\sigma_{cs}^2$  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., 2008a).

## 16.5 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 (F RCC) 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.*

## 16.6 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  $\theta$ . 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  $\theta_{ij}$  is estimated from the ratings data. Repeating this over all treatments and readers yields a matrix of observed values  $\theta_{ij}$ . 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  $\alpha$ , one declares the treatments different at the  $\alpha$  - significance level. The quantity  $\alpha$  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  $\alpha$  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  $\alpha_{emp}$  over many (typically 2000) independent NH datasets, that the p-value is smaller than  $\alpha$ , should equal  $\alpha$  to within statistical uncertainty.

## 16.7 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 Barrett<sup>19,20</sup>, implemented in open-source JAVA software by Gallas and colleagues<sup>22,44</sup> (<http://didsr.github.io/iMRMC/>) is available on the web. Clarkson et al<sup>19,20</sup> 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 validate<sup>3</sup> DBM is that proposed by Roe and Metz<sup>39</sup> 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 published<sup>3</sup>. (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.

## 16.8 References

# Chapter 17

## Significance Testing using the DBM Method

### 17.1 TBA How much finished

60%

### 17.2 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., 1992a) 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):

$$Y'_{ijk} = K\theta_{ij} - (K-1)\theta_{ij(k)} \quad (17.1)$$

Here  $\theta_{ij}$  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 (17.1) 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., 2008a) 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*):

$$Y_{ijk} = Y'_{ijk} + (\theta_{ij} - Y'_{ij\bullet}) \quad (17.2)$$

**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:

$$Y_{ij\bullet} = Y'_{ij\bullet} + (\theta_{ij} - Y'_{ij\bullet}) = \theta_{ij} \quad (17.3)$$

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):

$$Y_{n(ijk)} = \mu + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + (\tau RC)_{ijk} + \epsilon_{n(ijk)} \quad (17.4)$$

The term  $\mu$  is a constant. By definition, the treatment effect  $\tau_i$  is subject to the constraint:

$$\sum_{i=1}^I \tau_i = 0 \Rightarrow \tau_{\bullet} = 0 \quad (17.5)$$

This constraint ensures that  $\mu$  has the interpretation of the average of the pseudovalue 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  $\tau_i$  is estimated as follows:

$$Y_{ijk} \equiv Y_{1(ijk)}\tau_i = Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet} \quad (17.6)$$

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

### 17.2.1 Explanation of terms in the model

The right hand side of Eqn. (17.1) 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:

$$R_j, C_k, (\tau R)_{ij}, (\tau C)_{ik}, (RC)_{jk}, (\tau RC)_{ijk}, \epsilon_{ijk} \quad (17.7)$$

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

$$\left. \begin{array}{l} R_j \sim N(0, \sigma_R^2) \\ C_k \sim N(0, \sigma_C^2) \\ (\tau R)_{ij} \sim N(0, \sigma_{\tau R}^2) \\ (\tau C)_{ik} \sim N(0, \sigma_{\tau C}^2) \\ (RC)_{jk} \sim N(0, \sigma_{RC}^2) \\ (\tau RC)_{ijk} \sim N(0, \sigma_{\tau RC}^2) \\ \epsilon_{ijk} \sim N(0, \sigma_\epsilon^2) \end{array} \right\} \quad (17.8)$$

Equation (17.8) defines the meanings of the variance components appearing in Equation (17.7). 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 (17.8).

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

The variances defined in (17.8) 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  $\sigma_R^2$  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  $\sigma_{br+wr}^2$  and  $\sigma_{cs+wr}^2$  used in §9.3, which describe the variability of  $\theta$  measured under specified conditions. [A jackknife pseudovalue is a weighted difference of FOM like quantities, TBA (17.1). Its meaning will be explored later. For now, *a pseudovalue variance is distinct from a FOM variance*.]
- The term  $\sigma_C^2$  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  $\sigma_R^2$ . If one were to sample readers and treatments for the same case-set, the net variance would be  $\sigma_R^2 + \sigma_{\tau R}^2 + \sigma_\epsilon^2$ .

- The term  $\sigma_{\tau C}^2$  is the treatment-dependent variance of cases that was excluded in the definition of  $\sigma_C^2$ . So, if one were to sample cases and treatments for the same readers, the net variance would be  $\sigma_C^2 + \sigma_{\tau C}^2 + \sigma_\epsilon^2$ .
- The term  $\sigma_{RC}^2$  is the treatment-independent variance of readers and cases that were excluded in the definitions of  $\sigma_R^2$  and  $\sigma_C^2$ . So, if one were to sample readers and cases for the same treatment, the net variance would be  $\sigma_R^2 + \sigma_C^2 + \sigma_{RC}^2 + \sigma_\epsilon^2$ .
- The term  $\sigma_{\tau RC}^2$  is the variance of treatments, readers and cases that were excluded in the definitions of all the preceding terms in TBA (17.1). So, if one were to sample treatments, readers and cases the net variance would be  $\sigma_R^2 + \sigma_C^2 + \sigma_{\tau C}^2 + \sigma_{RC}^2 + \sigma_{\tau RC}^2 + \sigma_\epsilon^2$ .
- The last term,  $\sigma_\epsilon^2$  describes the variance arising from different replications of the study using the same treatments, readers and cases. Measuring this variance requires repeating the study several ( $N$ ) times with the same treatments, readers and cases, and computing the variance of  $Y_{n(ijk)}$ , where the additional  $n$ -index refers to true replications,  $n = 1, 2, \dots, N$ .

$$\sigma_\epsilon^2 = \frac{1}{IJK} \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K \frac{1}{N-1} \sum_{n=1}^N \left( Y_{n(ijk)} - Y_{\bullet(ijk)} \right)^2 \quad (17.9)$$

The right hand side of TBA (17.1) 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, 1995a).

An equation like TBA (17.1) 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 (17.1), it being understood that for the random effects, one estimates the variances of the zero-mean normal distributions, TBA (17.1)Eqn. (9.7), from which the samples are obtained (by assumption).

Estimating the fixed effects is trivial. The term  $\mu$  is estimated by averaging the left hand side of TBA (17.1)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 (17.1) 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  $\tau_i = Y_{\bullet\bullet\bullet} - \mu$ .

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.

### 17.2.3 Definitions of mean-squares

Again, to be clear, one should 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).

$$\left. \begin{aligned} MST &= \frac{JK \sum_{i=1}^I (Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet})^2}{I-1} \\ MSR &= \frac{IK \sum_{j=1}^J (Y_{\bullet j\bullet} - Y_{\bullet\bullet\bullet})^2}{J-1} \\ MS(C) &= \frac{IJ \sum_{k=1}^K (Y_{\bullet\bullet k} - Y_{\bullet\bullet\bullet})^2}{K-1} \\ MSTR &= \frac{K \sum_{i=1}^I \sum_{j=1}^J (Y_{ij\bullet} - Y_{i\bullet\bullet} - Y_{\bullet j\bullet} + Y_{\bullet\bullet\bullet})^2}{(I-1)(J-1)} \\ MSTC &= \frac{J \sum_{i=1}^I \sum_{k=1}^K (Y_{i\bullet k} - Y_{i\bullet\bullet} - Y_{\bullet\bullet k} + Y_{\bullet\bullet\bullet})^2}{(I-1)(K-1)} \\ MSRC &= \frac{I \sum_{j=1}^J \sum_{k=1}^K (Y_{\bullet jk} - Y_{\bullet j\bullet} - Y_{\bullet\bullet k} + Y_{\bullet\bullet\bullet})^2}{(J-1)(K-1)} \\ MSTRC &= \frac{\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^K (Y_{ijk} - Y_{ij\bullet} - Y_{i\bullet k} - Y_{\bullet jk} + Y_{i\bullet\bullet} + Y_{\bullet j\bullet} + Y_{\bullet\bullet k} - Y_{\bullet\bullet\bullet})^2}{(I-1)(J-1)(K-1)} \end{aligned} \right\} \quad (17.10)$$

Note the absence of  $MSE$ , corresponding to the  $\epsilon$  term on the right hand side of (17.10). With only one observation per treatment-reader-case combination,  $MSE$  cannot be estimated; it effectively gets absorbed into the  $MSTRC$  term.

### 17.3 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 Scheffé [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 (17.10) 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 (17.10). Assuming no replications, i.e.,  $N = 1$ , the expected mean squares are as follows, Table Table 17.1; 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., 1992a).

Table 17.1: Unconstrained expected values of mean-squares, as in (Dorfman et al., 1995)

Source	df	E(MS)
T	(I-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2 + J\sigma_{\tau C}^2 + JK\sigma_\tau^2$
R	(J-1)	$\sigma_\epsilon^2 + I\sigma_{RC}^2 + IK\sigma_R^2 + K\sigma_{\tau R}^2$
C	(K-1)	$\sigma_\epsilon^2 + I\sigma_{RC}^2 + IJ\sigma_C^2 + J\sigma_{\tau C}^2$
TR	(I-1)(J-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2$
TC	(I-1)(K-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2$
RC	(J-1)(K-1)	$\sigma_\epsilon^2 + I\sigma_{RC}^2$
TRC	(I-1)(J-1)(K-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2$
$\epsilon$	$N - 1 = 0$	$\sigma_\epsilon^2$

- In Table 17.1 the following notation is used as a shorthand:

$$\sigma_\tau^2 = \frac{1}{I-1} \sum_{i=1}^I (Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet})^2 \quad (17.11)$$

Since treatment is a fixed effect, the variance symbol  $\sigma_r^2$ , which is used for notational consistency in Table 17.1, 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.

## 17.4 Random-reader random-case (RRRC) analysis

Both readers and cases are regarded as random factors. The expected mean squares in Table Table 17.1 are variance-like quantities; specifically, they are weighted linear combinations of the variances appearing in (17.8). 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  $\epsilon$  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  $\epsilon$  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.

### 17.4.1 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 (17.10) 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[, , k]) - 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[, , ])) + 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

```

### 17.4.2 Significance testing

If the NH of no treatment effect is true, i.e., if  $\sigma_{\tau}^2 = 0$ , then according to Table 17.1 the following holds (the last term in the row labeled  $T$  in Table 17.1 drops out):

$$E(MST | NH) = \sigma_{\epsilon}^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2 + J\sigma_{\tau C}^2 \quad (17.12)$$

Also, the following linear combination is equal to  $E(MST | NH)$ :

$$\begin{aligned}
& E(MSTR) + E(MSTC) - E(MSTRC) \\
&= (\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2) + (\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2) - (\sigma_\epsilon^2 + \sigma_{\tau RC}^2) \\
&= \sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2 + K\sigma_{\tau R}^2 \\
&= E(MST | NH)
\end{aligned} \tag{17.13}$$

Therefore, under the NH, the ratio:

$$\frac{E(MST | NH)}{E(MSTR) + E(MSTC) - E(MSTRC)} = 1 \tag{17.14}$$

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 (17.10):

$$F_{DBM} = \frac{MST}{MSTR + MSTC - MSTRC} \tag{17.15}$$

$F_{DBM}$  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 17.1. 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. (17.15). 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,

$$\left. \begin{aligned} F_{DBM|NH} &\sim F_{\text{ndf}, \text{ddf}} \\ \text{ndf} &= I - 1 \end{aligned} \right\} \quad (17.16)$$

The next topic is estimating  $ddf$ .

#### 17.4.3 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 (17.10)). 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 (17.15) has an approximate degree of freedom defined by (this is called the *Satterthwaite's approximation*):

$$ddf_{Sat} = \frac{(MSTR + MSTC - MSTRC)^2}{\left( \frac{MSTR^2}{(I-1)(J-1)} + \frac{MSTC^2}{(I-1)(K-1)} + \frac{MSTRC^2}{(I-1)(J-1)(K-1)} \right)} \quad (17.17)$$

The subscript *Sat* is for Satterthwaite. From Equation (17.17) it should be fairly obvious that in general  $ddf_{Sat}$  is not an integer. To accommodate possible negative estimates of the denominator of Equation (17.17), the original DBM method (Dorfman et al., 1992a) proposed, depending on the signs of  $\sigma_{\tau R}^2$  and  $\sigma_{\tau C}^2$ , 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., 2008a).

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

$$F_{DBM} = \frac{MST}{MSTR + \max(MSTC - MSTRC, 0)} \quad (17.18)$$

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., 2005a) has shown that the left hand side of Equation (17.18) is distributed as an F-statistic with  $\text{ndf} = I - 1$  and denominator degrees of freedom  $\text{ddf}_H$  defined by:

$$\text{ddf}_H = \frac{(MSTR + \max(MSTC - MSTRC, 0))^2}{\text{MSTR}^2} (I - 1)(J - 1) \quad (17.19)$$

Summarizing,

$$F_{DBM} \sim F_{\text{ndf}, \text{ddf}_H} \quad (17.20)$$

Instead of 4 rules, as in the original DBM method, the Hillis modification involves just one rule, summarized by Equations (17.19) through (17.20). Moreover, the F-statistic is constrained to non-negative values. Using simulation testing (Hillis et al., 2008a) 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  $\alpha = 0.05$ .

#### 17.4.4 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:

$$\Pr(F \leq F_{1-\alpha, \text{ndf}, \text{ddf}_H} \mid F \sim F_{\text{ndf}, \text{ddf}_H}) = 1 - \alpha \quad (17.21)$$

The critical value  $F_{1-\alpha, \text{ndf}, \text{ddf}_H}$  increases as  $\alpha$  decreases. The value of  $\alpha$ , generally chosen to be 0.05, termed the *nominal*  $\alpha$ , 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 (17.18), 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 (17.1) Eqn. (9.23).
- $\alpha$  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  $F_{DBM}$ .
- $\text{ndf}$  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.

- $\text{ddf}_H$  is large: this causes the critical value to decrease, see below, and is more likely to be exceeded by the observed value of  $F_{DBM}$ .

#### 17.4.4.1 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  $F_{\text{ndf}, \text{ddf}}$  that lies to the right of the observed value of  $F_{DBM}$ :

$$p = \Pr(F > F_{DBM} \mid F \sim F_{\text{ndf}, \text{ddf}_H}) \quad (17.22)$$

#### 17.4.4.2 Confidence intervals for inter-treatment FOM differences

If  $p < \alpha$  then the NH that all treatments are identical is rejected at significance level  $\alpha$ . 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., 2005a) has shown that the  $(1 - \alpha)$  confidence interval for  $Y_{i\bullet\bullet} - Y_{i'\bullet\bullet}$  is given by:

$$CI_{1-\alpha} = (Y_{i\bullet\bullet} - Y_{i'\bullet\bullet}) \pm t_{\alpha/2; \text{ddf}_H} \sqrt{\frac{2}{JK} (MSTR + \max(MSTC - MSTRC, 0))} \quad (17.23)$$

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

$$\Pr(T > t_{\alpha/2; \text{ddf}_H}) = \alpha/2 \quad (17.24)$$

Since centered pseudovalues were used:

$$(Y_{i\bullet\bullet} - Y_{i'\bullet\bullet}) = (\theta_{i\bullet} - \theta_{i'\bullet}) \quad (17.25)$$

Therefore, Equation (17.23) can be rewritten:

$$CI_{1-\alpha} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2; \text{ddf}_H} \sqrt{\frac{2}{JK} (MSTR + \max(MSTC - MSTRC, 0))} \quad (17.26)$$

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.

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

Line 1 defines  $\alpha$ . Line 2 forms a data frame from previously calculated mean-squares. Line 3 calculates the denominator appearing in Equation (17.18). Line 4 computes the observed value of  $F_{DBM}$ , namely the ratio of the numerator and denominator in Equation (17.18). Line 5 sets ndf to  $I - 1$ . Line 6 computes  $\text{ddf}_H$ . 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 (17.22). 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

```

#### 17.4.4.4 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 (17.26), 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

```

## 17.5 Sample size estimation for random-reader random-case generalization

### 17.5.1 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 (17.20). *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 (17.15), is distributed as a *non-central* F-distribution  $F_{\text{ndf}, \text{ddf}_H, \Delta}$  with *non-centrality parameter*  $\Delta$ :

$$F_{DBM|AH} \sim F_{\text{ndf}, \text{ddf}_H, \Delta} \quad (17.27)$$

The non-centrality parameter  $\Delta$  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  $\sigma_\tau^2$ ,  $\sigma_{\tau R}^2$  and  $\sigma_{\tau C}^2$  appearing in this equation are identical to three of the six variances describing the DBM model, Equation (17.4). 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):

$$\Delta = \frac{JK\sigma_\tau^2}{\sigma_\epsilon^2 + \max(K\sigma_{\tau R}^2, 0) + \max(J\sigma_{\tau C}^2, 0)} \quad (17.28)$$

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

$$d = \theta_{1\bullet} - \theta_{2\bullet} \quad (17.29)$$

For two treatments, since the individual treatment effects must be the negatives of each other (because they sum to zero, see (17.5)), it follows that:

$$\sigma_\tau^2 = \frac{d^2}{2} \quad (17.30)$$

Therefore, for two treatments the numerator of the expression for  $\Delta$  is  $JKd^2/2$ . Dividing numerator and denominator of Equation (17.28) by  $K$ , one gets the final expression for  $\Delta$ , as coded in `RJafroc`, namely:

$$\Delta = \frac{Jd^2/2}{\max(\sigma_{\tau R}^2, 0) + (\sigma_\epsilon^2 + \max(J\sigma_{\tau C}^2, 0))/K} \quad (17.31)$$

The variances,  $\sigma_\tau^2$ ,  $\sigma_{\tau R}^2$  and  $\sigma_{\tau C}^2$ , appearing in Equation (17.31), can be calculated from the observed mean squares using the following equations, see (Hillis and Berbaum, 2004) Eqn. 4,

$$\left. \begin{aligned} \sigma_\epsilon^2 &= \text{MSTRC}^* \\ \sigma_{\tau R}^2 &= \frac{\text{MSTR}^* - \text{MSTRC}^*}{K^*} \\ \sigma_{\tau C}^2 &= \frac{\text{MSTC}^* - \text{MSTRC}^*}{J^*} \end{aligned} \right\} \quad (17.32)$$

- 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 (17.31), are the corresponding numbers for the *planned or pivotal study*.
- The three variances, determined from the pilot study via Equation (17.32), are assumed to apply unchanged to the pivotal study (as they are sample-size independent parameters of the DBM model).

### 17.5.2 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 (17.19).
- The starting point is Equation (17.19) 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 (17.19) can be related to the variances by an equation analogous to Equation (17.32), except that, again, all quantities in it apply to the *pivotal study* (note the absence of asterisks):

$$\left. \begin{array}{l} \sigma_{\epsilon}^2 = MSTRC \\ \sigma_{\tau R}^2 = \frac{MSTR - MSTRC}{K} \\ \sigma_{\tau C}^2 = \frac{MSTC - MSTRC}{J} \end{array} \right\} \quad (17.33)$$

Substituting from Equation (17.33) into Equation (17.19) 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 RJaafroc:

$$df_2 = \frac{(\max(\sigma_{\tau R}^2, 0) + (\max(J\sigma_{\tau C}^2, 0) + \sigma_{\epsilon}^2)/K)^2}{(\max(\sigma_{\tau R}^2, 0) + \sigma_{\epsilon}^2/K)^2} (J - 1) \quad (17.34)$$

### 17.5.3 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.,  $\text{VarTR} = \sigma_{\tau R}^2$ , 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$FOMs$trtMeanDiff["trt0-trt1", "Estimate"]
```

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

---

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

```

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

```

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

The extension to FRRC generalization is as follows. One sets  $\sigma_R^2 = 0$  and  $\sigma_{\tau R}^2 = 0$  in the DBM model (17.4). The F-statistic for testing the NH and its distribution under the NH is:

$$F = \frac{\text{MST}}{\text{MSTC}} \sim F_{I-1,(I-1)(K-1)} \quad (17.35)$$

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 (17.31) upon setting  $\sigma_{\tau R}^2 = 0$ .

$$\Delta = \frac{Jd^2/2}{(\sigma_\epsilon^2 + \max(J\sigma_{\tau C}^2, 0))/K} \quad (17.36)$$

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

```

## 17.7 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 (17.4). The F-statistic for testing the NH and its distribution under the NH is:

$$F = \frac{\text{MST}}{\text{MSTR}} \sim F_{I-1, (I-1)(J-1)} \quad (17.37)$$

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

$$\Delta = \frac{Jd^2/2}{\max(\sigma_{\tau R}^2, 0) + \sigma_e^2/K} \quad (17.38)$$

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.

## 17.8 Summary TBA

This chapter has detailed analysis of MRMIC 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 Barrett<sup>19,20</sup>, implemented in open-source JAVA software by Gallas and colleagues<sup>22,44</sup> (<http://didsr.github.io/iMRMC/>) is available on the web. Clarkson et al<sup>19,20</sup> 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 validate<sup>3</sup> DBM is that proposed by Roe and Metz<sup>39</sup> 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 published<sup>3</sup>. (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.

## 17.9 Things for me to think about

### 17.9.1 Expected values of mean squares

Assuming no replications the expected mean squares are as follows, Table Table 17.1; 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., 1992a), 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 17.1 for the subsequent development.**

Table 17.2: Table 9.1 Unconstrained expected values of mean-squares, as in (Dorfman et al., 1995)

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

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

Table 17.3: Table 9.2 Constrained expected values of mean-squares, as in (Dorfman et al., 1992a)

Source	df	E(MS)
T	(I-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2 + J\sigma_{\tau C}^2 + JK\sigma_\tau^2$
R	(J-1)	$\sigma_\epsilon^2 + I\sigma_{RC}^2 + IK\sigma_R^2$
C	(K-1)	$\sigma_\epsilon^2 + I\sigma_{RC}^2 + IJ\sigma_C^2$
TR	(I-1)(J-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2$
TC	(I-1)(K-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2$
RC	(J-1)(K-1)	$\sigma_\epsilon^2 + I\sigma_{RC}^2$
TRC	(I-1)(J-1)(K-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2$
$\epsilon$	0	$\sigma_\epsilon^2$

Table 17.4: Table 9.3 As in Hillis “marginal-means ANOVA paper” (Hillis, 2014)

Source	df	E(MS)
T	(I-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2 + J\sigma_{\tau C}^2 + JK\sigma_\tau^2$
R	(J-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + I\sigma_{RC}^2 + IK\sigma_R^2 + K\sigma_{\tau R}^2$
C	(K-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + I\sigma_{RC}^2 + IJ\sigma_C^2 + J\sigma_{\tau C}^2$
TR	(I-1)(J-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2$
TC	(I-1)(K-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2$
RC	(J-1)(K-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + I\sigma_{RC}^2$
TRC	(I-1)(J-1)(K-1)	$\sigma_\epsilon^2 + \sigma_{\tau RC}^2$
$\epsilon$	0	$\sigma_\epsilon^2$

## 17.10 References

# Chapter 18

## 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).

### 18.1 TBA How much finished

30%

### 18.2 Fixed-reader random-case (FRRC) analysis

The model is the same as in Eqn. (17.4) except one sets  $\sigma_R^2 = \sigma_{\tau R}^2 = 0$  in Table 17.1. The appropriate test statistic is:

$$\frac{E(MST)}{E(MSTC)} = \frac{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2 + JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2} \quad (18.1)$$

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

$$\frac{E(MST)}{E(MSTC)} = 1 \quad (18.2)$$

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

$$F_{DBM|R} = \frac{MST}{MSTC} \quad (18.3)$$

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 17.1. Therefore, the distribution of the observed value is (no Satterthwaite approximation needed this time as both numerator and denominator are simple mean-squares):

$$F_{DBM|R} \sim F_{I-1,(I-1)(K-1)} \quad (18.4)$$

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

$$F_{DBM|R} > F_{1-\alpha,I-1,(I-1)(K-1)} \quad (18.5)$$

The p-value of the test is the probability that a random sample from the F-distribution TBA (17.1) Eqn. (9.39), exceeds the observed value:

$$p = \Pr(F > F_{DBM|R} \mid F \sim F_{I-1,(I-1)(K-1)}) \quad (18.6)$$

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

$$CI_{1-\alpha} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,(I-1)(K-1)} \sqrt{2 \frac{MST}{JK}} \quad (18.7)$$

### 18.2.1 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.

#### 18.2.1.1 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.

### 18.3 Random-reader fixed-case (RRFC) analysis

The model is the same as in TBA (17.1) Eqn. (9.4) except one puts  $\sigma_C^2 = \sigma_{\tau C}^2 = 0$  in Table Table 17.1. It follows that:

$$\frac{E(MST)}{E(MSTR)} = \frac{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2 + JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2} \quad (18.8)$$

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

$$\frac{E(MST)}{E(MSTR)} = 1 \quad (18.9)$$

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

$$F_{DBM|C} \sim \frac{MST}{MSTR} \quad (18.10)$$

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 17.1.

$$F_{DBM|C} \sim F_{I-1, (I-1)(J-1)} \quad (18.11)$$

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

$$F_{DBM|C} > F_{1-\alpha, I-1, (I-1)(J-1)} \quad (18.12)$$

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

$$p = \Pr(F > F_{DBM|C} \mid F \sim F_{I-1, (I-1)(J-1)}) \quad (18.13)$$

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

$$CI_{1-\alpha} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2, (I-1)(J-1)} \sqrt{2 \frac{MSTR}{JK}} \quad (18.14)$$

### 18.3.0.1 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

```

## 18.4 References



# Chapter 19

## Introduction to the Obuchowski-Rockette method

### 19.1 TBA How much finished

70%

### 19.2 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"))
```

### 19.3 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.

## 19.4 Single-reader multiple-treatment

### 19.4.1 Overview

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

*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  $\theta$  is modeled as:

$$\theta_{i\{c\}} = \mu + \tau_i + \epsilon_{i\{c\}} \quad (19.1)$$

Eqn. (19.1) models the observed figure-of-merit  $\theta_{i\{c\}}$  as a constant term  $\mu$ , a treatment dependent term  $\tau_i$  (the treatment-effect), and a random term  $\epsilon_{i\{c\}}$ . The term  $\tau_i$  has the constraint:

$$\sum_{i=1}^I \tau_i = 0 \quad (19.2)$$

The left hand side of Eqn. (19.1) is the figure-of-merit  $\theta_{i\{c\}}$  for treatment  $i$  and case-set index  $\{c\}$ , where  $c = 1, 2, \dots, C$  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  $\theta_i$  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<sup>th</sup>* 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. (19.1) 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  $\Sigma$ :

$$\epsilon_{i\{c\}} \sim N_I(\vec{0}, \Sigma) \quad (19.3)$$

Here  $N_I$  is the I-variate normal distribution (i.e., each sample yields  $I$  random numbers). For the single-reader model Eqn. (19.1), the covariance matrix has the following structure :

$$\Sigma_{ii'} = Cov(\epsilon_{i\{c\}}, \epsilon_{i'\{c\}}) = \begin{cases} \text{Var} & (i = i') \\ Cov_1 & (i \neq i') \end{cases} \quad (19.4)$$

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  $\Sigma$  is:

$$\Sigma = \begin{pmatrix} \text{Var} & Cov_1 & \dots & Cov_1 & Cov_1 \\ Cov_1 & \text{Var} & \dots & Cov_1 & Cov_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ Cov_1 & Cov_1 & \dots & \text{Var} & Cov_1 \\ Cov_1 & Cov_1 & \dots & Cov_1 & \text{Var} \end{pmatrix} \quad (19.5)$$

If  $I = 2$  then  $\Sigma$  is a symmetric  $2 \times 2$  matrix, whose diagonal terms are the common variances in the two treatments (each assumed equal to  $\text{Var}$ ) and whose off-diagonal terms (each assumed equal to  $Cov_1$ ) are the co-variances. With  $I = 3$  one has a  $3 \times 3$  symmetric matrix with all diagonal elements equal to  $\text{Var}$  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<sub>1</sub> estimates need to be averaged over all distinct treatment-treatment pairings.*

<sup>1</sup>

Some elementary statistical results are presented in the Appendix.

---

<sup>1</sup>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.

### 19.4.2 Significance testing

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

$$MS(T) = \frac{1}{I-1} \sum_{i=1}^I (\theta_i - \theta_{\bullet})^2 \quad (19.6)$$

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

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

$$\chi_{1R} \equiv \frac{(I-1)MS(T)}{\text{Var} - \text{Cov1}} \sim \chi_{I-1}^2 \quad (19.7)$$

Eqn. (19.7) is from §5.4 in (Hillis, 2007b) 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., 2005a):

$$F_{1R} \equiv \frac{MS(T)}{\text{Var} - \text{Cov1}} \sim F_{I-1,\infty} \quad (19.8)$$

### 19.4.3 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:

$$p \equiv \Pr(f > F_{1R} \mid f \sim F_{I-1,\infty}) \quad (19.9)$$

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

$$CI_{1-\alpha,1R} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2,\infty} \sqrt{2(\text{Var} - \text{Cov1})} \quad (19.10)$$

Comparing Eqn. (19.10) to Eqn. (19.27) 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.]

#### 19.4.4 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  $\alpha$ . In the following, as is usual, we set  $\alpha = 0.05$ .

```

1  set.seed(seed = 201)
2  mu <- 0.8
3  vc <- UtilORVarComponentsFactorial(dataset02, FOM = "Wilcoxon")
4  trueVar <- vc$IndividualRdr$varEachRdr[1]
5  trueCov1 <- vc$IndividualRdr$cov1EachRdr[1]
6  sigma <- matrix(c(trueVar,
7                      trueCov1,
8                      trueCov1,
9                      trueVar),
10                     ncol = 2)
11 I <- 2
12 S <- 2000
13 # simulate foms for two modalities, S times
14 # using the sampling model
15 theta_i <- t(rmvnorm(n=S, mean=c(0,0), sigma=sigma) + mu)
16 # estimated variance covariances
17 vc <- VarCov1_FomInput(theta_i)
18 Var <- vc$Var
19 Cov1 <- vc$Cov1
20
21 # conduct NH testing
22 reject <- 0
23 for(s in 1:S) {
24
25   MS_T <- 0
26   for (i in 1:I) {
27     MS_T <- MS_T + (theta_i[i,s] - mean(theta_i[,s]))^2
28   }
29   MS_T <- MS_T/(I-1)
30
31   F_1R <- MS_T/(Var - Cov1)
32   pValue <- 1 - pf(F_1R, I-1, Inf)
33   if (pValue < 0.05) reject <- reject + 1
34 }
35 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. (19.1). Different values of `seed` correspond to different case samples.

Line 2 sets the value of  $\mu$  to 0.8, the average figure of merit, appearing in Eqn. (19.1).

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  $\Sigma$ . 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  $\Sigma$ .

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 19.2.

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 normal distribution (at line 14). Line 30 computes the F-statistic - compare to Eqn. (19.8). 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.

### 19.4.5 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
```

```

7 msT <- vc$IndividualRdr$msTEachRdr
8 I <- length(ds$ratings$NL[,1,1,1])
9 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]) + qt(0.025, Inf, lower.tail = F) * sqrt(2*(Var - Cov1))
13 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.

#### 19.4.6 Application 2

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

```

1 ds <- DfExtractDataset(dataset04, rdrs = 1, trts = c(4,5))
2 fom <- as.vector(UtilFigureOfMerit(ds, FOM = "wAFROC"))
3 fom <- t(fom)
4 vc <- UtilORVarComponentsFactorial(ds, FOM = "wAFROC")
5 Cov1 <- vc$IndividualRdr$cov1EachRdr
6 Var <- vc$IndividualRdr$varEachRdr
7 msT <- vc$IndividualRdr$msTEachRdr
8 I <- length(ds$ratings$NL[,1,1,1])
9 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.

## 19.5 Single-treatment multiple-reader

### 19.5.1 Overview

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

$$\theta_{j\{c\}} = \mu + R_j + \epsilon_{j\{c\}} \quad (19.11)$$

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  $\Sigma$ :

$$\epsilon_{j\{c\}} \sim N_J(\vec{0}, \Sigma) \quad (19.12)$$

The covariance matrix has the following structure:

$$\Sigma_{jj'} = \text{Cov}(\epsilon_{j\{c\}}, \epsilon_{j'\{c\}}) = \begin{cases} \text{Var} & (j = j') \\ \text{Cov}_2 & (j \neq j') \end{cases} \quad (19.13)$$

The reason for the subscript “2” in  $\text{Cov}_2$  will become clear when one extends this model to multiple-treatments and multiple-readers. The  $J \times J$  covariance matrix  $\Sigma$  is:

$$\Sigma = \begin{pmatrix} \text{Var} & \text{Cov}_2 & \dots & \text{Cov}_2 & \text{Cov}_2 \\ \text{Cov}_2 & \text{Var} & \dots & \text{Cov}_2 & \text{Cov}_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \text{Cov}_2 & \text{Cov}_2 & \dots & \text{Var} & \text{Cov}_2 \\ \text{Cov}_2 & \text{Cov}_2 & \dots & \text{Cov}_2 & \text{Var} \end{pmatrix} \quad (19.14)$$

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

### 19.5.2 Significance testing

Unlike the seemingly analogous single-reader multiple-treatment case addressed in Section 19.4.2, the single-treatment multiple-reader case is fundamentally different. This is because reader is a *random* factor while treatment, in Section 19.4.2, 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.,  $R \sim N(0, \sigma_R^2)$ .

### 19.5.3 Special case

If reader is regarded as a *fixed* factor significance testing between readers can be performed. The analysis presented in Section 19.4.2 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.

## 19.6 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. (19.1), 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:

$$\theta_{ij\{c\}} = \mu + \tau_i + R_j + (\tau R)_{ij} + \epsilon_{ij\{c\}} \quad (19.15)$$

- The fixed treatment effect  $\tau_i$  is subject to the usual constraint, Eqn. (19.2).
- The first two terms on the right hand side of Eqn. (19.15) have their usual meanings: a constant term  $\mu$  representing performance averaged over treatments and readers, and a treatment effect  $\tau_i$  ( $i = 1, 2, \dots, I$ ).
- 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  $j$  ( $j = 1, 2, \dots, J$ ), modeled by a zero-mean normal distribution with variance  $\sigma_R^2$ ;
  - $(\tau R)_{ij}$  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:

$$\left. \begin{aligned} R_j &\sim N(0, \sigma_R^2) \\ \tau R &\sim N(0, \sigma_{\tau R}^2) \end{aligned} \right\} \quad (19.16)$$

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

$$\mu = \theta_{\bullet\bullet\{1\}} \quad (19.17)$$

Averaging over the  $j$  index and performing a subtraction yields an estimate of  $\tau_i$ :

$$\tau_i = \theta_{i\bullet\{1\}} - \theta_{\bullet\bullet\{1\}} \quad (19.18)$$

The  $\tau_i$  estimates obey the constraint Eqn. (19.2). For example, with two treatments, the values of  $\tau_i$  must be the negatives of each other:  $\tau_1 = -\tau_2$ .

The error term on the right hand side of Eqn. (19.15) 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  $\Sigma$ . In other words,

$$\epsilon_{ij\{c\}} \sim N_{IJ}(\vec{0}, \Sigma) \quad (19.19)$$

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  $\Sigma$  of the distribution, is defined by 4 parameters,  $\text{Var}$ ,  $\text{Cov1}$ ,  $\text{Cov2}$ ,  $\text{Cov3}$ , defined as follows:

$$\text{Cov}(\epsilon_{ij\{c\}}, \epsilon_{i'j'\{c\}}) = \begin{cases} \text{Var}(i = i', j = j') \\ \text{Cov1}(i \neq i', j = j') \\ \text{Cov2}(i = i', j \neq j') \\ \text{Cov3}(i \neq i', j \neq j') \end{cases} \quad (19.20)$$

Apart from fixed effects, the model implied by Eqn. (19.15) and Eqn. (19.20) 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. (19.15) “looks” simpler because four covariance terms are encapsulated in the  $\epsilon$  term. As with the single-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  $\text{Var}$ ,  $\text{Cov1}$ ,  $\text{Cov}_2$ , and  $\text{Cov}_3$  estimates are averaged over all applicable treatment-reader combinations.

### 19.6.1 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  $6 \times 6$  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. (19.20) 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. (19.20) 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:

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

Similarly, cells with identical treatment indices but different reader indices represent Cov2:

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

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

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

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

$$Var = \left\langle \frac{1}{C-1} \sum_{c=1}^C (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})^2 \right\rangle_{ij} \epsilon_{ij\{c\}} \sim N_{IJ}(\vec{0}, \Sigma) \quad (19.21)$$

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. (19.21)). Notice that the left-hand-side of Eqn. (19.21) 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:

$$Cov = \begin{cases} Cov1 = \left\langle \frac{1}{C-1} \sum_{c=1}^C (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})(\theta_{i'j\{c\}} - \theta_{i'j\{\bullet\}}) \right\rangle_{ii', jj'} \\ Cov_2 = \left\langle \frac{1}{C-1} \sum_{c=1}^C (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})(\theta_{ij'\{c\}} - \theta_{ij'\{\bullet\}}) \right\rangle_{ii', jj'} \\ Cov_3 = \left\langle \frac{1}{C-1} \sum_{c=1}^C (\theta_{ij\{c\}} - \theta_{ij\{\bullet\}})(\theta_{i'j'\{c\}} - \theta_{i'j'\{\bullet\}}) \right\rangle_{ii', jj'} \end{cases} \quad (19.22)$$

In Eqn. (19.22) 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<sub>2</sub> one averages over all combinations of *same treatment and different readers*, as denoted by  $\langle \rangle_{ii,jj'}$ .
- For Cov<sub>3</sub> one averages over all combinations of *different treatments and different readers*, as denoted by  $\langle \rangle_{ii',jj'}$ .

### 19.6.2 Physical meanings of the covariance terms

The meanings of the different terms follow a similar description to that given in Eqn. 19.6.1. The diagonal term Var is the variance of the figures-of-merit when reader  $j$  interprets different case-sets  $\{c\}$  in treatment  $i$ . Each case-set yields a number  $\theta_{ij\{c\}}$  and the variance of the  $C$  numbers, averaged over the  $I \times J$  treatments and readers, is Var. It captures the total variability due to varying difficulty levels of the case-sets, inter-reader and within-reader variability.

It is easier to see the physical meanings of Cov1, Cov<sub>2</sub>, Cov<sub>3</sub> 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  $\rho_1$ . 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  $\rho_1$  approaches unity. If there is no common variation, the two arrays become independent, and  $\rho_1$  equals zero. Converting from correlation to covariance, see Eqn. (19.28), one has Cov1 < Var.
- $\rho_{2;iijj'}$  is the correlation of the figures-of-merit values when different readers  $j, j'$  interpret the same case-sets in the same treatment  $i$ . As before this yields two C-length arrays, whose correlation, upon averaging over all distinct treatment pairings and same readers, yields  $\rho_2$ . 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  $\rho_2$  will be smaller than  $\rho_1$ . 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<sub>2</sub> < Cov1 < Var.
- $\rho_{3;ii'jj'}$  is the correlation of the figure-of-merit values when different readers  $j, j'$  interpret the same case set in different treatments  $i, i'$ , etc., yielding  $\rho_3$ . This is expected to yield the least correlation.

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

$$\text{Cov}_3 \leq \text{Cov}_2 \leq \text{Cov}1 \leq \text{Var} \quad (19.23)$$

## 19.7 Summary

## 19.8 Discussion

## 19.9 Appendix: Covariance and correlation

Some elementary statistical results are reviewed here.

### 19.9.1 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  $(1 - \alpha)$ :

$$\Pr(d \leq D_{1-\alpha} \mid d \sim D) = 1 - \alpha \quad (19.24)$$

With definition Eqn. (19.24), the  $(1 - \alpha)$  quantile of the  $\chi^2_{I-1}$  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., 2005a, Eq. 22):

$$\frac{\chi^2_{1-\alpha, I-1}}{I-1} = F_{1-\alpha, I-1, \infty} \quad (19.25)$$

Eqn. (19.25) implies that the  $(1 - \alpha)$  quantile of the F-distribution with  $ndf = (I-1)$  and  $ddf = \infty$  equals the  $(1 - \alpha)$  quantile of the  $\chi^2_{I-1}$  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
```

### 19.9.2 Definitions of covariance and correlation

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

$$Cov(X, Y) = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{N - 1} = E(XY) - E(X)E(Y) \quad (19.26)$$

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  $x$ :

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

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

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

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

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

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

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

It has the property:

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

### 19.9.3 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:

$$\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y) = 2(\text{Var} - \text{Cov}) \quad (19.27)$$

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. (19.5) are equal. The correlation  $\rho_1$  is defined by (the reason for the subscript 1 on  $\rho$  is the same as the reason for the subscript 1 on  $\text{Cov}_1$ , which will be explained later):

$$\rho_1 = \frac{\text{Cov}_1}{\text{Var}}$$

The  $I \times I$  covariance matrix  $\Sigma$  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):

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ \rho_1\sigma^2 & \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ \rho_1\sigma^2 & \rho_1\sigma^2 & \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 \\ \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \sigma^2 & \rho_1\sigma^2 \\ \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \sigma^2 \end{bmatrix} \quad (19.28)$$

#### 19.9.4 Estimating the variance-covariance matrix

An unbiased estimate of the covariance matrix Eqn. (19.4) follows from:

$$\Sigma_{ii'}|_{ps} = \frac{1}{C-1} \sum_{c=1}^C (\theta_{i\{c\}} - \bar{\theta}_{i\{\bullet\}})(\theta_{i'\{c\}} - \bar{\theta}_{i'\{\bullet\}}) \quad (19.29)$$

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

$$\text{Var}_i|_{ps} = \frac{1}{C-1} \sum_{c=1}^C (\theta_{i\{c\}} - \bar{\theta}_{i\{\bullet\}})^2 \quad (19.30)$$

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

Sampling different case-sets, as required by Eqn. (19.29), 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:

$$\Sigma_{ii'}|_{bs} = \frac{1}{B-1} \sum_{b=1}^B (\theta_{i\{b\}} - \theta_{i\{\bullet\}}) (\theta_{i'\{b\}} - \theta_{i'\{\bullet\}}) \quad (19.31)$$

Eqn. (19.31), the bootstrap method of estimating the covariance matrix, is a direct translation of Eqn. (19.29). 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:

$$\Sigma_{ii'}|_{jk} = \frac{(K-1)^2}{K} \left[ \frac{1}{K-1} \sum_{k=1}^K (\theta_{i(k)} - \theta_{i(\bullet)}) (\theta_{i'(k)} - \theta_{i'(\bullet)}) \right] \quad (19.32)$$

[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. (19.29) and Eqn. (19.32). 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. (19.31) and Eqn. (19.32). 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.

### 19.9.5 The variance inflation factor

In Eqn. (19.32), the expression for the jackknife covariance estimate contains a *variance inflation factor*:

$$\frac{(K-1)^2}{K} \quad (19.33)$$

This factor multiplies the traditional expression for the covariance, shown in square brackets in Eqn. (19.32). 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.

### 19.9.6 Meaning of the covariance matrix

With reference to Eqn. (19.5), 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  $i = 1, 2, \dots, I$  and  $c = 1, 2, \dots, C$ . Considering treatment  $i$ , the variance of the FOM-values for the different case-sets  $c = 1, 2, \dots, C$ , is an estimate of  $Var_i$  for this treatment:

$$\sigma_i^2 \equiv Var_i = \frac{1}{C-1} \sum_{c=1}^C (\theta_{i\{c\}} - \bar{\theta}_{i\{\bullet\}}) (\theta_{i\{c\}} - \bar{\theta}_{i\{\bullet\}}) \quad (19.34)$$

The process is repeated for all treatments and the  $I$ -variance values are averaged. This is the final estimate of Var appearing in Eqn. (19.3).

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:

$$\text{Cov}_{1;ii'} = \frac{1}{C-1} \sum_{c=1}^C (\theta_{i\{c\}} - \bar{\theta}_{i\{\bullet\}}) (\theta_{i'\{c\}} - \bar{\theta}_{i'\{\bullet\}}) \quad (19.35)$$

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  $\text{Cov}_1$ . [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:

$$\rho_{1;ii'} = \frac{1}{C-1} \sum_{c=1}^C \frac{(\theta_{i\{c\}} - \bar{\theta}_{i\{\bullet\}})(\theta_{i'\{c\}} - \bar{\theta}_{i'\{\bullet\}})}{\sigma_i \sigma_{i'}} \quad (19.36)$$

Averaging over all different-treatment pairings yields the final estimate of the correlation  $\rho_1$ . Since the covariance is smaller than the variance, the magnitude of the correlation is smaller than 1. In most situations one expects  $\rho_1$  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. (19.31) and Eqn. (19.32) 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).

### 19.9.7 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 analyzed
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 `Util0RVarComponentsFactorial`:

```

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 one 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.

### 19.9.8 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     1.3028587      1     0.2536917

```

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  $\infty$  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\$FRRRC\$FTests[“Treatment”, “p”]
- CI Lower: ret\_rj\$FRRRC\$ciDiffTrt[1, “CILower”]
- Mid Value: ret\_rj\$FRRRC\$ciDiffTrt[1, “Estimate”]
- CI Upper: ret\_rj\$FRRRC\$ciDiffTrt[1, “CIUpper”]

#### 19.9.8.1 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\$TTranova[“R”, “MS”]
- msT: ret\_rj\$ANOVA\$TTranova[“T”, “MS”]
- msTR: ret\_rj\$ANOVA\$TTranova[“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.

## 19.10 References

# Chapter 20

## Obuchowski Rockette (OR) Analysis

### 20.1 TBA How much finished

80%

### 20.2 Introduction

In previous chapters the DBM significance testing procedure (Dorfman et al., 1992a) 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., 2005a; Hillis, 2007b; Hillis et al., 2008a). 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, 1995a), 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.

### 20.3 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  $\epsilon$  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.,

$$NH : \tau_i = 0 \quad (i = 1, 2, \dots, I) \quad (20.1)$$

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

$$F_{ORH} = \frac{MS(T)}{MS(TR) + J \max(\text{Cov2} - \text{Cov3}, 0)} \quad (20.2)$$

Eqn. (20.2) incorporates Hillis' modification of the original OR F-statistic. The modification ensures that the constraint Eqn. (19.23) 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*):

$$\left. \begin{aligned} MS(T) &= \frac{J}{I-1} \sum_{i=1}^I (\theta_{i\bullet} - \theta_{\bullet\bullet})^2 \\ MS(R) &= \frac{I}{J-1} \sum_{j=1}^J (\theta_{\bullet j} - \theta_{\bullet\bullet})^2 \\ MS(TR) &= \frac{1}{(I-1)(J-1)} \sum_{i=1}^I \sum_{j=1}^J (\theta_{ij} - \theta_{i\bullet} - \theta_{\bullet j} + \theta_{\bullet\bullet}) \end{aligned} \right\} \quad (20.3)$$

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

$$F_{OR} = \frac{MS(T)}{MS(TR) + J(\text{Cov2} - \text{Cov3})} \quad (20.4)$$

Note that Eqn. (20.4) lacks the constraint, subsequently proposed by Hillis, which ensures that the denominator cannot be negative. The following distribution was proposed for the test statistic.

$$F_{OR} \sim F_{\text{ndf}, \text{ddf}} \quad (20.5)$$

The original degrees of freedom were defined by:

$$\begin{aligned} \text{ndf} &= I - 1 \\ \text{ddf} &= (I - 1) \times (J - 1) \end{aligned} \quad (20.6)$$

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.

### 20.3.1 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. (20.2) 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  $100 \pm 20$  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.

### 20.3.2 Hillis ddf

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

$$ddf_H = \frac{[MS(TR) + J \max(Cov2 - Cov3, 0)]^2}{\frac{[MS(TR)]^2}{(I-1)(J-1)}} \quad (20.7)$$

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

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

If  $\text{Cov2} < \text{Cov3}$  (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. (20.2), is distributed as an F-statistic with  $ndf = I-1$  and  $ddf = ddf_H$  degrees of freedom (Hillis et al., 2005a; Hillis, 2007b; Hillis et al., 2008a):

$$F_{ORH} \sim F_{ndf, ddf_H} \quad (20.8)$$

If the expected ordering is true, i.e.,  $\text{Cov2} > \text{Cov3}$ , 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. (2008a)).

### 20.3.3 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. (20.2):

- 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.
- $\alpha$  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.
- $\text{ddf}_H$  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  $F_{ORH}$ :

$$p = \Pr(F > F_{ORH} \mid F \sim F_{\text{ndf}, \text{ddf}_H}) \quad (20.9)$$

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

$$\begin{aligned} CI_{1-\alpha, RRR, \theta_{i\bullet} - \theta_{i'\bullet}} = & \theta_{i\bullet} - \theta_{i'\bullet} \\ & \pm t_{\alpha/2, \text{ddf}_H} \sqrt{\frac{2}{J} (MS(TR) + J \max(\text{Cov2} - \text{Cov3}, 0))} \end{aligned} \quad (20.10)$$

Define  $\text{df}_i$ , the degrees of freedom for modality  $i$ :

$$\text{df}_i = (\text{MS(R)}_i + J \max(\text{Cov2}_i, 0))^2 / \text{MS(R)}_i^2 * (J - 1) \quad (20.11)$$

Here  $\text{MS(R)}_i$  is the reader mean-square for modality  $i$ , and  $\text{Cov2}_i$  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:

$$CI_{1-\alpha, RRRC, \theta_{i\bullet}} = \theta_{i\bullet} \pm t_{\alpha/2, df_i} \sqrt{\frac{1}{J}(\text{MS}(\mathbf{R})_i + J \max(\text{Cov}2_i, 0))} \quad (20.12)$$

## 20.4 Fixed-reader random-case

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

$$F_{ORH|R} = \frac{MS(T)}{\text{Var} - \text{Cov}1 + (J-1) \max(\text{Cov}2 - \text{Cov}3, 0)} \quad (20.13)$$

[For  $J = 1$ , Eqn. (20.13) reduces to Eqn. (19.8), i.e., the single-reader analysis described in the previous chapter.]

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

$$F_{ORH|R} \sim F_{I-1, \infty} \quad (20.14)$$

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  $\chi^2$  distribution with  $I - 1$  degrees of freedom, i.e., as  $\chi^2_{I-1}$ . Therefore, one has, analogous to Eqn. (19.7),

$$\chi^2_{ORH|R} \equiv (I-1)F_{ORH|R} \sim \chi^2_{I-1} \quad (20.15)$$

The critical value of the  $\chi^2$  statistic is  $\chi^2_{1-\alpha, I-1}$ , which is that value such that fraction  $(1 - \alpha)$  of the area under the  $\chi^2_{I-1}$  distribution lies to the left of the critical value. The null hypothesis is rejected if the observed value of the  $\chi^2$  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  $\chi^2_{I-1}$  exceeds the observed value of the test statistic  $\chi^2_{ORH|R}$  statistic defined in Eqn. (20.15):

$$p = \Pr(\chi^2 > \chi^2_{ORH|R} | \chi^2 \sim \chi^2_{I-1}) \quad (20.16)$$

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

$$\begin{aligned} CI_{1-\alpha, FRRC, \theta_{i\bullet} - \theta_{i'\bullet}} = & (\theta_{i\bullet} - \theta_{i'\bullet}) \\ & \pm t_{\alpha/2, \infty} \sqrt{\frac{2}{J} (\text{Var} - \text{Cov1} + (J-1) \max(\text{Cov2} - \text{Cov3}, 0))} \end{aligned} \quad (20.17)$$

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  $\chi^2$  statistic exceeds the critical value  $\chi^2_{1-\alpha, I-1}$ .
- The p-value is less than  $\alpha$ .
- 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  $CI_{1-\alpha, FRRC, \theta_{ij} - \theta_{i'j}}$ .

$$CI_{1-\alpha, FRRC, \theta_{i\bullet}} = \theta_{i\bullet} \pm z_{\alpha/2} \sqrt{\frac{1}{J} (\text{Var}_i + (J-1) \max(\text{Cov2}_i, 0))} \quad (20.18)$$

$$CI_{1-\alpha, FRRC, \theta_{ij} - \theta_{i'j}} = (\theta_{ij} - \theta_{i'j}) \pm z_{\alpha/2} \sqrt{2(\text{Var}_j - \text{Cov1}_j)} \quad (20.19)$$

In these equations  $\text{Var}_i$  and  $\text{Cov2}_i$  are computed using the data for treatment  $i$  only, and  $\text{Var}_j$  and  $\text{Cov1}_j$  are computed using the data for reader  $j$  only.

## 20.5 Random-reader fixed-case

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

$$F_{ORH|C} = \frac{MS(T)}{MS(TR)} \quad (20.20)$$

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

$$\left. \begin{array}{rcl} \text{ndf} & = & I - 1 \\ \text{ddf} & = & (I - 1)(J - 1) \\ F_{ORH|C} & \sim & F_{\text{ndf}, \text{ddf}} \end{array} \right\} \quad (20.21)$$

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:

$$CI_{1-\alpha, RRFC, \theta_{i\bullet} - \theta_{i'\bullet}} = (\theta_{i\bullet} - \theta_{i'\bullet}) \pm t_{\alpha/2, (I-1)(J-1)} \sqrt{\frac{2}{J} MS(TR)} \quad (20.22)$$

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

$$CI_{1-\alpha, RRFC, \theta_{i\bullet}} = \theta_{i\bullet} \pm t_{\alpha/2, J-1} \sqrt{\frac{1}{J} MS(R)_i} \quad (20.23)$$

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

## 20.6 Single treatment analysis

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

# Chapter 21

## Obuchowski Rockette Applications

### 21.1 TBA How much finished

80%

### 21.2 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).

## 21.3 Hand calculation

Dataset `dataset02` is well-known 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.

### 21.3.1 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 calculate the NH testing statistic.
- The relevant equation is Eqn. (20.2).
- `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  $5.5103062 \times 10^{-4}$ .
  - $Cov2$  is in `vc$VarCom["Cov2", "Estimates"]`, whose value is  $3.4407483 \times 10^{-4}$ .
  - $Cov3$  is in `vc$VarCom["Cov3", "Estimates"]`, whose value is  $2.3902837 \times 10^{-4}$ .

Applying Eqn. (20.2) 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. (20.7)). 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. (20.9) 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  $I = 2$ , there 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. (20.10), in which equation the expression inside the square-root symbol is  $2/J*den$ .
- $\alpha$ , 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. (20.11) and Eqn. (20.12). 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

```

### 21.3.2 Fixed-Reader Random-Case (FRRC) analysis

- The chi-square statistic is calculated using Eqn. (20.13) and Eqn. (20.15).
- 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  $8.0228827 \times 10^{-4}$ .
- $Cov1$  is in `vc$VarCom["Cov1", "Estimates"]`, whose value is  $3.4661371 \times 10^{-4}$ .
- 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.,  $CI_{1-\alpha, FRRC, \theta_i - \theta_j}$ .
- The relevant equation is Eqn. (20.17), 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.,  $CI_{1-\alpha, FRRC, \theta_i}$ .
- The relevant formula is in Eqn. (20.18), 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. (20.19), whose implementation follows:

```

trtMeanDiffss1 <- 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)))
CIRreader <- 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) {
    trtMeanDiffss1[j, pair] <- foms[i, j] - foms[ip, j]
    trDiffNames[j,pair] <- diffTRName[pair]
    zStat[j,pair] <- trtMeanDiffss1[j,pair]/stdErr[j,pair]
    PrGTz[j,pair] <-
      2 * pnorm(abs(zStat[j,pair]), lower.tail = FALSE)
    CIRreader[j, pair,] <-
      c(trtMeanDiffss1[j,pair] +
        qnorm(alpha/2) * stdErr[j,pair],
        trtMeanDiffss1[j,pair] +
        qnorm(1-alpha/2) * stdErr[j,pair])
    rowName <-
      paste0("rdr", Reader[j,pair], ":", trDiffNames[j, pair])
    ciDiffTrtEachRdr <- rbind(
      ciDiffTrtEachRdr,
      data.frame(Estimate = trtMeanDiffss1[j, pair],
                  StdErr = stdErr[j,pair],
                  z = zStat[j, pair],
                  PrGTz = PrGTz[j, pair],
                  CILower = CIRreader[j, pair,1],
                  CIUpper = CIRreader[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.

### 21.3.3 Random-Reader Fixed-Case (RRFC) analysis

The F-statistic is shown in Eqn. (20.20). This time  $\text{ndf} = I - 1$  and  $\text{ddf} = (I - 1) \times (J - 1)$ , 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. (20.22), 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.

## 21.4 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.,  $\theta_{ij}$ ,
  - `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:
  - `Tanova`, 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  $CI_{1-\alpha,FRRRC,\theta_{i\bullet}-\theta_{i'\bullet}}$  in the previous chapter,
  - `ciAvgRdrEachTrt`, the confidence intervals for individual treatment FOMs, averaged over readers, denoted  $CI_{1-\alpha,FRRRC,\theta_{i\bullet}}$  in the previous chapter,
  - `ciDiffTrtEachRdr`, the confidence intervals for inter-treatment FOM differences for individual readers, denoted  $CI_{1-\alpha,FRRRC,\theta_{ij}-\theta_{i'j}}$  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 individual 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.

### 21.4.1 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.,  $CI_{1-\alpha, RRRC, \theta_i \bullet}$ .

### 21.4.2 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  $\alpha$ .

```
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.,  $CI_{1-\alpha, FRRC, \theta_{ij} - \theta_{i'j}}$ .

### 21.4.3 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.,  $CI_{1-\alpha, RRFC, \theta_{i\bullet}}$ .

## 21.5 RJa froc: 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
#>
#>
#> $FRRRC
#> $FRRRC$FTests
#>          MS Chisq DF      p
#> Treatment 0.001897 13.6 4 0.00868
#> Error     0.000558   NA NA    NA
#>

```

```

#> $FRRRC$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
#>
#> $FRRRC$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
#>
#> $FRRRC$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
```

### 21.5.1 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.

### 21.5.2 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.

### 21.5.3 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).

## 21.6 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

```

### 21.6.1 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 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.,  $CI_{1-\alpha, RRRC, \theta_{i\bullet} - \theta_{i'}}$ .
- 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.,  $CI_{1-\alpha, RRRC, \theta_{ij} - \theta_{i'j}}$ .

### 21.6.2 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.,  $CI_{1-\alpha, FRRC, \theta_{ij} - \theta_{i'j}}$ .

### 21.6.3 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).

## 21.7 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
#>
#>
#> $FRRRC
#> $FRRRC$FTests
#>           DF   MS FStat      p
#> Treatment  4 0.463 3.85 0.00416
#> Error      796 0.120   NA      NA
#>
#> $FRRRC$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
#>
#> $FRRRC$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

```

### 21.7.1 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.,  $CI_{1-\alpha, RRRC, \theta_i - \theta_{i'}}$ .
- `st4$RRRC$ciAvgRdrEachTrt` contains confidence intervals for each treatment, averaged over readers, i.e.,  $CI_{1-\alpha, RRRC, \theta_i}$ .

### 21.7.2 Fixed-Reader Random-Case (FRRRC) analysis

- `st4$FRRRC$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.,  $CI_{1-\alpha, FRRC, \theta_{ij} - \theta_{i'j}}$ .

### 21.7.3 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).

## 21.8 Summary

## 21.9 Discussion

## 21.10 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  $\infty$ , 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`.

## 21.11 References

# Chapter 22

## Sample size estimation for ROC studies DBM method

### 22.1 TBA How much finished

80%

### 22.2 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 MRMIC 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)  $\theta$  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  $\theta$ 's 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  $\alpha$ , typically set at 5%, one rejects the NH and declares the treatments different at the  $\alpha$  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  $\theta$ . 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.

## 22.3 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  $\alpha$ . If the NH is false and one fails to reject it, the probability of the ensuing Type II- error is denoted  $\beta$ . Statistical power is the complement of  $\beta$ , i.e.,

$$\text{Power} = 1 - \beta \quad (22.1)$$

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.

### 22.3.1 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.*

### 22.3.2 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  $\sigma_R^2$  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  $\sigma_R^2$  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  $\sigma_R^2$ , making the sample size estimation procedure less accurate.
- The final term in the denominator is  $J\sigma_{\tau C}^2$ . The variance  $\sigma_C^2$  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  $J \ll K$ , 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).

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

#### 22.3.4 Significance testing

#### 22.3.5 p-value and confidence interval

#### 22.3.6 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 anticipated 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  $\sigma$  is one-fourth the width of the 95% confidence interval for  $d_{obs}$ . Anticipating more than  $2\sigma$  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.

## 22.4 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:

$$\Delta = \frac{JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + J\sigma_{\tau C}^2} \quad (22.2)$$

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

$$F_{AH|R} \equiv \frac{MST}{MSTC} \sim F_{I-1, (I-1)(K-1), \Delta} \quad (22.3)$$

### 22.4.1 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:

$$\Delta = \frac{JK\sigma_\tau^2}{\sigma_\epsilon^2 + \sigma_{\tau RC}^2 + K\sigma_{\tau R}^2} \quad (22.4)$$

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

$$F_{AH|C} \equiv \frac{MST}{MSTR} \sim F_{I-1, (I-1)(J-1), \Delta} \quad (22.5)$$

### 22.4.2 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.

**22.4.3 Random-reader fixed-case (RRFC) analysis**

**22.4.4 Single-treatment multiple-reader analysis**

**22.5 Discussion/Summary/2**

**22.6 References**

# Chapter 23

## Sample size estimation for ROC studies OR method

### 23.1 TBA How much finished

70%

### 23.2 Introduction

### 23.3 Statistical Power

$$Power = 1 - \beta \quad (23.1)$$

#### 23.3.1 Sample size estimation for random-reader random-cases

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

$$Y_{n(ijk)} = \mu + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + (\tau RC)_{ijk} + \epsilon_{n(ijk)} \quad (23.2)$$

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

$$d = \theta_{1\bullet} - \theta_{2\bullet} \quad (23.3)$$

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  $\Delta$ :

$$F_{DBM|AH} \sim F_{ndf,ddf,\Delta} \quad (23.4)$$

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

$$\Delta = \frac{JK\sigma_{Y;\tau}^2}{(\sigma_{Y;\epsilon}^2 + \sigma_{Y;\tau RC}^2) + K\sigma_{Y;\tau R}^2 + J\sigma_{Y;\tau C}^2} \quad (23.5)$$

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  $\sigma_{Y;\tau C}^2$  can turn out to be negative. To avoid a negative denominator, Hillis suggests the following modification:

$$\Delta = \frac{JK\sigma_{Y;\tau}^2}{(\sigma_{Y;\epsilon}^2 + \sigma_{Y;\tau RC}^2) + K\sigma_{Y;\tau R}^2 + \max(J\sigma_{Y;\tau C}^2, 0)} \quad (23.6)$$

This expression depends on three variance components,  $(\sigma_{Y;\epsilon}^2 + \sigma_{Y;\tau RC}^2)$  - the two terms are inseparable -  $\sigma_{Y;\tau R}^2$  and  $\sigma_{Y;\tau C}^2$ . 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:

$$ddf_H = \frac{[MSTR + \max(MSTR - MSTRC, 0)]^2}{\frac{[MSTR]^2}{(I-1)(J-1)}} \quad (23.7)$$

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). RJafrro 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:

$$\sigma_{Y;\tau}^2 = \frac{d^2}{2} \quad (23.8)$$

### 23.3.2 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  $(\sigma_{Y;\epsilon}^2 + \sigma_{Y;\tau RC}^2)$ . 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  $K\sigma_{Y;\tau R}^2$ , the treatment-reader variance component multiplied by the total number of cases. The reader variance  $\sigma_{Y;R}^2$  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  $\sigma_{Y;R}^2$  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  $\sigma_{Y;R}^2$ , making the sample size estimation procedure less accurate.
- The final term in the denominator is  $J\sigma_{Y;\tau C}^2$ . The variance  $\sigma_{Y;C}^2$  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  $J \ll K$ , 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).

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

#### 23.3.4 Significance testing

#### 23.3.5 p-value and confidence interval

#### 23.3.6 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 anticipated 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  $\sigma$  is one-fourth the width of the 95% confidence interval for  $d_{obs}$ . Anticipating more than  $2\sigma$  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.

## 23.4 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:

$$\Delta = \frac{JK\sigma_{Y;\tau}^2}{\sigma_{Y;\epsilon}^2 + \sigma_{Y;\tau RC}^2 + J\sigma_{Y;\tau C}^2} \quad (23.9)$$

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

$$F_{AH|R} \equiv \frac{MST}{MSTC} \sim F_{I-1,(I-1)(K-1),\Delta} \quad (23.10)$$

### 23.4.1 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:

$$\Delta = \frac{JK\sigma_{Y;\tau}^2}{\sigma_{Y;\epsilon}^2 + \sigma_{Y;\tau RC}^2 + K\sigma_{Y;\tau R}^2} \quad (23.11)$$

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

$$F_{AH|C} \equiv \frac{MST}{MSTR} \sim F_{I-1,(I-1)(J-1),\Delta} \quad (23.12)$$

### 23.4.2 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" = Power)
#>   J   K   FCrit      ddf      ncp RRRCPower
#> 1 10 163 4.1270572 34.334268 8.1269825 0.79111255

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

23.4.3 Fixed-reader random-case (FRRC) analysis

23.4.4 Random-reader fixed-case (RRFC) analysis

23.4.5 Single-treatment multiple-reader analysis

### **23.5 Discussion/Summary/3**

### **23.6 References**



# **FROC paradigm**



# Chapter 24

## The FROC paradigm

### 24.1 TBA How much finished

70%

### 24.2 Introduction

Until now the focus has been on the receiver operating characteristic (ROC) paradigm. For diagnostic tasks such as detecting diffuse interstitial lung disease<sup>1</sup>, or diseases similar to it, where *disease location is implicit* – by definition diffuse interstitial lung disease is spread through, and confined to, lung tissues – this is an appropriate paradigm in the sense that essential information is not being lost by limiting the radiologist’s response in the study to a single rating categorizing the likelihood of presence of interstitial disease.

In clinical practice it is not only important to identify if the patient is diseased, but also to offer further guidance to subsequent care-givers regarding other characteristics (such as type, location, size, extent) of the disease. In most clinical tasks, if the radiologist believes the patient may be diseased, there is a location (or more than one location) associated with the manifestation of the suspected disease. Physicians have a term for this: *focal disease: disease located at a specific region of the image*.

---

<sup>1</sup>Diffuse interstitial lung disease refers to disease within both lungs that affects the interstitium or connective tissue that forms the support structure of the lungs’ air sacs or alveoli. When one inhales, the alveoli fill with air and pass oxygen to the blood stream. When one exhales, carbon dioxide passes from the blood into the alveoli and is expelled from the body. When interstitial disease is present, the interstitium becomes inflamed and stiff, preventing the alveoli from fully expanding. This limits both the delivery of oxygen to the blood stream and the removal of carbon dioxide from the body. As the disease progresses, the interstitium scars with thickening of the walls of the alveoli, which further hampers lung function.

For focal disease, the ROC paradigm restricts the collected information to a single rating representing the confidence level that there is disease *somewhere* in the patient's imaged anatomy. The emphasis on "somewhere" is because it begs the question: if the radiologist believes the disease is somewhere, why not have them to point to it? In fact they do "point to it" in the sense that they record the location(s) of suspect regions in their clinical report, but the ROC paradigm cannot use this information. Clinicians have long recognized problems with ignoring location (Black and Dwyer, 1990; Black, 2000). Neglect of location information leads to loss of statistical power: the basic reason for this is that there is additional noise in the measurement due to crediting the reader for correctly detecting the diseased patient but getting the wrong lesion location - i.e., being right for the wrong reason. One way of compensating for reduced statistical power is to increase the sample size, which increases the cost of the study and is also unethical because, by not using the optimal paradigm and analysis, one is subjecting more patients to imaging procedures (Halpern et al., 2002).

Here is an outline of this chapter. Four observer performance paradigms are compared, using a visual schematic, as to the kinds of information collected and ignored. An essential characteristic of the FROC paradigm, namely *visual search*, is introduced. The FROC paradigm and its historical context is described. A pioneering FROC study using phantom images is described. Key differences between FROC ratings and ROC data are noted. The FROC plot is introduced. The dependence of population and empirical FROC plots on a variable identified as *perceptual signal-to-noise-ratio* (*pSNR*) is shown. The expected dependence of the FROC curve on pSNR is illustrated with a "solar" analogy – understanding this is key to obtaining a good intuitive feel for this paradigm. The finite extent of the FROC curve, characterized by an *end-point*, is noted.

The starting point is a comparison of four observer performance paradigms.

### 24.3 Location specific paradigms

Location-specific paradigms take into account, to varying degrees, information regarding the locations of perceived lesions, so they are sometimes referred to as lesion-specific (or lesion-level) paradigms (Alberdi et al., 2008). Usage of this term is discouraged. In this book the term "lesion" is reserved for true malignant lesions (as distinct from "perceived lesions" or "suspicious regions" that may not be true lesions).

All observer performance methods involve detecting the presence of true lesions; so ROC methodology is, in this sense, also lesion-specific. On the other hand location is a characteristic of true and perceived focal lesions, and methods that account for location are better termed *location-specific* than lesion-specific. There are three location-specific paradigms:

- the free-response ROC (FROC) (Bunch et al., 1977b; Chakraborty, 1989);
- the location ROC (LROC) (Starr et al., 1977; Swensson, 1996a);
- the region of interest (ROI) (Obuchowski et al., 2000).

Fig. 24.1 shows a schematic mammogram interpreted according to current observer performance paradigms. The arrows point to two real lesions and the three light crosses indicate suspicious regions. Evidently the radiologist saw one of the lesions, missed the other lesion and mistook two normal structures for lesions.

- ROC (top-left): the radiologist assigns a single rating that the image contains at least one lesion, somewhere.
- FROC (top-right): the dark crosses indicate suspicious regions that are marked and the accompanying numerals are the FROC ratings.
- LROC (bottom-left): the radiologist provides a single rating that the image contains at least one lesion and marks the most suspicious region.
- ROI (bottom-right): the image is divided – by the researcher – into a number of regions-of-interest (ROIs) and the radiologist rates each ROI for presence of at least one lesion somewhere within the ROI.

The numbers and locations of suspicious regions depend on the case and the radiologists' expertise level. Some images can be correctly perceived as obviously non-diseased so that expert radiologists perceive nothing suspicious in them, or they can be correctly perceived as obviously diseased and the suspicious regions are so conspicuous that they are correctly localized by the expert radiologist. There is the gray area – especially when lesions are of low conspicuity – where two radiologists may perceive different suspicious regions, not all lesions present are perceived, and/or false regions are perceived as lesions.

In Fig. 24.1, evidently the radiologist found one of the lesions (the lightly shaded cross near the left most arrow), missed the other one (pointed to by the second arrow) and mistook two normal structures for lesions (the two lightly shaded crosses that are relatively far from any true lesion). The term *lesion* always refers to a true or real lesion. The prefix "true" or "real" is implicit. The term *suspicious region* is reserved for any region that, as far as the observer is concerned, has "lesion-like" characteristics. *A lesion is a real while a suspicious region is perceived.*

- In the ROC paradigm, Fig. 24.1 (top-left), the radiologist assigns a single rating indicating the confidence level that there is at least one lesion somewhere in the image. Assuming a 1 – 5 positive directed integer rating scale, if the left-most lightly shaded cross is a highly suspicious region then the ROC rating might be 5 (highest confidence for presence of disease).
- In the free-response (FROC) paradigm, Fig. 24.1 (top-right), the dark shaded crosses indicate suspicious regions that were *marked* (or *reported*

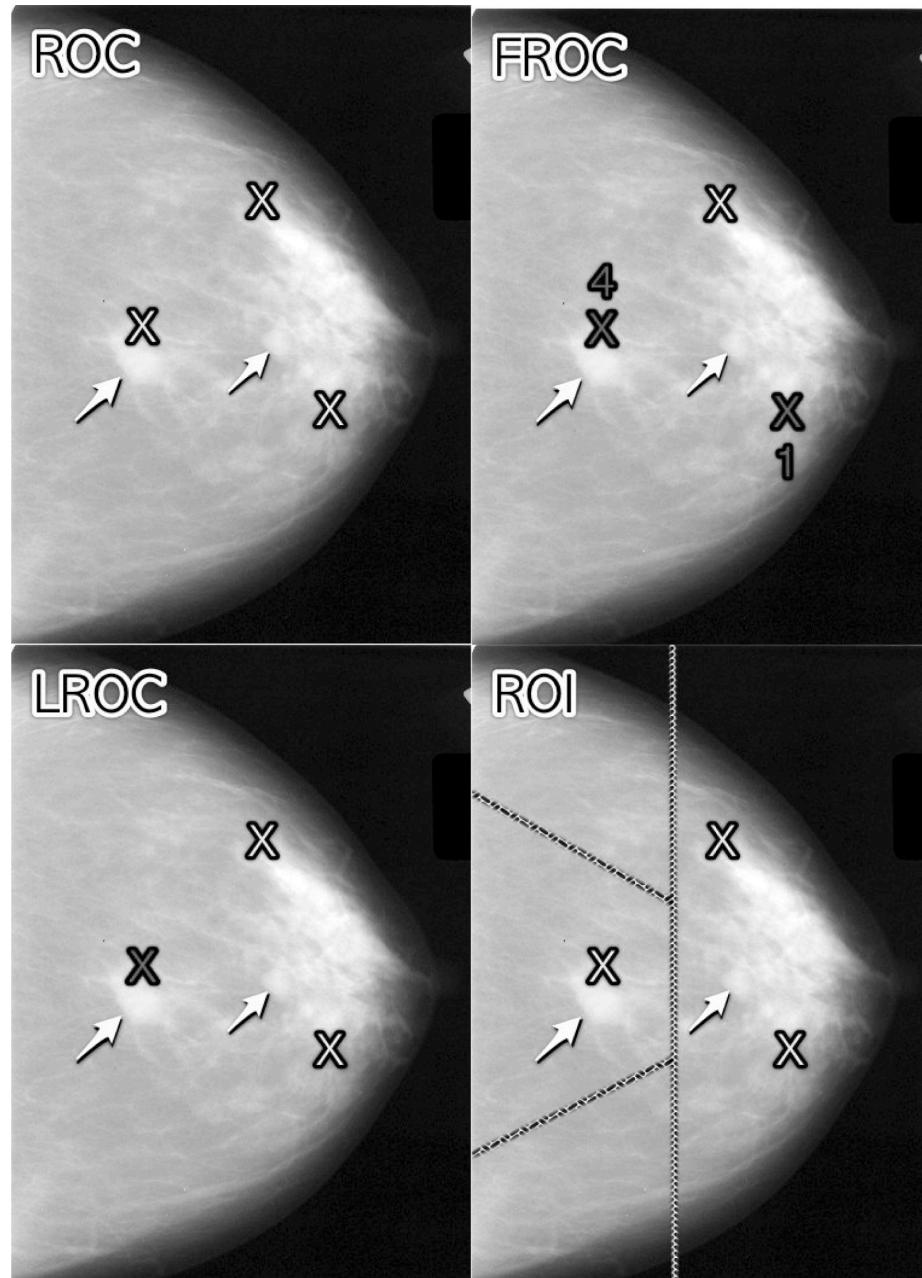


Figure 24.1: Upper Left: ROC, Upper Right: FROC, Lower Left: LROC, Lower Right: ROI

in the clinical report), and the adjacent numbers are the corresponding ratings, which now apply to specific regions in the image, unlike ROC where the rating applies to the whole image. Assuming the allowed FROC ratings are 1 through 4, two marks are shown, one rated FROC-4, which is close to a true lesion, and the other rated FROC-1, which is not close to any true lesion. The third suspicious region, indicated by the lightly shaded cross, was not marked, implying its confidence level did not exceed the lowest reporting threshold. The marked region rated FROC-4 (highest FROC confidence) is likely what caused the radiologist to assign the ROC-5 rating to this image in the top-left ROC paradigm figure.

- In the LROC paradigm, Fig. 24.1 (bottom-left), the radiologist provides a rating summarizing confidence that there is at least one lesion somewhere in the image (as in the ROC paradigm) and marks the most suspicious region in the image. In this example the rating might be LROC-5, the five rating being the same as in the ROC paradigm, and the mark may be the suspicious region rated FROC-4 in the FROC paradigm, and, since it is close to a true lesion, in LROC terminology it would be recorded as a *correct localization*. If the mark were not near a lesion it would be recorded as an *incorrect localization*. Only one mark is allowed in this paradigm, and in fact one mark is *required* on every image, even if the observer does not find any suspicious regions to report. The late Prof. Swensson has been the prime contributor to this paradigm.
- In the region of interest (ROI) paradigm, the researcher segments the image into a number of regions-of-interest (ROIs) and the radiologist rates each ROI for presence of at least one suspicious region somewhere within the ROI. The rating is similar to the ROC rating, except it applies to the segmented ROI, not the whole image. Assuming a 1 – 5 positive directed integer rating scale in Fig. 24.1 (bottom-right) there are four ROIs. The ROI at ~9 o'clock might be rated ROI-5 as it contains the most suspicious light cross, the one at ~11 o'clock might be rated ROI-1 as it does not contain any light crosses, the one at ~3 o'clock might be rated LROC-2 or 3 (the unmarked light cross would tend to increase the confidence level) and the one at ~7 o'clock might be rated ROI-1. In the example shown in Fig. 24.1 (bottom-right), each case yields 4 ratings. Prof. Obuchowski has been the principal contributor to this paradigm.<sup>2</sup>

The rest of this book part focuses on the FROC paradigm.

---

<sup>2</sup>When different views of the same patient anatomy (perhaps in different modalities) are available, it is assumed that all images are segmented consistently, and the rating for each ROI takes into account all views of that ROI in the different views (or modalities). The segmentation shown in the figure is a schematic. In fact the ROIs could be clinically driven descriptors of location, such as "apex of lung" or "mediastinum", and the image does not have to have lines showing the ROIs (which would be distracting to the radiologist). The number of ROIs per image can be at the researcher's discretion and there is no requirement that every case have a fixed number of ROIs.

## 24.4 Visual search

The FROC paradigm in medical imaging is equivalent to a visual search task. Any search task has two components: (i) finding something and (ii) acting on it. An example of a search task is looking for lost car-keys or a milk carton in the refrigerator. Success in a search task is finding the searched for object. Acting on it could be driving to work or drinking milk from the carton. There is search-expertise associated with any search task. Husbands are notoriously bad at finding the milk carton in the refrigerator (analogy due to Dr. Elizabeth Krupinski at an SPIE course taught jointly with me). Like anything else, search expertise is honed by experience, i.e., lots of practice.

Likewise, a medical imaging search task has two components (i) finding suspicious regions and (ii) acting on each finding (“finding”, used as a noun, is the actual term used by clinicians in their reports), i.e., determining the relevance of each finding to the health of the patient, and whether to report it. A general feature of a medical imaging search task is that the radiologist does not know a-priori if the patient is diseased and, if diseased, how many lesions are present. In the breast-screening context, it is known a-priori that about 5 out of 1000 cases have cancers, so 99.5% of the time odds are that the case has no malignant lesions.<sup>3</sup> **The radiologist searches each image for lesions.** If a suspicious region is found, and provided it is sufficiently suspicious, the relevant location is marked and rated for confidence in being a lesion. The process is repeated for each suspicious region found in the case. A screening mammography report consists of a listing of search related findings. To summarize: **Free-response data consists of a variable number of mark-rating pairs per case.**

### 24.4.1 Proximity criterion and scoring the data

In the first two clinical applications of the FROC paradigm (Chakraborty et al., 1986; Niklason et al., 1986) the marks and ratings were indicated by a grease pencil on an acrylic overlay aligned, in a reproducible way, to the CRT displayed chest image. Credit for a correct detection and localization, termed a lesion-localization or LL-event<sup>4</sup>, was given only if a mark was sufficiently close (as per proximity criterion, see below) to an actual diseased region; otherwise, the observer’s mark-rating pair was scored as a non-lesion localization or NL-event.

- The use of ROC terminology, such as true positives or false positives to describe FROC data is not conducive to clarity, and is strongly discouraged.

---

<sup>3</sup>The probability of benign suspicious regions is much higher (Ernster, 1981), about 13% for women aged 40-45.

<sup>4</sup>The proper terminology for this paradigm has evolved. Older publications and some newer ones refer to this as a true positive (TP) event, thereby confusing a ROC related term that does not involve search with one that does.

- The classification of each mark as either a LL or a NL is referred to as **scoring** the marks.

Definition:

- NL = non-lesion localization, i.e., a mark that is *not* close to any lesion
- LL = lesion localization, i.e., a mark that is close to a lesion

What is meant by sufficiently close? One adopts an acceptance radius (for spherical lesions) or *proximity criterion* (the more general case). What constitutes “close enough” is a clinical decision the answer to which depends on the application. It is not necessary for two radiologists to point to the same pixel in order for them to agree that they are seeing the same suspicious region. Likewise, two physicians – e.g., the radiologist finding the lesion on an x-ray and the surgeon responsible for resecting it – do not have to agree on the exact center of a lesion in order to appropriately assess and treat it. More often than not, “clinical common sense” can be used to determine if a mark actually localized the real lesion. *When in doubt, the researcher should ask an independent radiologist (i.e., not one used in the observer study) how to score ambiguous marks. A rigid definition of the proximity criterion should not be used.*

For roughly spherical nodules a simple rule can be used. If a circular lesion is 10 mm in diameter, one can use the “touching-coins” analogy to determine the criterion for a mark to be classified as lesion localization. Each coin is 10 mm in diameter, so if they touch, their centers are separated by 10 mm, and the rule is to classify any mark within 10 mm of an actual lesion center as a LL mark, and if the separation is greater, the mark is classified as a NL mark. A recent paper (Dobbins III et al., 2016) using FROC analysis gives more details on appropriate proximity criteria in the clinical context.<sup>5</sup>

#### 24.4.2 Multiple marks in the same vicinity

Multiple marks near the same vicinity are rarely encountered with radiologists, especially if the perceived lesion is mass-like.<sup>6</sup> However, algorithmic readers, such as CAD, tend to find multiple regions in the same area. Algorithm designers generally incorporate a clustering step to reduce overlapping regions to a single region and assign the highest rating to it (i.e., the rating of the highest rated mark, not the rating of the closest mark.)<sup>7</sup>

<sup>5</sup>Generally the proximity criterion is more stringent for smaller lesions than for larger ones. However, for very small lesions allowance is made so that the criterion does not penalize the radiologist for normal marking “jitter”. For 3D images the proximity criteria is different in the x-y plane vs. the slice thickness axis.

<sup>6</sup>The exception would be if the perceived lesions were speck-like objects in a mammogram, and even here radiologists tend to broadly outline the region containing perceived specks – they do not mark individual specks with great precision.

<sup>7</sup>The reason for using the highest rating is that this gives full and deserved credit for the localization. Other marks in the same vicinity with lower ratings need to be discarded from the

#### 24.4.3 Historical context

The term “free-response” was coined by (Egan et al., 1961) to describe a task involving the detection of brief audio tone(s) against a background of white-noise (white-noise is what one hears if an FM tuner is set to an unused frequency). The tone(s) could occur at any instant within an active listening interval, defined by an indicator light bulb that is turned on. The listener’s task was to respond by pressing a button at the specific instant(s) when a tone(s) was perceived (heard). The listener was uncertain how many true tones could occur in an active listening interval and when they might occur. Therefore, the number of responses (button presses) per active interval was a priori unpredictable: it could be zero, one or more. The Egan et al study did not require the listener to rate each button press, but apart from this difference and with a two-dimensional image replacing the listening interval, the acoustic signal detection study is similar to a common task in medical imaging, namely, prior to interpreting a screening case for possible breast cancer, the radiologist does not know how many diseased regions are actually present and, if present, where they are located. Consequently the case (all 4 views and possibly prior images) is searched for regions that appear to be suspicious for cancer. If one or more suspicious regions are found, and the level of suspicion of at least one of them exceeds the radiologists’ minimum reporting threshold, the radiologist reports the region(s). At my former institution (University of Pittsburgh, Department of Radiology) the radiologists digitally outline and annotate (describe) suspicious region(s) that are found. As one would expect from the low prevalence of breast cancer, in the screening context about 5 per 1000 cases in the US, and assuming expert-level radiologist interpretations, about 90% of breast cases do not generate any marks, implying case-level specificity of 90%. About 10% of cases generate one or more marks and are recalled for further comprehensive imaging (termed diagnostic workup). Of marked cases about 90% generate one mark, about 10% generate 2 marks, and a rare case generates 3 or more marks. Conceptually, a mammography report consists of the locations of regions that exceed the threshold and the corresponding levels of suspicion, reported as a Breast Imaging Reporting and Data System (BIRADS) rating (the BIRADS rating is actually assigned after the diagnostic workup following a 0-screening rating; the screening rating itself is binary: 0 for recall or 1 for normal).

Described next is the first medical imaging application of this paradigm.

---

analysis; specifically, they should not be classified as NLs, because each mark has successfully located the true lesion to within the clinically acceptable criterion, i.e., any one of them is a good decision because it would result in a patient recall and point further diagnostics to the true location.

## 24.5 A pioneering FROC study in medical imaging

This section details an FROC paradigm phantom study with x-ray images conducted in 1978 that is often overlooked. With the obvious substitution of clinical images for the phantom images, this study is a template for how an FROC experiment should be conducted. A detailed description of it is provided to set up the paradigm, the terminology used to describe it, and to introduce the FROC plot.

### 24.5.1 Image preparation

Bunch et al. conducted the first radiological free-response paradigm study using simulated lesions. They drilled 10-20 small holes (the simulated lesions) at random locations in ten 5 cm x 5 cm x 1.6 mm Teflon sheets. A Lucite plastic block 5 cm thick was placed on top of each Teflon sheet to decrease contrast and increase scatter, thereby appropriately reducing visibility of the holes (otherwise the hole detection task would be too easy; as in ROC it is important that the task not be too easy or too difficult). Imaging conditions (kVp, mAs) were chosen such that, in preliminary studies, approximately 50% of the simulated lesions were correctly localized at the observer's lowest confidence level. To minimize memory effects, the sheets were rotated, flipped or replaced between exposures. Six radiographs of 4 adjacent Teflon sheets, arranged in a 10 cm x 10 cm square, were obtained. Of these six radiographs one was used for training purposes, and the remaining five for actual data collection. Contact radiographs (i.e., with high visibility of the simulated lesions) of the sheets were obtained to establish the true lesion locations. Observers were told that each sheet contained from 0 to 30 simulated lesions. A mark had to be within about 1 mm to count as a correct localization; *a rigid definition was deemed unnecessary*. Once images had been prepared, observers interpreted them.

### 24.5.2 Image Interpretation and the 1-rating

Observers viewed each film and marked and rated any visible holes with a felt-tip pen on a transparent overlay taped to the film at one edge (this allowed the observer to view the film directly without the distracting effect of previously made marks – in digital interfaces it is important to implement a show/hide feature in the user interface). The record of mark-rating pairs generated by the observer constitutes free-response data.

The observers used a 4-point ordered rating scale with 4 representing “most likely a simulated lesion” to 1 representing “least likely a simulated lesion”. Note the meaning of the 1 rating: least likely a simulated lesion. There is confusion with some using the FROC-1 rating to mean “definitely not a lesion”. If that

Table 24.1: Comparison of ROC and FROC rating scales: note that the FROC rating is one less than the corresponding ROC rating and that there is no rating corresponding to ROC-1.

ROC Rating	Observer's Description	FROC Rating	Observer's Description
1	Definitely not diseased	NA	Image is not marked
2	Probably not diseased	1	Just possible it is a lesion
3	Possibly diseased	2	Possibly a lesion
4	Probably diseased	3	Probably a lesion
5	Definitely diseased	4	Definitely a lesion

were the observer's understanding, then logically the observer would "fill up" the entire image, especially parts outside the patient anatomy, with 1's, as each of these regions is "definitely not a lesion". Since the observer did not behave in this unreasonable way, the meaning of the FROC-1 rating, as they interpreted it, or were told, must have been "I have nothing further to report on this image".

When correctly used, the 1-rating means there is some finite, perhaps small, probability that the marked region is a lesion. In this sense the free-response rating scale is *asymmetric*. Compare the 5 rating ROC scale, where ROC-1 = "patient is definitely not diseased" and ROC-5 = "patient definitely diseased". This is a symmetric confidence level scale. In contrast the free-response confidence level scale labels different confidence levels of positivity in presence of disease. Table 24.1 compares the ROC 5-rating study to a FROC 4-rating study.

Table 24.1: comparison of ROC and FROC rating scales: note that the FROC rating is one less than the corresponding ROC rating and that there is no rating corresponding to ROC-1. The observer's way of indicating definitely non-diseased images is by simply not marking them. (NA = not available.)

The FROC rating is one less than the corresponding ROC rating because the ROC-1 rating is not used by the observer; the observer indicates such images by the simple expedient of *not* marking them.

### 24.5.3 Scoring the data

Scoring the data was defined 24.4.1 as the process of classifying each mark-rating pair as NL or LL according to the chosen proximity criterion. In the Bunch et al study, after each case was read the person running the study (i.e., Dr. Phil Bunch) compared the marks on the overlay to the true lesion locations on the contact radiographs and scored the marks as lesion localizations (LLs: lesions correctly localized to within about 1 mm radius) or non-lesion localizations (NLs: all other marks). <sup>8</sup>

---

<sup>8</sup>Bunch et al actually used the terms "true positive" and "false positive" to describe these events. This practice, still used in publications in this field, is confusing because there is

## 24.6 The free-response receiver operating characteristic (FROC) plot

The free-response receiver operating characteristic (FROC) plot was introduced, also in an auditory detection task, by Miller (Miller, 1969) as a way of visualizing performance in the free-response auditory tone detection task. In the medical imaging context, assume the mark rating pairs have been classified as NLs (non-lesion localizations) or LLs (lesion localizations).

- Non-lesion localization fraction (NLF) is defined as the total number of NLs rated at or above a threshold rating divided by the total number of cases.
- Lesion localization fraction (LLF) is defined as the total number of LLs rated at or above the same threshold rating divided by the total number of lesions in the case set.
- The FROC plot is defined as that of LLF (ordinate) vs. NLF as the threshold is varied. *Unlike the ROC plot which is completely contained in the unit square, the FROC plot is not.*
- The upper-right most operating point is termed the *observed end-point* and its coordinates are denoted ( $NLF_{max}$ ,  $LLF_{max}$ ).

While the ordinate LLF is a proper fraction, e.g., 30/40 assuming 30 LLs and 40 true lesions, the abscissa is an improper fraction that can exceed unity, like 35/21 assuming 35 NLs on 21 cases). The NLF notation is not ideal: as will become evident in the next chapter, it is used for notational symmetry and compactness.

Following Miller's suggestion, (Bunch et al., 1977b, Bunch et al. (1977a)) plotted lesion localization fraction (LLF) along the ordinate vs. non-lesion localization fraction (NLF) along the abscissa. Corresponding to the different threshold ratings, pairs of (NLF, LLF) values, or operating points on the FROC, were plotted. For example, in a positive directed four-rating FROC study, such as employed by Bunch et al, 4 FROC operating points resulted: that corresponding to marks rated 4s; that corresponding to marks rated 4s or 3s; the 4s, 3s, or 2s; and finally the 4s, 3s, 2s, or 1s. An R-rating (integer  $R > 0$ ) FROC study yields at most  $R$  operating points. So Bunch et al were able to plot only 4 operating points per reader, Fig. 6 ibid. Lacking a method of fitting a continuous FROC curve to the operating points, they did the best they could, and manually "French-curved" fitted curves. In 1986, I followed the same practice in my first paper on this topic (Chakraborty et al., 1986). In 1989 I described (Chakraborty, 1989) a method for fitting such operating points, and developed

---

ambiguity about whether these terms, commonly used in the ROC paradigm, are being applied to the case as a whole or to specific regions in the case.

software called FROCFIT, but the fitting method is obsolete, as the underlying statistical model has been superseded, and moreover, it is now known that the FROC plot is a poor visual descriptor of performance.

If continuous ratings are used, the procedure is to start with a very high threshold so that none of the ratings exceed the threshold, and one gradually lowers the threshold. Every time the threshold crosses the rating of a mark, or possibly multiple marks, the total count of LLs and NLs exceeding the threshold is divided by the appropriate denominators yielding the “raw” FROC plot. For example, when an LL rating just exceeds the threshold, the operating point jumps up by  $1/(\text{total number of lesions})$ , and if two LLs simultaneously just exceed the threshold, the operating point jumps up by  $2/(\text{total number of lesions})$ . If an NL rating just exceeds the threshold, the operating point jumps to the right by  $1/(\text{total number of cases})$ . If an LL rating and a NL rating simultaneously just exceed the threshold, the operating point moves diagonally, up by  $1/(\text{total number of lesions})$  and to the right by  $1/(\text{total number of cases})$ . The reader should get the general idea by now and recognize that the cumulating procedure is very similar to the manner in which ROC operating points were calculated, the only differences being in the quantities being cumulated and the relevant denominators.

Having seen how a binned data FROC study is conducted and scored, and the results “French-curved” as an FROC plot, typical simulated plots, generated under controlled conditions, are shown next, both for continuous ratings data and for binned rating data. Such demonstrations, that illustrate basic principles and trends, are impossible using real datasets. The reader should take my word for it (for now) that the *radiological search model (RSM)* simulator used is the simplest one possible that incorporates key elements of the search process. Details of the simulator are given in TBA Chapter 16, but for now the following summary should suffice.

## 24.7 Preview of the RSM data simulator

The RSM simulator is characterized by three parameters  $\mu$ ,  $\lambda$  and  $\nu$ . The parameter  $\nu$  characterizes the ability of the observer to *find* lesions (larger values preferred), the  $\lambda$  parameter characterizes the ability of the observer to *avoid finding* non-lesions (smaller values preferred) and parameter  $\mu$  characterizes the ability of the observer to *correctly classify* a found suspicious region as a true lesion or a non-lesion (larger values preferred). The reader may think of  $\mu$  as a *perceptual signal-to-noise ratio (pSNR)* or *conspicuity* of the lesion (similar to the separation parameter of the binormal model) that separates two normal distributions describing the z-sampling of ratings of NLs and LLs. Finally, there is a threshold parameter  $\zeta_1$  that determines if a found suspicious region is actually marked. If  $\zeta_1$  is negative infinity, then all found suspicious regions are marked and conversely, as  $\zeta_1$  increases, only those suspicious regions whose

$z$ -samples exceed  $\zeta_1$  are marked.

## 24.8 Population and binned FROC plots

Fig. 24.3 (A - C) shows simulated population FROC plots when the ratings are not binned, i.e., *raw FROC plots*, where the ratings were generated by the RJafroc function `SimulateFrocDataset()`. The help page for this function is shown below.

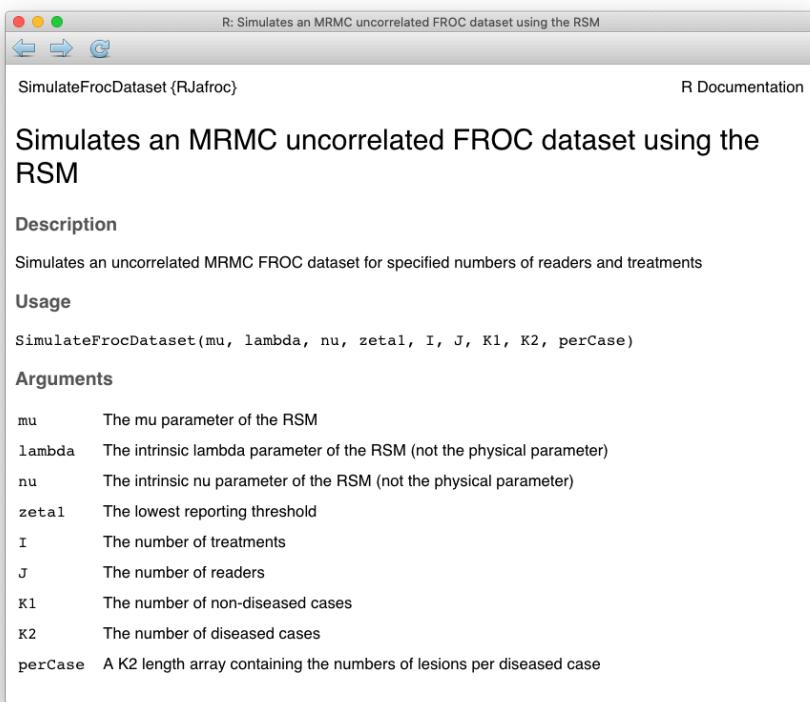


Figure 24.2: Help page for RJafroc function `SimulateFrocDataset`

For now ignore the distinction between *intrinsic* and *physical* parameters. As evident from the following code, one supplies the function with the parameters of the RSM:  $\mu, \lambda, \nu$ , the threshold parameter  $\zeta_1 = -\infty$ , the number of treatments  $I = 1$ , the number of readers  $J = 1$ , the number of non-diseased cases  $K_1$ , the number of diseased cases  $K_2$ , and the number of lesions per each diseased case  $L_{k2}$ . In this example the maximum number of lesions per case  $L_{max}$  has been specified to be two.

Single modality single reader FROC data from 10,000 non-diseased and 10,000 diseased were generated by `SimulateFrocDataset()` (the code takes a while to finish). The very large number of cases minimizes sampling variability, thereby approximating “population” curves. Additionally, the reporting threshold was set to negative infinity to ensure that all suspicious regions were marked. Plots (A) – (C) correspond to  $\mu$  equal to 0.5, 1 and 2, respectively, were generated by `PlotEmpiricalOperatingCharacteristics()`.

```

seed <- 1
set.seed(seed)
mu_arr <- c(0.5, 1, 2) # the three selected values of mu
lambda <- 1
nu <- 1
zeta1 <- -Inf
K1 <- 1000
K2 <- 1000
Lmax <- 2 # maximum number of lesions per case
Lk2 <- floor(runif(K2, 1, Lmax + 1)) # no. les. per dis. case

for (i in 1:3) {
  mu <- mu_arr[i]
  frocDataRaw <- SimulateFrocDataset(
    mu = mu,
    lambda = lambda,
    nu = nu,
    zeta1 = zeta1,
    I = 1,
    J = 1,
    K1 = K1,
    K2 = K2,
    perCase = Lk2
  )

  frocRaw <- PlotEmpiricalOperatingCharacteristics(
    dataset = frocDataRaw,
    trts= 1,
    rdrs = 1,
    opChType = "FROC",
    legend.position = "NULL"
  )

  if (i == 1) figA <- frocRaw$Plot + ggtitle("A")
  if (i == 2) figB <- frocRaw$Plot + ggtitle("B")
  if (i == 3) figC <- frocRaw$Plot + ggtitle("C")
}

```

Plots (D) – (F) correspond to 5-ratings binned data for 50 non-diseased and 70

diseased cases, and the same values of the RSM parameters as in the preceding example. The binning was performed using function `DfBinDataset()`. [Binning 20,000 cases requires much more time and is not useful.]

```

K1 <- 50
K2 <- 70
Lk2 <- floor(runif(K2, 1, Lmax + 1))
for (i in 1:3) {
  mu <- mu_arr[i]
  frocDataRaw1 <- SimulateFrocDataset(
    mu = mu,
    lambda = lambda,
    nu = nu,
    zeta1 = zeta1,
    I = 1,
    J = 1,
    K1 = K1,
    K2 = K2,
    perCase = Lk2
  )
  frocDataBin <- DfBinDataset(frocDataRaw1, desiredNumBins = 5, opChType = "FROC")

  frocBin <- PlotEmpiricalOperatingCharacteristics(
    dataset = frocDataBin,
    trts= 1,
    rdrs = 1,
    opChType = "FROC",
    legend.position = "NULL"
  )

  if (i == 1) figD <- frocBin$Plot + ggtitle("D")
  if (i == 2) figE <- frocBin$Plot + ggtitle("E")
  if (i == 3) figF <- frocBin$Plot + ggtitle("F")
}

```

Fig. 24.3: Plots (A) – (C): Population FROC plots for  $\mu = 0.5, 1, 2$ ; the other parameters are  $\lambda = 1$ ,  $\nu = 1$ ,  $\zeta_1 = -\infty$  and  $L_{\max} = 2$  is the maximum number of lesions per case in the dataset. Plots (D) – (F) correspond to 50 non-diseased and 70 diseased cases, where the data was binned into 5 bins, and other parameters are unchanged. As  $\mu$  increases, the uppermost point moves upwards and to the left (the latter trend is somewhat hidden by the changing scale factor of the x-axis).

Points to note:

- Plots (A) – (C) show quasi-continuous plots, while (D) – (F) show operating points, five per plot, connected by straight line segments, so they are

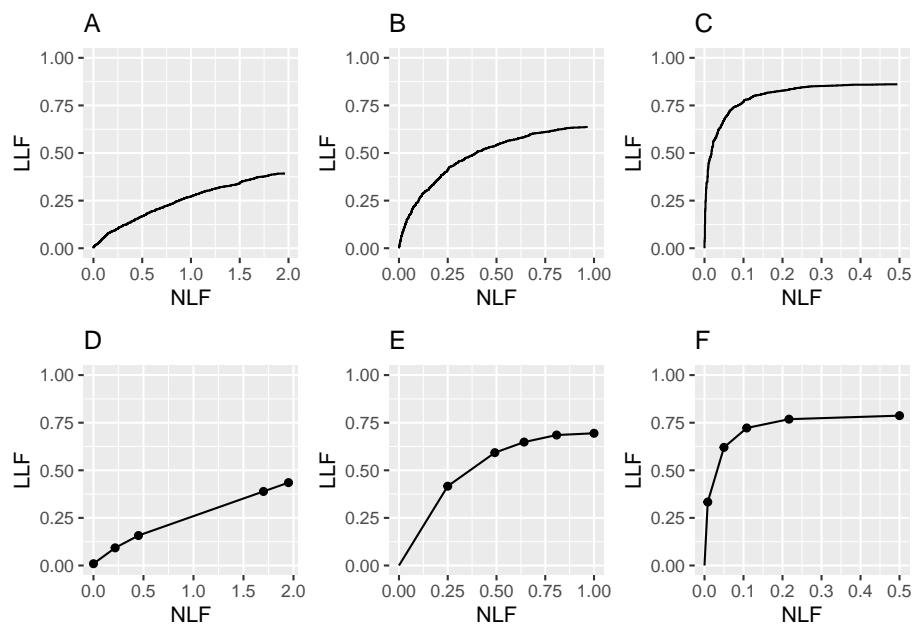


Figure 24.3: FROC plots: A, B, C correspond to raw population plots and D, E, F to binned plots with fewer cases.

termed *empirical FROC curves*, analogous to the empirical ROC curves encountered in previous chapters. At a “microscopic level” plots (A) – (C) are also discrete, but one would need to “zoom in” to see the discrete behavior (upward and rightward jumps) as each rating crosses a sliding threshold.

- The empirical plots in the bottom row (D - F) are subject to sampling variability and will not, in general, match the population plots. The reader may wish to experiment with different values of the `seed` variable in the code.
- In general FROC plots do not extend indefinitely to the right.<sup>9</sup>
- Like an ROC plot, the population FROC curve rises monotonically from the origin, initially with infinite slope (this may not be visually evident for Fig. 24.3 (A), but it is true, see next code segment). If all suspicious regions are marked, i.e.,  $\zeta_1 = -\infty$ , the plot reaches its upper-right most limit, termed the end-point, with zero slope (again, this may not be visually evident for (A), but it is true). In general these characteristics, i.e., initial infinite slope and zero final slope, are not true for the empirical plots Fig. 24.3 (D – F).

```
y <- frocRaw$Points$genOrdinate
x <- frocRaw$Points$genAbscissa
str(x)
#> num [1:2264] 0 0 0 0 0 0 0 0 0 ...
(y[2]-y[1])/(x[2]-x[1]) # slope at origin
#> [1] Inf
(y[2264]-y[2264-1])/(x[2264]-x[2264-1]) # slope at end-point
#> [1] 0
```

- Assuming all suspicious regions are marked, the end-point represents a literal end of the extent of the population FROC curve. This will become clearer in following chapters, but for now it should suffice to note that the region of the population FROC plot to the upper-right of the end-point is inaccessible to both the observer and the data analyst. [If sampling variability is significant it is possible for the observed end-point to randomly extend into this inaccessible region.]
- There is an inverse correlation between  $LLF_{max}$  and  $NLF_{max}$ , analogous to that between sensitivity and specificity in the ROC paradigm. As the perceptual SNR  $\mu$  of the lesions approaches infinity the end-point of the

---

<sup>9</sup>Fig. 5 in (Bunch et al., 1977b) is incorrect in implying, with the arrows, that the plots extend indefinitely to the right. Also there is a notation differences:  $P(TP)$  in Bunch et. al. is equivalent to  $LLF$  in this book. To avoid confusion with the  $\lambda$  parameter of the radiological search model, the variable Bunch et al. call  $\lambda$  is equivalent to  $NLF$  in this book.

FROC approaches the point (0,1), as in the next coded example, Fig. 24.4 (A). As  $\mu$  decreases the FROC curve approaches the x-axis and extends to large values along the abscissa, as in Fig. 24.4 (B). This is the “chance-level” FROC, where the reader detects few lesions, and makes many NL marks.

```

mu_arr <- c(10, 0.01)
K1 <- 1000
K2 <- 1000
Lk2 <- floor(runif(K2, 1, Lmax + 1))
for (i in 1:2) {
  mu <- mu_arr[i]
  frocDataRaw <- SimulateFrocDataset(
    mu = mu,
    lambda = lambda,
    nu = nu,
    zeta1 = zeta1,
    I = 1,
    J = 1,
    K1 = K1,
    K2 = K2,
    perCase = Lk2
  )

  frocLimits <- PlotEmpiricalOperatingCharacteristics(
    dataset = frocDataRaw,
    trts= 1,
    rdrs = 1,
    opChType = "FROC",
    legend.position = "NULL"
  )

  if (i == 1) figG <- frocLimits$Plot + ggtitle("A")
  if (i == 2) figH <- frocLimits$Plot + ggtitle("B")
}

```

Fig. 24.4: (A): FROC plot for  $\mu = 10$ . Note the small range of the NLF axis (it only extends to 0.1). In this limit the ordinate reaches unity, but the abscissa is limited to a small value. (B): This plot corresponds to  $\mu = 0.01$ , depicting near chance-level performance. Note the greatly increased traverse in the x-directions and the slight upturn in the plot near NLF = 100.

- The slope of the population FROC, assuming all suspicious regions are marked, decreases monotonically as the operating point moves up the curve, always staying non-negative, and it approaches zero, flattening out at an ordinate generally less than unity. LLF reaches unity for large  $\mu$ ,

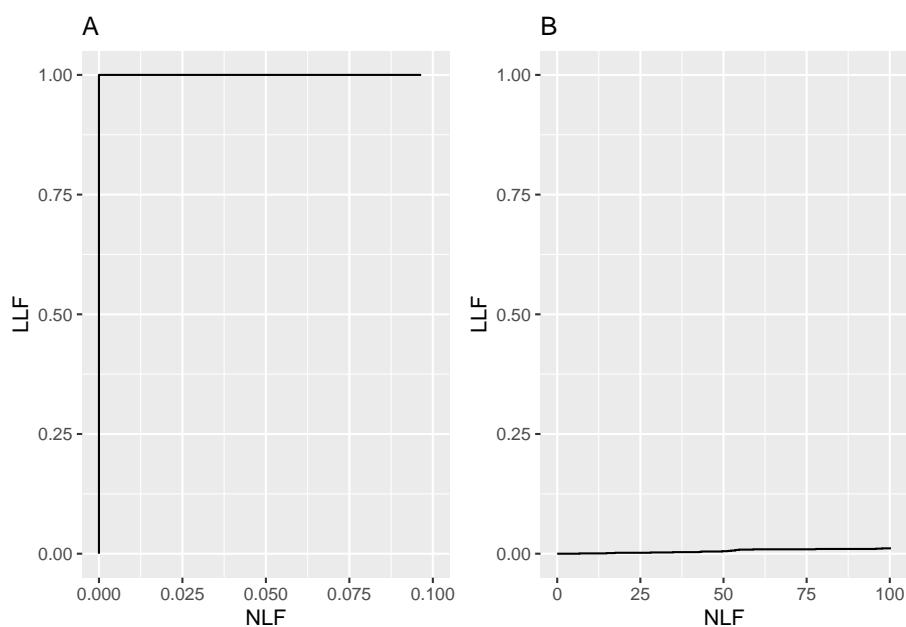


Figure 24.4: A: raw FROC curve for  $\mu = 10$ , B: raw FROC for  $\mu = 0.01$ .

which can be confirmed by setting  $\mu$  to a large value, e.g.,  $\mu = 10$ , as in Fig. 24.4 plot (A). [On the unit variance normal distribution scale, a value of 10, equivalent to 10 standard deviations, is effectively infinite.]

## 24.9 Perceptual SNR

Most readers, especially those with engineering backgrounds, are familiar with the concept of signal-to-noise-ratio, SNR. The shape and extent of the FROC plot is to a large extent determined by the *perceptual*<sup>10</sup> SNR of the lesions, pSNR, modeled by the  $\mu$  parameter. Perceptual SNR is the ratio of perceptual signal to perceptual noise. To get to perceptual variables one needs a model of the eye-brain system that transforms physical image brightness variations to corresponding perceived brightness variations, and such models exist (Van den Branden Lambrecht and Verscheure, 1996; Daly, 1993; Lubin, 1995). For uniform background images, like the phantom images used by Bunch et al, physical signal can be measured by a *template* function that has the same attenuation profile as the true lesion. Assuming the template is aligned with the lesion the *cross-correlation* between the template function and the image pixel values is related to the numerator of SNR. The cross correlation is defined as the summed product of template function pixel values times the corresponding pixel values in the actual image. Next, one calculates the cross-correlation between the template function and the pixel values in the image when the template is centered over regions known to be *lesion free*. Subtracting the mean of these values (over several lesion free regions) from the centered value gives the numerator of SNR. The denominator is the standard deviation of the cross correlation values in the lesion free areas. Details on calculating *physical* SNR are in my CAMPI (computer analysis of mammography phantom images) work (Chakraborty et al., 1999; Chakraborty and Fatouros, 1998; Chakraborty, 1997a,b). To calculate perceptual SNR one repeats these measurements but the visual process, or some model of it (e.g., the Sarnoff JNDMetrix visual discrimination model (Lubin, 1995; Siddiqui et al., 2005; Chakraborty, 2006a)), is used to filter the image prior to calculation of the cross-correlations.

An analogy may be helpful at this point. *Finding the sun in the sky is a search task, so it can be used to illustrate important concepts.*

## 24.10 The “solar” analogy: search vs. classification performance

Consider the sun, regarded as a “lesion” to be detected, with two daily observations spaced 12 hours apart, so that at least one observation period is bound

---

<sup>10</sup>Since humans make the decisions, it would be incorrect to label these as physical signal-to-noise-ratios; that is the reason for qualifying them as perceptual SNRs.

to have the sun “somewhere up there”. Furthermore, the observer is assumed to know their GPS coordinates and have a watch that gives accurate local time, from which an accurate location of the sun can be deduced. Assuming clear skies and no obstructions to the view, the sun will always be correctly located and no reasonable observer will ever generate a non-lesion localization or NL, i.e., no region of the sky will be erroneously “marked”.

FROC curve implications of this analogy are:

- Each 24-hour day corresponds to two “trials” in the (Egan et al., 1961) sense, or two cases – one diseased and one non-diseased – in the medical imaging context.
- The denominator for calculating LLF is the total number of AM days, and the denominator for calculating NLF is twice the total number of 24-hour days.
- Most important,  $LLF_{max} = 1$  and  $NLF_{max} = 0$ .

In fact, even when the sun is not directly visible due to heavy cloud cover, since the actual location of the sun can be deduced from the local time and GPS coordinates, the rational observer will still “mark” the correct location of the sun and not make any false sun localizations or “non-lesion localizations”, NLs. Consequently, even in this example  $LLF_{max} = 1$  and  $NLF_{max} = 0$ .

The conclusion is that in a task where a target is known to be present in the field of view and its location is known, the observer will always reach  $LLF_{max} = 1$  and  $NLF_{max} = 0$ . Why are LLF and NLF subscripted *max*? By randomly not marking the position of the sun even though it is visible, for example, using a coin toss to decide whether or not to mark the sun, the observer can “walk down” the y-axis of the FROC plot, reaching  $LLF = 0$  and  $NLF = 0$ . Alternatively, the observer uses a very large threshold for reporting the sun, and as this threshold is lowered the operating point “walks down” the curve. The reason for allowing the observer to “walk down” the vertical is simply to demonstrate that a continuous FROC curve from the origin to the highest point (0,1) can, in fact, be realized.

Now consider a fictitious otherwise earth-like planet where the sun can be at random positions, rendering GPS coordinates and the local time useless. All one knows is that the sun is somewhere, in the upper or lower hemispheres subtended by the sky. If there are no clouds and consequently one can see the sun clearly during daytime, a reasonable observer would still correctly locate the sun while not marking the sky with any incorrect sightings, so  $LLF_{max} = 1$  and  $NLF_{max} = 0$ . This is because, in spite of the fact that the expected location is unknown, the high contrast sun is enough to trigger the peripheral vision system, so that even if the observer did not start out looking in the correct direction, peripheral vision will drag the observer’s gaze to the correct location for foveal viewing.

The implication of this is that fundamentally different mechanisms from that considered in conventional observer performance methodology, namely *search* and *lesion-classification*, are involved. Search describes the process of *finding* the lesion while *not finding* non-lesions. Once a possible sun location has been found, classification describes the process, of recognizing that it is indeed the sun and marking it. Recall that search involves two steps: finding the object of the search and acting on it. Search and lesion-classification performances describe the abilities of an observer to efficiently perform these steps.

Think of the eye as two cameras: a low-resolution camera (peripheral vision) with a wide field-of-view plus a high-resolution camera (foveal vision) with a narrow field-of-view. If one were limited to viewing with the high-resolution camera one would spend so much time steering the high-resolution narrow field-of-view camera from spot-to-spot that one would have a hard time finding the desired stellar object. Having a single high-resolution narrow field of view vision would also have negative evolutionary consequences as one would spend so much time scanning and processing the surroundings with the narrow field of view vision that one would miss dangers or opportunities. Nature has equipped us with essentially two cameras; the first low-resolution camera is able to “digest” large areas of the surround and process it rapidly so that if danger (or opportunity) is sensed, then the eye-brain system rapidly steers the second high-resolution camera to the location of the danger (or opportunity). This is Nature’s way of optimally using the eye-brain system. For a similar reason astronomical telescopes come with a wide field of view lower resolution “spotter scope”.

Since the large field-of-view low-resolution peripheral vision system has complementary properties to the small field-of-view high-resolution foveal vision system, one expects an inverse correlation between search and lesion-classification performances. Stated generally, search involves two complementary processes: finding the suspicious regions and deciding if the found region is actually a lesion, and that there should be an inverse correlation between performance in the two tasks, see TBA Chapter 19.

When cloud cover completely blocks the fictitious random-position sun there is no stimulus to trigger the peripheral vision system to guide the fovea to the correct location. Lacking any stimulus, the observer is reduced to guessing and is led to different conclusions depending upon the benefits and costs involved. If, for example, the guessing observer earns a dollar for each LL and is fined a dollar for each NL, then the observer will likely not make any marks as the chance of winning a dollar is much smaller than losing many dollars. For this observer  $LLF_{max} = 0$  and  $NLF_{max} = 0$ , and the operating point is “stuck” at the origin. If, on the other hand, the observer is told every LL is worth a dollar and there is no penalty to NLs, then with no risk of losing the observer will “fill up” the sky with marks. In either situation the locations of the marks will lie on a grid determined by the ratio of the  $4\pi$  solid angle (subtended by the spherical sky) and the solid angle  $\Omega$  subtended by the sun. By marking every possible grid location the observer is trivially guaranteed to “detect” the sun and earn

a dollar irrespective of its random location and reach  $LLF = 1$ , but now the observer will generate lots of non-lesion localizations, so  $NLF_{\max}$  will be large:

$$NLF_{\max} = 4\pi/\Omega$$

The FROC plot for this guessing observer is the straight line joining (0,0) to  $(NLF_{\max}, 1)$ . For example, if the observer fills up half the sky then the operating point, averaged over many trials, is

$$(0.5 \times NLF_{\max}, 0.5)$$

Radiologists do not guess – there is much riding on their decisions – so in the clinical situation, if the lesion is not seen, the radiologist will not mark the image at random.

The analogy is not restricted to the sun, which one might argue is an almost infinite SNR object and therefore atypical. As another example, consider finding stars or planets. In clear skies, if one knows the constellations, one can still locate bright stars and planets like Venus or Jupiter. With less bright stars and / or obscuring clouds, there will be false-sightings and the FROC plot could approach a flat horizontal line at ordinate equal to zero, but the observer will not fill up the sky with false sightings of a desired star.

False sightings of objects in astronomy do occur. Finding a new astronomical object is a search task, where as always one can have two outcomes, correct localization (LL) or incorrect localizations (NLs). At the time of writing there is a hunt for a new planet, possibly a gas giant, that is much further than even the newly demoted Pluto. There is an astronomer in Australia who is particularly good at finding super novae (an exploding star; one has to be looking in the right region of the sky at the right time to see the relatively brief explosion). His equipment is primitive by comparison to the huge telescope at Mt. Palomar, but his advantage is that he can rapidly point his 15" telescope at a new region of the sky and thereby cover a lot more sky, in a given unit of time, than is possible with the 200" Mt. Palomar telescope. His search expertise is particularly good. Once correctly pointed at the Mt. Palomar telescope will reveal a lot more detail about the object than is possible with the smaller telescope, i.e., the analogy is to high lesion-classification accuracy. In the medical imaging context this detail (the shape of the lesion, its edge characteristics, presence of other abnormal features, etc.) allows the radiologist to diagnose whether the lesion is malignant or benign. Once again one sees that there should be an inverse correlation between search and lesion-classification performances.

## 24.11 Discussion and suggestions

This chapter has introduced the FROC paradigm, the terminology used to describe it and a common operating characteristic associated with it, namely the FROC. There are several areas of possible confusion to avoid which consider the following suggestions:

- Avoid using the term “lesion-specific” to describe location-specific paradigms.
- Avoid using the term “lesion” when one means a “suspicious region” that may or may not be a true lesion.
- Avoid using ROC-specific terms, such as true positive and false positive, that apply to the whole case, to describe location-specific terms such as lesion and non-lesion localization, that apply to localized regions of the image. This issue will come up in later chapters.
- Avoid using the FROC-1 rating to mean in effect “I see no signs of disease in this image”, when in fact it should be used as the lowest level of a reportable suspicious region. The former usage amounts to wasting a confidence level.
- Do not show FROC curves as reaching the unit ordinate, as this is the exception rather than the rule.
- Do not conceptualize FROC curves as extending to large values to the right.
- Arbitrariness of the proximity criterion and multiple marks in the same region are not clinically important. Interactions with clinicians will allow selection of an appropriate proximity criterion for the task at hand and the multiple mark problem only occurs with algorithmic observers and is readily fixed.

Additional points made in this chapter are: There is an inverse correlation between  $\text{LLF}_{\max}$  and  $\text{NLF}_{\max}$ , analogous to that between sensitivity and specificity in ROC analysis. The observed end-point ( $\text{NLF}_{\max}, \text{LLF}_{\max}$ ) of the FROC curve tends to approach the point (0,1) as the perceptual SNR of the lesions approaches infinity. The solar analogy is relevant to understanding the search task. In search tasks two types of expertise are at work: search and lesion-classification performances, and there exists an expected inverse correlation between them.

The FROC plot is the first proposed way of visually summarizing FROC data. The next chapter deals with all empirical operating characteristics that can be defined from an FROC dataset.

## 24.12 References

# Chapter 25

## Empirical plots

### 25.1 TBA How much finished

70%

### 25.2 Introduction

Operating characteristics are visual depicters of performance. If properly defined, scalar quantities derived from operating characteristics can serve as quantitative measures of performance, termed figures of merit (FOMs). The previous chapter defined the FROC curve and suggested the area under this curve as a possible FOM. This chapter introduces mathematical expressions for empirical operating characteristics (FROC and others) possible with FROC data and associated FOMs.

A distinction between latent and actual marks is made followed by a summary of FROC notation applicable to a single modality single reader dataset. This is a key table, which will be used in later chapters. Following this, different empirical operating characteristics proposed for FROC data are described. Formulae are given for calculating each empirical operating characteristic.

The observed end-point of an operating characteristic is defined as that operating point achieved by cumulating all the ratings. For the FROC plot it is demonstrated that the observed FROC curve is not contained in the unit square, unlike the other operating characteristics, which are contained in the unit square.

## 25.3 Mark rating pairs

*FROC* data consists of mark-rating pairs. Each mark indicates the location of a region suspicious enough to warrant reporting and the rating is the associated confidence level. A mark is recorded as *lesion localization* (LL) if it is sufficiently close to a true lesion, according to the adopted proximity criterion, and otherwise it is recorded as *non-lesion localization* (NL).

In an FROC study the number of marks on an image is an *a-priori* unknown modality-reader-case dependent non-negative random integer. It is incorrect to estimate it by dividing the image area by the lesion area because not all regions of the image are equally likely to have lesions, lesions do not have the same size, and perhaps most important, clinicians don't assign equal attention units to all areas of the image. The best insight into the number of marks per case is obtained from eye-tracking studies (Duchowski, 2002), but even here the information is incomplete, as eye-tracking studies can only measure foveal gaze and not lesions found by peripheral vision, not to mention that such studies are very difficult to conduct in a clinical setting.

Experts tend to have smaller numbers of NL marks per case than non-experts while maintaining equal or more LL marks per case. As an example, in screening mammography, the number of marks per case (a case is defined as 4-views, two of each breast) that an expert will consider for marking to typically less than three. About 80% on non-diseased cases have no marks. The reason is that because of the low disease prevalence marking too many cases would result in unacceptably high recall rates.

### 25.3.1 Latent vs. actual marks

To distinguish between suspicious regions that were considered for marking and regions that were actually marked, it is necessary to introduce the distinction between *latent* marks and *actual* marks.

- A *latent* mark is defined as a suspicious region, regardless of whether or not it was marked. A latent mark becomes an *actual* mark if it is marked.
- A latent mark is a latent LL if it is close to a true lesion and otherwise it is a latent NL.
- A non-diseased case can only have latent NLs. A diseased case can have latent NLs and latent LLs.
- If marked, a latent NL is recorded as an actual NL.
- If not marked, a latent NL is an *unobservable event*.
- In contrast, unmarked lesions are observable events – one knows (trivially) which lesions were not marked.

### 25.3.2 Binning rule

Recall from Section 10.3 that ROC data modeling requires the existence of a *case-dependent* decision variable, or z-sample  $z$ , and case-independent decision thresholds  $\zeta_r$ , where  $r = 0, 1, \dots, R_{ROC} - 1$  and  $R_{ROC}$  is the number of ROC study bins, and the rule that if  $\zeta_r \leq z < \zeta_{r+1}$  the case is rated  $r + 1$ . Dummy cutoffs are defined as  $\zeta_0 = -\infty$  and  $\zeta_{R_{ROC}} = \infty$ . The z-sample applies to the whole case. To summarize:

$$\left. \begin{array}{l} \text{if } (\zeta_r \leq z < \zeta_{r+1}) \Rightarrow \text{rating} = r + 1 \\ r = 0, 1, \dots, R_{ROC} - 1 \\ \zeta_0 = -\infty \\ \zeta_{R_{ROC}} = \infty \end{array} \right\} \quad (25.1)$$

Analogously, FROC data modeling requires the existence of a *case and location-dependent* z-sample for each latent mark and *case and location-independent* reporting thresholds  $\zeta_r$ , where  $r = 1, \dots, R_{FROC}$  and  $R_{FROC}$  is the number of FROC study bins, and the rule that a latent mark is marked and rated  $r$  if  $\zeta_r \leq z < \zeta_{r+1}$ . Dummy cutoffs are defined as  $\zeta_0 = -\infty$  and  $\zeta_{R_{FROC}+1} = \infty$ . For the same numbers of non-dummy cutoffs, the number of FROC bins is one less than the number of ROC bins. For example, 4 non-dummy cutoffs  $\zeta_1, \zeta_2, \zeta_3, \zeta_4$  can correspond to a 5-rating ROC study or to a 4-rating FROC study. To summarize:

$$\left. \begin{array}{l} \text{if } (\zeta_r \leq z < \zeta_{r+1}) \Rightarrow \text{rating} = r \\ r = 1, 2, \dots, R_{FROC} \\ \zeta_0 = -\infty \\ \zeta_{R_{FROC}+1} = \infty \end{array} \right\} \quad (25.2)$$

## 25.4 FROC notation

*Clear notation is vital to understanding this paradigm.* The notation needs to account for case and location dependencies of ratings and the distinction between case-level and location-level ground truth. For example, a diseased case can have several regions that are non-diseased and a few diseased regions (the lesions). The notation also has to account for cases with no marks.

FROC notation is summarized in Table 25.1, in which **all marks are latent marks**. The table is organized into three columns, the first column is the row number, the second column has the symbol(s), and the third column has the meaning(s) of the symbol(s).

Table 25.1: FROC notation; all marks refer to latent marks; see comments

Row	Symbol	Meaning
1	$t$	Case-level truth: 1 for non-diseased and 2 for diseased
2	$K_t$	Number of cases with case-level truth $t$
3	$k_t t$	Case $k_t$ in case-level truth $t$
4	$s$	Mark-level truth: 1 for NL and 2 for LL
5	$l_s s$	Mark $l_s$ in mark-level truth $s$
6	$z_{k_t t l_1 1}$	z-sample for case $k_t t$ and mark $l_1 1$
7	$z_{k_2 2 l_2 2}$	z-sample for case $k_2 2$ and mark $l_2 2$
8	$R_{FROC}$	Number of FROC bins
9	$\zeta_1$	Lowest reporting threshold
10	$\zeta_r$	Other non-dummy reporting thresholds
11	$\zeta_0, \zeta_{R_{FROC}+1}$	Dummy thresholds
12	$N_{k_t t}$	Number of NLs on case $k_t t$
13	$L_{k_2 2}$	Number of lesions on case $k_2 2$
14	$W_{k_2 2 l_2}$	Weight of lesion $l_2 2$ on case $k_2 2$
15	$L_{max}$	Maximum number of lesions per case in dataset
16	$L_T$	Total number of lesions in dataset

#### 25.4.1 Comments on Table 25.1

- Row 1: The case-truth index  $t$  refers to the case (or patient), with  $t = 1$  for non-diseased and  $t = 2$  for diseased cases. As a useful mnemonic,  $t$  is for *truth*.
- Row 2:  $K_t$  is the number of cases with truth state  $t$ ; specifically,  $K_1$  is the number of non-diseased cases and  $K_2$  the number of diseased cases.
- Row 3: Two indices  $k_t t$  are needed to select case  $k_t$  in truth state  $t$ . As a useful mnemonic,  $k$  is for *case*.
- Rows 4 and 5: For a similar reason, two indices  $l_s s$  are needed to select latent mark  $l_s$  in location level truth state  $s$ , where  $s = 1$  corresponds to a latent NL and  $s = 2$  corresponds to a latent LL. One can think of  $l_s$  as indexing the locations of different latent marks with location-level truth state  $s$ . As a useful mnemonic,  $l$  is for *location*.
  - $l_1 = \{1, 2, \dots, N_{k_t t}\}$  indexes latent NL marks, provided the case has at least one NL mark, and otherwise  $N_{k_t t} = 0$  and  $l_1 = \emptyset$ , the null set.
  - The possible values of  $l_1$  are  $l_1 = \{\emptyset\} \oplus \{1, 2, \dots, N_{k_t t}\}$ . The null set applies when the case has no latent NL marks and  $\oplus$  is the “exclusive-or” symbol (“exclusive-or” is used in the English sense: “one or the

other, but not neither nor both"). In other words,  $l_1$  can *either* be the null set or take on values  $1, 2, \dots, N_{k_t t}$ .

- Likewise,  $l_2 = \{1, 2, \dots, L_{k_2 2}\}$  indexes latent LL marks. Unmarked LLs are assigned negative infinity ratings. The null set notation is not needed for latent LLs.
- Row 6: The z-sample for case  $k_t t$  and **latent NL mark**  $l_1 1$  is denoted  $z_{k_t t l_1 1}$ . Latent NL marks are possible on non-diseased and diseased cases (both values of  $t$  are allowed). The range of a z-sample is  $-\infty < z_{k_t t l_1 1} < \infty$ , provided  $l_1 \neq \emptyset$ ; otherwise, it is an *unobservable event*.
- Row 7: The z-sample of a **latent LL** is  $z_{k_2 2 l_2 2}$ . Unmarked lesions are assigned negative infinity ratings and are observable events. The null-set notation is unnecessary for them.
- Row 8:  $R_{FROC}$  is the number of bins in the FROC study.
- Rows 9, 10 and 11: The cutoffs in the FROC study. The lowest threshold is  $\zeta_1$ . The other non-dummy thresholds are  $\zeta_r$  where  $r = 2, 3, \dots, R_{FROC}$ . The dummy thresholds are  $\zeta_0 = -\infty$  and  $\zeta_{R_{FROC}+1} = \infty$ .
- Row 12:  $N_{k_t t}$  is the total number of latent NL marks on case  $k_t t$ .
- Row 13:  $L_{k_2 2}$  is the number of lesions in diseased case  $k_2 2$ .
- Row 14:  $W_{k_2 l_2}$  is the weight (i.e., clinical importance) of lesion  $l_2 2$  in diseased case  $k_2 2$ . The weights of lesions on a case sum to one:  $\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} = 1$ .
- Row 15:  $L_{max}$  is the maximum number of lesions per case in the dataset.
- Row 16:  $L_T$  is the total number of lesions in the dataset.

#### 25.4.2 Discussion: cases with zero latent NL marks

An aspect of FROC data, **that there could be cases with no NL marks, no matter how low the reporting threshold**, has created problems both from conceptual and notational viewpoints. Taking the conceptual issue first, my thinking (prior to 2004) was that as the reporting threshold  $\zeta_1$  is lowered, the number of NL marks per case increases almost indefinitely. I visualized this process as each case "filling up" with NL marks<sup>1</sup>. In fact the first modeling of FROC data (Chakraborty, 1989) predicts that, as the reporting threshold is lowered to  $\zeta_1 = -\infty$ , the number of NL marks per case approaches  $\infty$ . However, observed FROC curves end with a finite value of NLs per case. This

---

<sup>1</sup>I expected the number of NL marks per image to be limited only by the ratio of image size to lesion size, i.e., larger values for smaller lesions.

mismatch between observation and theory is one reason I introduced the radiological search model (RSM) (Chakraborty, 2006b). I will have much more to say about this in a subsequent chapter, but for now I state one prediction (actually an assumption) of the RSM: the number of latent NL marks is a Poisson distributed random integer with a finite value for the mean parameter of the Poisson distribution. This means that the actual number of latent NL marks per case can be 0, 1, 2, ..., whose average (over cases) is a finite number. With this background, let us return to the conceptual issue: why does the observer not keep “filling-up” the image with NL marks? The answer is that **the observer can only mark regions that have a non-zero chance of being a lesion**. For example, if the actual number of latent NLs on a particular case is 2, then, as the reporting threshold is lowered, the observer will make at most two NL marks. Having exhausted these two regions the observer will not mark any more regions because there are no more regions to be marked - *all other regions in the image have, in the perception of the observer, zero chance of being a lesion.*

The notational issue is how to handle images with no latent NL marks. Basically it involves restricting summations over cases  $k_t t$  to those cases which have at least one latent NL mark, i.e.,  $N_{k_t t} \neq 0$ . This is illustrated in the next section.

## 25.5 The empirical FROC

The FROC was defined, Chapter 24, as the plot of LLF (along the ordinate) vs. NLF (along the abscissa).

Using the notation of Table 25.1 and assuming binned data<sup>2</sup>, then, corresponding to the operating point determined by threshold  $\zeta_r$ , the FROC abscissa is  $\text{NLF}_r \equiv \text{NLF}(\zeta_r)$ , the total number of NLs rated  $\geq$  threshold  $\zeta_r$  divided by the total number of cases, and the corresponding ordinate is  $\text{LLF}_r \equiv \text{LLF}(\zeta_r)$ , the total number of LLs rated  $\geq$  threshold  $\zeta_r$  divided by the total number of lesions:

$$\text{NLF}_r = \frac{n(\text{NLs rated } \geq \zeta_r)}{n(\text{cases})} \quad (25.3)$$

and

$$\text{LLF}_r = \frac{n(\text{LLs rated } \geq \zeta_r)}{n(\text{lesions})} \quad (25.4)$$

---

<sup>2</sup>This is not a limiting assumption: if the data is continuous, for finite numbers of cases, no ordering information is lost if the number of ratings is chosen large enough. This is analogous to Bamber’s theorem in Chapter 05, where a proof, although given for binned data, is applicable to continuous data.

The observed operating points correspond to the following values of  $r$ :

$$r = 1, 2, \dots, R_{FROC} \quad (25.5)$$

Due to the ordering of the thresholds, i.e.,  $\zeta_1 < \zeta_2 \dots < \zeta_{R_{FROC}}$ , higher values of  $r$  correspond to lower operating points. The uppermost operating point, i.e., that defined by  $r = 1$ , is referred to as the *observed end-point*.

Equations (25.3) and (25.4) are equivalent to:

$$NLF_r = \frac{1}{K_1 + K_2} \sum_{t=1}^2 \sum_{k_t=1}^{K_t} \mathbb{I}(N_{k_t t} \neq 0) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}(z_{k_t t l_1 1} \geq \zeta_r) \quad (25.6)$$

and

$$LLF_r = \frac{1}{L_T} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}(z_{k_2 2 l_2 2} \geq \zeta_r) \quad (25.7)$$

Each indicator function,  $\mathbb{I}()$ , yields unity if the argument is true and zero otherwise.

In Eqn. (25.6)  $\mathbb{I}(N_{k_t t} \neq 0)$  ensures that **only cases with at least one latent NL** are counted. Recall that  $N_{k_t t}$  is the total number of latent NLs in case  $k_t t$ . Not including this term would cause the summation over  $l_1$  to be undefined for cases with zero latent NLs. The term  $\mathbb{I}(z_{k_t t l_1 1} \geq \zeta_r)$  counts over all NL marks with ratings  $\geq \zeta_r$ . The three summations yield the total number of NLs in the dataset with z-samples  $\geq \zeta_r$  and dividing by the total number of cases yields  $NLF_r$ . This equation also shows explicitly that NLs on both non-diseased ( $t = 1$ ) and diseased ( $t = 2$ ) cases contribute to NLF.

In Eqn. (25.7) a summation over  $t$  is not needed as only diseased cases contribute to LLF. Analogous to the first indicator function term in Eqn. (25.6), a term like  $\mathbb{I}(L_{k_2 2} \neq 0)$  would be superfluous since  $L_{k_2 2} > 0$ , as each diseased case must have at least one lesion. The term  $\mathbb{I}(z_{k_2 2 l_2 2} \geq \zeta_r)$  counts over all LL marks with ratings  $\geq \zeta_r$ . Dividing by  $L_T$ , the total number of lesions in the dataset, yields  $LLF_r$ .

### 25.5.1 Definition

The empirical FROC plot connects adjacent operating points  $(NLF_r, LLF_r)$ , including the origin  $(0,0)$  and the observed end-point, with straight lines. The area under this plot is the empirical FROC AUC, denoted  $A_{FROC}$ .

### 25.5.2 The origin, a trivial point

Since  $\zeta_{R_{FROC}+1} = \infty$  according to Eqn. (25.6) and Eqn. (25.7),  $r = R_{FROC} + 1$  yields the trivial operating point  $(0,0)$ .

### 25.5.3 The observed end-point and its semi-constrained property

The abscissa of the observed end-point  $NLF_1$ , is defined by:

$$NLF_1 = \frac{1}{K_1 + K_2} \sum_{t=1}^2 \sum_{k_t=1}^{K_t} \mathbb{I}(N_{k_t t} \neq 0) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}(z_{k_t t l_1 1} \geq \zeta_1) \quad (25.8)$$

Since each case could have an arbitrary number of NLs,  $NLF_1$  need not equal unity, except fortuitously.

The ordinate of the observed end-point  $LLF_1$ , is defined by:

$$\left. \begin{aligned} LLF_1 &= \frac{\sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}(z_{k_2 2 l_2 2} \geq \zeta_1)}{L_T} \\ &\leq 1 \end{aligned} \right\} \quad (25.9)$$

The numerator is the total number of lesions that were actually marked. The ratio is the fraction of lesions that are marked, which is  $\leq 1$ .

This is the **semi-constrained property of the observed end-point**, namely, while the observed end-point *ordinate* is constrained to the range  $(0,1)$  the corresponding *abscissa* is not so constrained.

### 25.5.4 Futility of extrapolation outside the observed end-point

To understand this consider the expression for  $NLF_0$ , i.e., using Eqn. (25.6) with  $r = 0$ :

$$NLF_0 = \frac{1}{K_1 + K_2} \sum_{t=1}^2 \sum_{k_t=1}^{K_t} \mathbb{I}(N_{k_t t} \neq 0) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}(z_{k_t t l_1 1} \geq -\infty) \quad (25.10)$$

The right hand side of this equation can be separated into two terms, the contribution of latent NLs with z-samples in the range  $z \geq \zeta_1$  and those in the range  $-\infty \leq z < \zeta_1$ . The first term yields the abscissa of the observed end-point, Eqn. (25.8). The 2nd term is:

$$\left. \begin{aligned}
 \text{2nd term} &= \left( \frac{1}{K_1 + K_2} \right) \sum_{t=1}^2 \sum_{k_t=1}^{K_t} \mathbb{I}(N_{k_t t} \neq 0) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}(-\infty \leq z_{k_t t l_1} < \zeta_1) \\
 &= \frac{\text{unknown number}}{K_1 + K_2}
 \end{aligned} \right\} \quad (25.11)$$

It represents the contribution of unmarked NLs, i.e., latent NLs whose z-samples were below  $\zeta_1$ . It determines how much further to the right the observer's NLF would have moved, relative to  $NLF_1$ , if one could get the observer to lower the reporting criterion to  $-\infty$ . **Since the observer may not oblige, this term cannot, in general, be evaluated.** Therefore  $NLF_0$  cannot be evaluated. The basic problem is that **unmarked latent NLs represent unobservable events.**

Turning our attention to  $LLF_0$ :

$$\left. \begin{aligned}
 LLF_0 &= \frac{\sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}(z_{k_2 2 l_2} \geq -\infty)}{L_T} \\
 &= 1
 \end{aligned} \right\} \quad (25.12)$$

Unlike unmarked latent NLs, **unmarked lesions can safely be assigned the  $-\infty$  rating, because an unmarked lesion is an observable event.** The right hand side of Eqn. (25.12) evaluates to unity. However, since the corresponding abscissa  $NLF_0$  is undefined, one cannot plot this point. It follows that one cannot extrapolate outside the observed end-point.

The formalism should not obscure the fact that the futility of extrapolation outside the observed end-point of the FROC is a fairly obvious property: one does not know how far to the right the abscissa of the observed end-point might extend if one could get the observer to report every latent NL, no matter how low its z-sample.

## 25.6 The inferred ROC plot

By adopting a sensible rule for converting the zero or more mark-rating data per case to a single rating per case, and commonly the highest rating rule is used<sup>3</sup>, it is possible to infer ROC data from FROC mark-rating data.

---

<sup>3</sup>The highest rating method was used in early FROC modeling in (Bunch et al., 1977b) and in (Swensson, 1996b), the latter in the context of LROC paradigm modeling.

### 25.6.1 Inferred-ROC rating

The rating of the highest rated mark on a case, or  $-\infty$  if the case has no marks, is defined as the inferred-ROC rating for the case. Inferred-ROC ratings on non-diseased cases are referred to as inferred-FP ratings and those on diseased cases as inferred-TP ratings.

When there is little possibility for confusion, the prefix “inferred” is suppressed. Using the by now familiar cumulation procedure, FP counts are cumulated to calculate FPF and likewise, TP counts are cumulated to calculate TPF.

Definitions:

- $FPF(\zeta)$  = cumulated inferred FP counts with z-sample  $\geq$  threshold  $\zeta$  divided by total number of non-diseased cases.
- $TPF(\zeta)$  = cumulated inferred TP counts with z-sample  $\geq$  threshold  $\zeta$  divided by total number of diseased cases

Definition of ROC plot:

- The ROC is the plot of inferred  $TPF(\zeta)$  vs. inferred  $FPF(\zeta)$ .
- The plot includes a **straight line extension from the observed endpoint to (1,1)**.

The mathematical definition of the ROC follows.

### 25.6.2 Inferred FPF

The highest z-sample ROC false positive (FP) rating for non-diseased case  $k_1$  is defined by:

$$FP_{k_11} = \max_{l_1} \left( z_{k_11l_11} \mid l_1 \neq \emptyset \right) \quad \left. \begin{array}{l} \\ = -\infty \mid l_1 = \emptyset \end{array} \right\} \quad (25.13)$$

If the case has at least one latent NL mark, then  $l_1 \neq \emptyset$ , where  $\emptyset$  is the null set, and the first definition applies. If the case has no latent NL marks, then  $l_1 = \emptyset$ , and the second definition applies.  $FP_{k_11}$  is the maximum z-sample over all latent marks occurring on non-diseased case  $k_1$ , or  $-\infty$  if the case has no latent marks. The corresponding false positive fraction is defined by:

$$FPF_r \equiv FPF(\zeta_r) = \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}(FP_{k_11} \geq \zeta_r) \quad (25.14)$$

### 25.6.3 Inferred TPF

The inferred true positive (TP) z-sample for diseased case  $k_2 2$  is defined by:

$$TP_{k_2 2} = \max_{l_1 l_2} ((z_{k_2 2 l_1 2}, z_{k_2 2 l_2 2}) \mid l_1 \neq \emptyset) \quad (25.15)$$

or

$$TP_{k_2 2} = \max_{l_2} (z_{k_2 2 l_2 2}) \mid (l_1 = \emptyset \wedge (\max_{l_2} (z_{k_2 2 l_2 2}) \neq -\infty)) \quad (25.16)$$

or

$$TP_{k_2 2} == -\infty \mid (l_1 = \emptyset \wedge (\max_{l_2} (z_{k_2 2 l_2 2}) = -\infty)) \quad (25.17)$$

Here  $\wedge$  is the logical AND operator.

- If  $l_1 \neq \emptyset$  then Eqn. (25.15) applies, i.e., one takes the maximum over all ratings, NLs and LLs, whichever is higher, occurring on the diseased case.
- If  $l_1 = \emptyset$  and at least one lesion is marked, then Eqn. (25.16) applies, i.e., one takes the maximum over all marked LLs.
- If  $l_1 = \emptyset$  and no lesions are marked, then Eqn. (25.17) applies; this represents an unmarked diseased case; the  $-\infty$  rating assignment is justified because an unmarked diseased case is an observable event.

The inferred true positive fraction  $TPF_r$  is defined by:

$$TPF_r \equiv TPF(\zeta_r) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \mathbb{I}(TP_{k_2 2} \geq \zeta_r) \quad (25.18)$$

### 25.6.4 Definition

The inferred empirical ROC plot connects adjacent points  $(FPF_r, TPF_r)$ , including the origin  $(0,0)$ , with straight lines plus a straight-line segment connecting the observed end-point to  $(1,1)$ . Like a real ROC, this plot is constrained to lie within the unit square. The area under this plot is the empirical inferred ROC AUC, denoted  $A_{ROC}$ .

## 25.7 The alternative FROC (AFROC) plot

- Fig. 4 in (Bunch et al., 1977b) anticipated another way of visualizing FROC data. I subsequently termed<sup>4</sup> this the *alternative FROC (AFROC)* plot (Chakraborty, 1989).
- The empirical AFROC is defined as the plot of  $\text{LLF}(\zeta_r)$  along the ordinate vs.  $\text{FPF}(\zeta_r)$  along the abscissa.
- $\text{LLF}_r \equiv \text{LLF}(\zeta_r)$  was defined in Eqn. (25.7).
- $\text{FPF}_r \equiv \text{FPF}(\zeta_r)$  was defined in Eqn. (25.14).

### 25.7.1 Definition

The empirical AFROC plot connects adjacent operating points  $(\text{FPF}_r, \text{LLF}_r)$ , including the origin  $(0,0)$  and  $(1,1)$ , with straight lines. The area under this plot is the empirical inferred AFROC AUC, denoted  $A_{\text{AFROC}}$ .

Key points:

- The ordinates LLF of the FROC and AFROC are identical.
- The abscissa FPF of the ROC and AFROC are identical.
- The AFROC is, in this sense, a hybrid plot, incorporating aspects of both ROC and FROC plots.
- Unlike the empirical FROC, whose observed end-point has the semi-constrained property, **the AFROC end-point is constrained to within the unit square**.

### 25.7.2 The constrained observed end-point of the AFROC

Since  $\zeta_{R_{\text{FROC}}+1} = \infty$ , according to Eqn. (25.7) and Eqn. (25.14),  $r = R_{\text{FROC}} + 1$  yields the trivial operating point  $(0,0)$ . Likewise, since  $\zeta_0 = -\infty$ ,  $r = 0$  yields the trivial point  $(1,1)$ :

$$\left. \begin{aligned} \text{FPF}_{R_{\text{FROC}}+1} &= \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}(FP_{k_11} \geq \infty) \\ &= 0 \\ \text{LLF}_{R_{\text{FROC}}+1} &= \frac{1}{L_T} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_22}} \mathbb{I}(LL_{k_22l_22} \geq \infty) \\ &= 0 \end{aligned} \right\} \quad (25.19)$$

---

<sup>4</sup>The late Prof. Richard Swensson did not like my choice of the word “alternative” in naming this operating characteristic. I had no idea in 1989 how important this operating characteristic would later turn out to be, otherwise a more meaningful name might have been proposed.

and

$$\left. \begin{aligned} \text{FPF}_0 &= \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}(FP_{k_11} \geq -\infty) \\ &= 1 \\ \text{LLF}_0 &= \frac{1}{L_T} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_22}} \mathbb{I}(LL_{k_22l_22} \geq -\infty) \\ &= 1 \end{aligned} \right\} \quad (25.20)$$

Because every non-diseased case is assigned a rating, and is therefore counted, the right hand side of the first equation in (25.20) evaluates to unity. This is obvious for marked cases. Since each unmarked case also gets a rating, albeit a  $-\infty$  rating, it is also counted (the argument of the indicator function in Eqn. (25.20) is true even when the inferred FP rating is  $-\infty$ ).

## 25.8 The weighted-AFROC (wAFROC) plot

The AFROC ordinate defined in Eqn. (25.7) gives equal importance to every lesion on a case. Therefore, a case with more lesions will have more influence on the AFROC (see TBA Chapter 14 for an explicit demonstration of this fact). This is undesirable since each case (i.e., patient) should get equal importance in the analysis. As with ROC analysis, one wishes to draw conclusions about the population of cases and each case is regarded as an equally valid sample from the population. In particular, one does not want the analysis to be skewed towards cases with greater than the average number of lesions.<sup>5</sup>

Another issue is that the AFROC assigns equal clinical importance to each lesion in a case. Lesion weights were introduced (Chakraborty and Berbaum, 2004) to allow for the possibility that the clinical importance of finding a lesion might be lesion-dependent (Chakraborty and Yoon, 2009). For example, it is possible that a diseased cases has lesions of two types with differing clinical importance; the figure-of-merit should give more credit to finding the more clinically important one. Clinical importance could be defined as the mortality associated with the specific lesion type; these can be obtained from epidemiological studies (DeSantis et al., 2011).

Let  $W_{k_2l_2} \geq 0$  denote the **weight** (i.e., clinical importance) of lesion  $l_2$  in diseased case  $k_2$  (since weights are only applicable to diseased cases, one can, without ambiguity, drop the case-level and location-level truth subscripts, i.e.,

---

<sup>5</sup>Historical note: I became aware of how serious this issue could be when a researcher contacted him about using FROC methodology for nuclear medicine bone scan images, where the number of lesions on diseased cases can vary from a few to a hundred!

the notation  $W_{k_2l_22}$  would be superfluous). For each diseased case  $k_22$  the weights are subject to the constraint:

$$\sum_{l_2=1}^{L_{k_22}} W_{k_2l_2} = 1 \quad (25.21)$$

The constraint assures that the each diseased case exerts equal importance in determining the weighted-AFROC (wAFROC) operating characteristic, regardless of the number of lesions in it (see TBA Chapter 14 for a demonstration of this fact).

The weighted lesion localization fraction  $wLLF_r$  is defined by (Chakraborty and Zhai, 2016):

$$wLLF_r \equiv wLLF(\zeta_r) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_22}} W_{k_2l_2} \mathbb{I}(z_{k_2l_2} \geq \zeta_r) \quad (25.22)$$

### 25.8.1 Definition

The empirical wAFROC plot connects adjacent operating points  $(FPF_r, wLLF_r)$ , including the origin  $(0,0)$ , with straight lines plus a straight-line segment connecting the observed end-point to  $(1,1)$ . The area under this plot is the empirical weighted-AFROC AUC, denoted  $A_{wAFROC}$ .

## 25.9 The AFROC1 plot

Historically the AFROC originally used a different definition of FPF, which is retrospectively termed the AFROC1 plot. Since NLs can occur on diseased cases, it is possible to define an inferred “FP” rating on a *diseased case* as the maximum of all NL ratings on the case, or  $-\infty$  if the case has no NLs. The quotes emphasize that this is non-standard usage of ROC terminology: in an ROC study, a FP can only occur on a *non-diseased case*. Since both case-level truth states are allowed, the highest false positive (FP) z-sample for case  $k_t t$  is [the “1” superscript below is necessary to distinguish it from Eqn. (25.13)]:

$$\left. \begin{aligned} FP_{k_t t}^1 &= \max_{l_1} (z_{k_t t l_1 1} \mid l_1 \neq \emptyset) \\ &= -\infty \mid l_1 = \emptyset \end{aligned} \right\} \quad (25.23)$$

$FP_{k_t t}^1$  is the maximum over all latent NL marks, labeled by the location index  $l_1$ , occurring on case  $k_t t$ , or  $-\infty$  if  $l_1 = \emptyset$ . The corresponding false positive

fraction  $FPF_r^1$  is defined by [the “1” superscript is necessary to distinguish it from Eqn. (25.14)]:

$$FPF_r^1 \equiv FPF_r^1(\zeta_r) = \frac{1}{K_1 + K_2} \sum_{t=1}^2 \sum_{k_t=1}^{K_t} \mathbb{I}(FP_{k_t t}^1 \geq \zeta_r) \quad (25.24)$$

Note the subtle differences between Eqn. (25.14) and Eqn. (25.24). The latter counts “FPs” on non-diseased and diseased cases while Eqn. (25.14) counts FPs on non-diseased cases only, and for that reason the denominators in the two equations are different. The advisability of allowing a diseased case to be both a TP and a FP is questionable from both clinical and statistical considerations. However, this operating characteristic can be useful in applications where all cases contain lesions, for example lesion localization plus classification tasks (See Chapter TBA).

### 25.9.1 Definition

The empirical AFROC1 plot connects adjacent operating points  $(FPF_r^1, LLF_r)$ , including the origin  $(0,0)$  and  $(1,1)$ , with straight lines. The only difference between AFROC1 and the AFROC plot is in the x-axis. The area under this plot is the empirical AFROC1 AUC, denoted  $A_{\text{AFROC1}}$ .

## 25.10 The weighted-AFROC1 (wAFROC1) plot

### 25.10.1 Definition

The empirical weighted-AFROC1 (wAFROC1) plot connects adjacent operating points  $(FPF_r^1, wLLF_r)$ , including the origin  $(0,0)$  and  $(1,1)$ , with straight lines. The only difference between it and the wAFROC plot is in the x-axis. The area under this plot is the empirical weighted-AFROC AUC, denoted  $A_{\text{wAFROC1}}$ .

## 25.11 The EFROC plot

An *exponentially transformed FROC* (EFROC) plot has been proposed (Popescu, 2011) that, like the AFROC, is contained within the unit square. The EFROC inferred FPF is defined by (this represents another way of inferring ROC data, albeit only FPF, from FROC data):

$$FPF_r = 1 - \exp(NLF(\zeta_r)) \quad (25.25)$$

In other words, one computes  $NLF_r$  using NLs rated  $\geq \zeta_r$  on all cases and then transforms it to  $FPF_r$  using the exponential transformation shown. Note that  $FPF_r$  so defined is in the range  $(0,1)$ .

### 25.11.1 Definition

The empirical EFROC plot connects adjacent operating points  $(FPF_r^1, LLF_r)$ , including the origin  $(0,0)$  and  $(1,1)$ , with straight lines. The only difference between it and the AFROC plot is in the x-axis. The area under this plot is the empirical EFROC AUC, denoted  $A_{\text{EFROC}}$ .

$A_{\text{EFROC}}$  has the advantage, compared to  $A_{\text{FROC}}$ , of being defined by points contained within the unit square. It has the advantage over the AFROC of using all NL ratings, not just the highest rated ones. In my opinion this is a mixed blessing. The effect on statistical power compared to  $A_{\text{AFROC}}$  has not been studied, but I expect the advantage to be minimal (because the highest rated NL contains more information than a randomly selected NL mark). A disadvantage is that cases with more LLs get more importance in the analysis; this can be corrected by replacing LLF with wLLF, essentially yielding a weighted version of the EFROC AUC. Another disadvantage is that inclusion of NLs on diseased cases causes the EFROC AUC to depend on diseased prevalence. *The EFROC represents the first recognition by someone other than me, of significant limitations of the FROC curve, and that an operating characteristic for FROC data that is completely contained within the unit square is highly desirable.*

## 25.12 Discussion

TBA This chapter started with the difference between latent and actual marks and the notation to describe FROC data. The notation is used in deriving formulae for FROC, inferred ROC, AFROC, wAFROC, AFROC1, wAFROC1 and EFROC operating characteristics. In each case an area measure was defined. With the exception of the FROC plot, all operating characteristics defined in this chapter are contained in the unit square. Discussion of the preferred operating characteristic is deferred to a subsequent chapter TBA.

## 25.13 References

# Chapter 26

## Empirical plot examples

### 26.1 TBA How much finished

50%

### 26.2 Introduction

The previous chapter introduced definitions and formulae for the various operating characteristics possible with FROC data. This chapter illustrates these definitions with numerical values and plots. The RSM simulator, introduced in Section 24.7, is used to generate FROC datasets under controlled conditions. Structure of the FROC dataset. TBA.

The starting point is the FROC plot.

### 26.3 Raw FROC/AFROC/ROC plots

*Raw plots* correspond to the actual simulator generated floating-point ratings, prior to any binning operation. If binning is employed the plots are termed *binned plots*. The FROC plots shown below were generated using the data simulator introduced in Chapter 24. The examples are similar to the population FROC curves shown in that chapter but the emphasis here is on understanding the FROC data structure. To this end smaller numbers of cases, not 20,000 as in the previous chapter, are used. Examples are given using smaller datasets. With a very small dataset, the logic of constructing the plot is more transparent but the operating points are more susceptible to sampling variability. The examples illustrate key points distinguishing the free-response paradigm from ROC. TBA

### 26.3.1 Code for raw plots

```

1  seed <- 1; set.seed(seed)
2  mu <- 1
3  lambda <- 1
4  nu <- 1
5  zeta1 <- -1
6  K1 <- 5
7  K2 <- 7
8  Lmax <- 2
9  Lk2 <- floor(runif(K2, 1, Lmax + 1))
10
11 frocDataRaw <- SimulateFrocDataset(
12   mu = mu,
13   lambda = lambda,
14   nu = nu,
15   I = 1,
16   J = 1,
17   K1 = K1,
18   K2 = K2,
19   perCase = Lk2,
20   zeta1 = zeta1,
21   seed = seed
22 )
23
24 p1A <- PlotEmpiricalOperatingCharacteristics(
25   dataset = frocDataRaw,
26   trts= 1, rdrs = 1, opChType = "FROC",
27   legend.position = "NULL")$Plot + ggtitle("A")
28
29 p1B <- PlotEmpiricalOperatingCharacteristics(
30   dataset = frocDataRaw,
31   trts= 1, rdrs = 1, opChType = "AFROC",
32   legend.position = "NULL")$Plot + ggtitle("B")
33
34 p1C <- PlotEmpiricalOperatingCharacteristics(
35   dataset = frocDataRaw,
36   trts= 1, rdrs = 1, opChType = "ROC",
37   legend.position = "NULL")$Plot + ggtitle("C")
38
39 frocDataRaw_1_5_7 <- frocDataRaw # seed 1, K1 = 5, K2 = 7

```

### 26.3.2 Explanation of the code

Line 1 sets the seed of the random number generator. Lines 2-5 set the simulator parameters  $\mu = 1$ ,  $\lambda = 1$ ,  $\nu = 1$ ,  $\zeta_1 = -1$ . Briefly,  $\mu$  determines the separation of two unit variance Gaussians, the one centered at zero determines the z-samples of latent NLs, while the one centered at  $\mu$  determines the z-samples of latent LLs.  $\lambda$  is the mean parameter of a Poisson distribution determining the number (a random non-negative integer) of latent NLs on each case while  $\nu$ , the success probability of a binomial distribution, determines the number of latent LLs on each diseased case. A latent NL or LL is marked if its z-sample  $\geq \zeta_1$ .

Lines 6-7 set the number of non-diseased cases  $K_1 = 5$  and the number of diseased cases  $K_2 = 7$ .

Line 8 sets the maximum number of lesions per diseased case  $L_{max} = 2$ . Line 9 randomly samples a uniform distribution to obtain the actual number of lesions per diseased case Lk2. The following code illustrates the process.

#### 26.3.2.1 Number of lesions per diseased case

```
Lk2
#> [1] 1 1 2 2 1 2 2
sum(Lk2)
#> [1] 11
max(floor(runif(1000, 1, Lmax + 1)))
#> [1] 2
```

This shows that the first two diseased cases have one lesion each, the third and fourth have two lesions each, etc. The total number of lesions in the dataset is 11. The last two lines of the code snippet show that, even with a thousand simulations, the number of lesions per diseased case is indeed limited to  $L_{max} = 2$ .

#### 26.3.2.2 The structure of the FROC dataset

Lines 11-21 uses the function `SimulateFrocDataset` to simulate the dataset object `frocDataRaw`. Its structure is examined next:

```
str(frocDataRaw)
#> List of 3
#> $ ratings      :List of 3
#> ..$ NL    : num [1, 1, 1:12, 1:3] -Inf 0.487 0.738 0.576 -Inf ...
#> ..$ LL    : num [1, 1, 1:7, 1:2] -Inf -Inf -0.238 1.919 -Inf ...
#> ..$ LL_IL: logi NA
```

```
#> $ lesions      :List of 3
#>   ..$ perCase: num [1:7] 1 1 2 2 1 2 2
#>   ..$ IDs     : num [1:7, 1:2] 1 1 1 1 1 ...
#>   ..$ weights: num [1:7, 1:2] 1 1 0.5 0.5 1 ...
#> $ descriptions:List of 7
#>   ..$ fileName    : chr "NA"
#>   ..$ type        : chr "FROC"
#>   ..$ name        : logi NA
#>   ..$ truthTableStr: logi NA
#>   ..$ design       : chr "FCTRL"
#>   ..$ modalityID  : chr "1"
#>   ..$ readerID    : chr "1"
```

It is seen to consist of three list members: `ratings`, `lesions` and `descriptions`.

#### 26.3.2.3 The structure of the `ratings` member

The `ratings` member is itself a list of 3, consisting of `NL` the non-lesion localization ratings, `LL` the lesion localization ratings and `LL_IL` the incorrect localization ratings. The last member is needed for LROC datasets and can be ignored for now.

#### 26.3.2.4 The structure of the `NL` member

```
frocDataRaw$ratings$NL[1,1,,]
#>           [,1]      [,2] [,3]
#> [1,]      -Inf      -Inf -Inf
#> [2,]  0.48742905      -Inf -Inf
#> [3,]  0.73832471      -Inf -Inf
#> [4,]  0.57578135 -0.3053884 -Inf
#> [5,]      -Inf      -Inf -Inf
#> [6,]  1.51178117  0.3898432 -Inf
#> [7,]  1.12493092 -0.6212406 -Inf
#> [8,] -0.04493361      -Inf -Inf
#> [9,] -0.01619026      -Inf -Inf
#> [10,]      -Inf      -Inf -Inf
#> [11,]      -Inf      -Inf -Inf
#> [12,]      -Inf      -Inf -Inf
```

- It is seen to be an array with dimensions [1,1,1:12,1:4].

- Note that all listed ratings are greater than  $\zeta_1 = -1$ . Unmarked locations are assigned the  $-\infty$  rating.
- Case 1, the first non-diseased case, has a single NL mark rated  $-\infty$  and the remaining 3 locations are filled with  $-\infty$ .
- Case 6, the first diseased case, has zero NL marks and all 4 locations for it are filled with  $-\infty$ . [As seen below, this case actually generated a rating in the first location, but it fell below  $\zeta_1 = -1$ .]
- Case 11, the sixth diseased case, has three NL marks rated  $-\infty$ ,  $-\infty$ ,  $-\infty$  and the remaining location for it is  $-\infty$ . As noted below, this case generated a fourth rating that fell below  $\zeta_1 = -1$ .
- The first dimension corresponds to the number of modalities, one in this example, the second dimension corresponds to the number of readers, also one in this example.
- The third dimension is the total number of cases,  $K_1 + K_2 = 12$  in this example, because NLs are possible on *both* non-diseased and diseased cases.
- The fourth dimension is 4, as the simulator generates, over 12 cases, a maximum of 4 latent NLs per case. This can be demonstrated (see below) by running the preceding code where one temporarily sets  $\zeta_1 = -\infty$ , which results in all latent marks being marked: one sees that case 11, the sixth diseased case, actually generates 4 NLs, but one of them, at position 4, has rating equal to -1.237538, which is less than  $\zeta_1 = -1$ , and is consequently not marked in the original example, i.e., this location is assigned a rating of  $-\infty$ .

```
frocDataRaw1$ratings$NL[1,1,,]
#>      [,1]     [,2]     [,3]
#> [1,]    -Inf    -Inf    -Inf
#> [2,] 0.48742905    -Inf    -Inf
#> [3,] 0.73832471    -Inf    -Inf
#> [4,] 0.57578135 -0.3053884    -Inf
#> [5,]    -Inf    -Inf    -Inf
#> [6,] 1.51178117  0.3898432    -Inf
#> [7,] 1.12493092 -0.6212406 -2.2147
#> [8,] -0.04493361    -Inf    -Inf
#> [9,] -0.01619026    -Inf    -Inf
#> [10,]    -Inf    -Inf    -Inf
#> [11,]    -Inf    -Inf    -Inf
#> [12,]    -Inf    -Inf    -Inf
```

### 26.3.2.5 The structure of the LL member

```
frocDataRaw$ratings$LL[1,1,,]
#>      [,1]      [,2]
#> [1,] -Inf      -Inf
#> [2,] -Inf      -Inf
#> [3,] -0.2375384 -Inf
#> [4,]  1.9189774 -Inf
#> [5,] -Inf      -Inf
#> [6,]  1.0745650 -Inf
#> [7,]  1.5036080  0.9428932
```

- It is seen to be an array with dimensions  $[1, 1, 1:7, 1:2]$ .
- The first dimension corresponds to the number of modalities, one in this example, the second dimension corresponds to the number of readers, also one in this example.
- The third dimension is the total number of diseased cases,  $K_2 = 7$ , because LLs are only possible on diseased cases.
- The fourth dimension is 2, as the maximum number of lesions per diseased case is  $L_{\max} = 2$ .
- Note that all listed ratings are greater than  $\zeta_1 = -1$ .
- Case 1, the first diseased case, has zero LL marks and both locations are filled with  $-\infty$ .
- Case 2, the second diseased case, has one LL mark rated  $-\infty$  and the remaining location is  $-\infty$ .
- Case 7, the seventh diseased case, has two LL marks rated 1.503608, 0.9428932 and zero locations with  $-\infty$ .
- The following output shows that setting  $\zeta_1 = -\infty$  does not reveal any more latent LLs.

```
frocDataRaw1$ratings$LL[1,1,,]
#>      [,1]      [,2]
#> [1,] -Inf      -Inf
#> [2,] -Inf      -Inf
#> [3,] -0.2375384 -Inf
#> [4,]  1.9189774 -Inf
#> [5,] -Inf      -Inf
#> [6,]  1.0745650 -Inf
#> [7,]  1.5036080  0.9428932
```

- Lines 23 - 25 use the `PlotEmpiricalOperatingCharacteristics` function to calculate the FROC plot `ggplot` object, which is saved to `p1A`. Note the argument `opChType = "FROC"`, for the desired FROC plot.
- Lines 28 - 31 use the `PlotEmpiricalOperatingCharacteristics` function to calculate the AFROC plot object, which is saved to `p1B`. Note the argument `opChType = "AFROC"`.
- Finally, lines 33 - 35 use the `PlotEmpiricalOperatingCharacteristics` function to calculate the ROC plot object, which is saved to `p1C`. Note the argument `opChType = "ROC"`.

In summary, the code generates FROC, AFROC and ROC plots shown in the top row of Fig. 26.1, labeled A, B and C. The discreteness, i.e., the relatively big jumps between data points, is due to the small numbers of cases. Increasing the numbers of cases to  $K_1 = 50$  and  $K_2 = 70$  yields the lower row of plots in Fig. 26.1, labeled D, E and F. The fact that the upper row left plot does not seem to match the lower row left plot, especially near  $NLF = 0.25$ , is due to sampling variability with few cases.

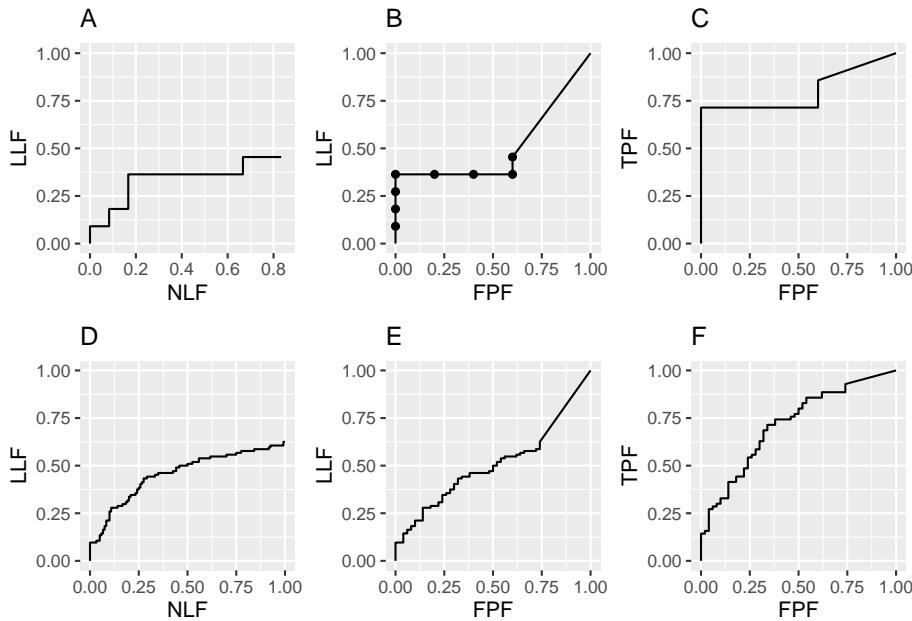


Figure 26.1: Raw FROC, AFROC and ROC plots with `seed = 1`: Plots A, B, C are for  $K_1 = 5$  and  $K_2 = 7$  cases while plots D, E, F are for  $K_1 = 50$  and  $K_2 = 70$  cases.

Fig. 26.1 Raw FROC, AFROC and ROC plots with `seed = 1`: Plots A, B and C are for  $K_1 = 5$  and  $K_2 = 7$  cases while D, E and F are for  $K_1 = 50$  and  $K_2 = 70$  cases. Model parameters are  $\mu = 1$ ,  $\lambda = 1$ ,  $\nu = 1$  and  $\zeta_1 = -1$ . The discreteness (jumps) in A, B and C is due to the small number of cases. The decreased discreteness in D, E and F is due to the larger numbers of cases. If the number of cases is increased further, the plots will approach continuous plots, like those shown in Chapter 24. Note that the AFROC (B and E) and ROC plots (C and F), are each contained within unit squares, unlike the semi-constrained FROC plots A and D.

#### 26.3.2.6 Effect of `seed` on raw plots

Shown next are similar plots but this time `seed = 2`.

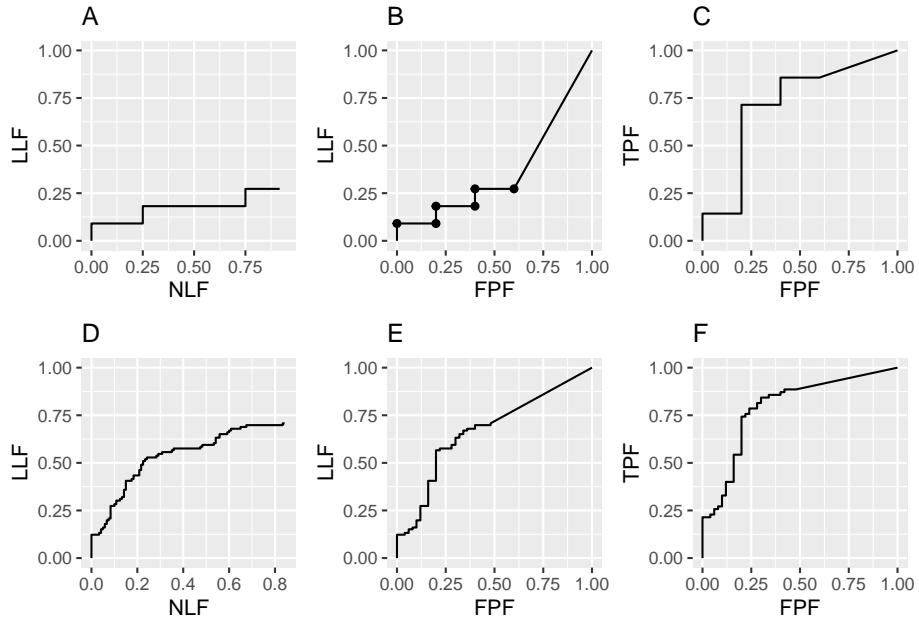


Figure 26.2: Raw FROC, AFROC and ROC plots with `seed = 2`: Plots A, B, C are for  $K_1 = 5$  and  $K_2 = 7$  cases while plots D, E, F are for  $K_1 = 50$  and  $K_2 = 70$  cases.

Fig. 26.2 Raw FROC, AFROC and ROC plots with `seed = 2`: Plots A, B, C are for  $K_1 = 5$  and  $K_2 = 7$  cases while plots D, E, F are for  $K_1 = 50$  and  $K_2 = 70$  cases. Model parameters are  $\mu = 1$ ,  $\lambda = 1$ ,  $\nu = 1$  and  $\zeta_1 = -1$ . Note the large variability in the upper row plots as compared to those in Fig. 26.1.

### 26.3.3 Key differences from the ROC paradigm:

- In a ROC study, each case generates exactly one rating.
- In a FROC study, each case can generate zero or more (0, 1, 2, ...) mark-rating pairs.
- The number of marks per case is a random variable as is the rating of each mark.
- Each mark corresponds to a distinct location on the image and associated with it is a rating, i.e., confidence level in presence of disease at the region indicated by the mark.
- In the ROC paradigm, each non-diseased case generates one FP and each diseased case generates one TP.
- In a FROC study, each non-diseased case can generate zero or more NLs and each diseased case can generate zero or more NLs and zero or more LLs.
- The number of lesions in the case limits the number of LLs.

## 26.4 The chance level FROC and AFROC

The chance level FROC was addressed in the previous chapter; it is a “flat-liner”, hugging the x-axis, except for a slight upturn at large NLF.

Fig. 26.3 shows “near guessing” FROC (plot A) and AFROC (plot B) plots for  $\mu = 0.1$ . These plots were generated by the code with  $\mu = 0.1$ ,  $\lambda = 1$ ,  $\nu = 0.1$ ,  $\zeta_1 = -1$ ,  $K_1 = 50$ ,  $K_2 = 70$ .

The AFROC of a guessing observer is not the line connecting (0,0) to (1,1). A guessing observer will also generate a “flat-liner”, but this time the plot ends at FPF = 1, and the straight line extension will be a vertical line connecting this point to (1,1). In the limit  $\mu \rightarrow 0+$ , AFROC-AUC tends to zero.

*To summarize, AFROC AUC of a guessing observer is zero.* On the other hand, suppose an expert radiologist views screening images and the lesions on diseased cases are very difficult, even for the expert, and the radiologist does not find any of them. Being an expert the radiologist successfully screens out non-diseased cases and sees nothing suspicious in any of them – this is a measure of the expertise of the radiologist, not mistaking variants of normal anatomy for false lesions on non-diseased cases. Accordingly, the expert radiologist does not report anything, and the operating point is “stuck” at the origin. Even in this unusual situation, one would be justified in connecting the origin to (1,1) and claiming area under AFROC is 0.5. The extension gives the radiologist credit for not marking any non-diseased case; of course, the radiologist does not get

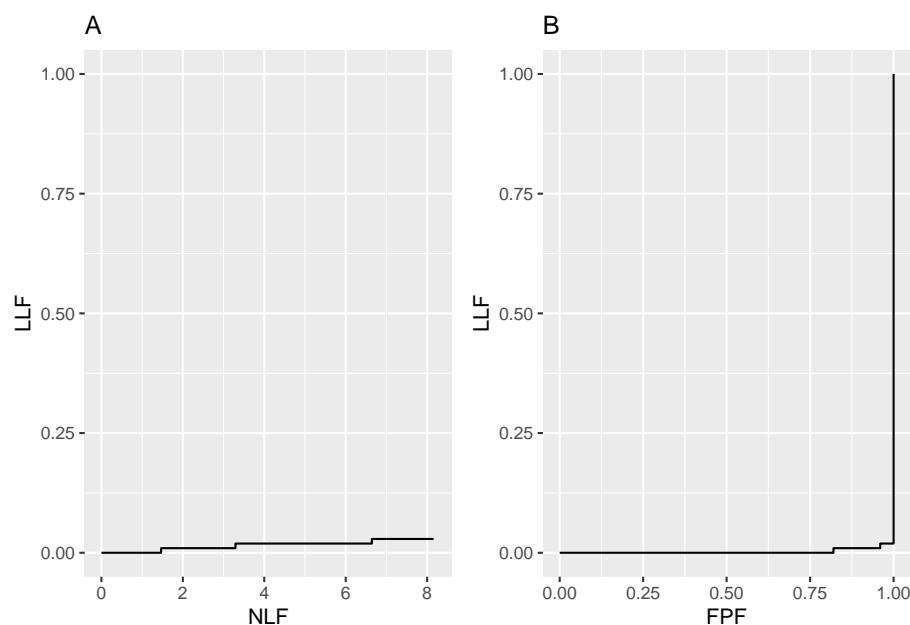


Figure 26.3: Plot A is the near guessing observer's FROC and plot B is the corresponding AFROC for  $\mu = 0.01$ .

any credit for marking any of the lesions. An even better radiologist, who finds and marks some of the lesions, will score higher, and AFROC-AUC will exceed 0.5. See TBA §17.7.4 for a software demonstration of this unusual situation.

## 26.5 Location-level “true-negatives”

The quotes are intended to draw attention to confusion that can result when one inappropriately applies ROC terminology to the FROC paradigm. For the 5 / 7 dataset, seed = 1, and reporting threshold set to -1, the first non-diseased case has one NL rated  $-\infty$ . The remaining three entries for this case are filled with  $-\infty$ .

What really happened is only known if one has access to the internals of the simulator. To the data analyst the following possibilities are indistinguishable:

- Four latent NLs, one of whose ratings exceeded  $\zeta_1$ , i.e., three location-level “true negatives” occurred on this case.
- Three latent NLs, one of whose ratings exceeded  $\zeta_1$ , i.e., two location-level “true negatives” occurred on this case.
- Two latent NLs, one of whose ratings exceeded  $\zeta_1$ , i.e., one location-level “true negative” occurred on this case.
- One latent NL, whose rating exceeded  $\zeta_1$ , i.e., 0 location-level “true negatives” occurred on this case.

The second non-diseased case has one NL mark rated 0.4874291 and similar ambiguities occur regarding the number of latent NLs. The third, fourth and fifth non-diseased cases have no marks. All four locations-holders on each of these cases are filled with  $-\infty$ , which indicates un-assigned values corresponding to either absence of any latent NL or presence of one or more latent NLs that did not exceed  $\zeta_1$  and therefore did not get marked.

To summarize: absence of an actual NL mark, indicated by a  $-\infty$  rating, could be due to either (i) non-occurrence of the corresponding latent NL or (ii) occurrence of the latent NL but its rating did not exceed  $\zeta_1$ . One cannot distinguish between the two possibilities, as in either scenario, the corresponding rating is assigned the  $-\infty$  value and either scenario would explain the absence of a mark.

For those who insist on using ROC terminology to describe FROC data the second possibility would be termed a location level True Negative (“TN”). Their “logic” is as follows: there was the possibility of a NL mark, which they term a “FP”, but the observer did not make it. Since the complement of a FP event is a TN event, this was a TN event. However, as just shown, one cannot tell if it was a “TN” event or there was no latent event in the first place. Here is the conclusion: there is no place in the FROC lexicon for a location level “TN”.

If  $\zeta_1 = -\infty$  then all latent marks are actually marked and the ambiguities mentioned above disappear. As noted previously, when this change is made one confirms that there were actually four latent NLs on the sixth diseased case (the eleventh sequential case), but the one rated  $-1.237538$  fell below  $\zeta_1 = -1$  and was consequently not marked.

So one might wonder, why not ask the radiologists to report everything they see, no matter how low the confidence level? Unfortunately, that would be contrary to their clinical task, where there is a price to pay for excessive NLs. It would also be contrary to a principle of good experimental design: one should keep interference with actual clinical practice, designed to make the data easier to analyze, to a minimum.

## 26.6 Binned FROC/AFROC/ROC plots

In the preceding example, continuous ratings data was available and data binning was not employed. Shown next is the code for generating the plots when the data is binned.

### 26.6.1 Code for binned plots

```

1  seed <- 1; set.seed(seed)
2  mu <- 1
3  zeta1 <- -1
4  K1 <- 5
5  K2 <- 7
6  Lmax <- 2
7  Lk2 <- floor(runif(K2, 1, Lmax + 1))
8
9  frocDataRaw <- SimulateFrocDataset(
10    mu = mu,
11    lambda = lambda,
12    nu = nu,
13    I = 1,
14    J = 1,
15    K1 = K1,
16    K2 = K2,
17    perCase = Lk2,
18    zeta1 = zeta1,
19    seed = seed
20  )
21
22  frocDataBinned <- DfBinDataset(

```

```

23   frocDataRaw,
24   desiredNumBins = 5,
25   opChType = "FROC")
26
27 p4A <- PlotEmpiricalOperatingCharacteristics(
28   dataset = frocDataBinned,
29   trts= 1, rdrs = 1, opChType = "FROC",
30   legend.position = "NULL")$Plot + ggtitle("A")
31
32 p4B <- PlotEmpiricalOperatingCharacteristics(
33   dataset = frocDataBinned,
34   trts= 1, rdrs = 1, opChType = "AFROC",
35   legend.position = "NULL")$Plot + ggtitle("B")
36
37 p4C <- PlotEmpiricalOperatingCharacteristics(
38   dataset = frocDataBinned,
39   trts= 1, rdrs = 1, opChType = "ROC",
40   legend.position = "NULL")$Plot + ggtitle("C")

```

This is similar to the code for the raw plots except that at lines 21-24 we have used the function `DfBinDataset` to bin the raw data `frocDataRaw` and the binned data is saved to `frocDataBinned`, which is used in the subsequent plotting routines. Note the arguments `desiredNumBins` and `opChType`. The binning function needs to know the desired number of bins (set to 5 in this example) and the operating characteristic that the binning is aimed at (here set to “FROC”).

### 26.6.2 Effect of `seed` on binned plots

Shown next are corresponding plots with `seed = 2`.

## 26.7 Structure of the binned data

```

str(frocDataBinnedSeed1$ratings$NL)
#> num [1, 1, 1:120, 1:4] -Inf 4 2 3 -Inf ...
table(frocDataBinnedSeed1$ratings$NL)
#>
#> -Inf     1     2     3     4
#> 376    35    30    23    16
sum(as.numeric(table(frocDataBinnedSeed1$ratings$NL)))
#> [1] 480

```

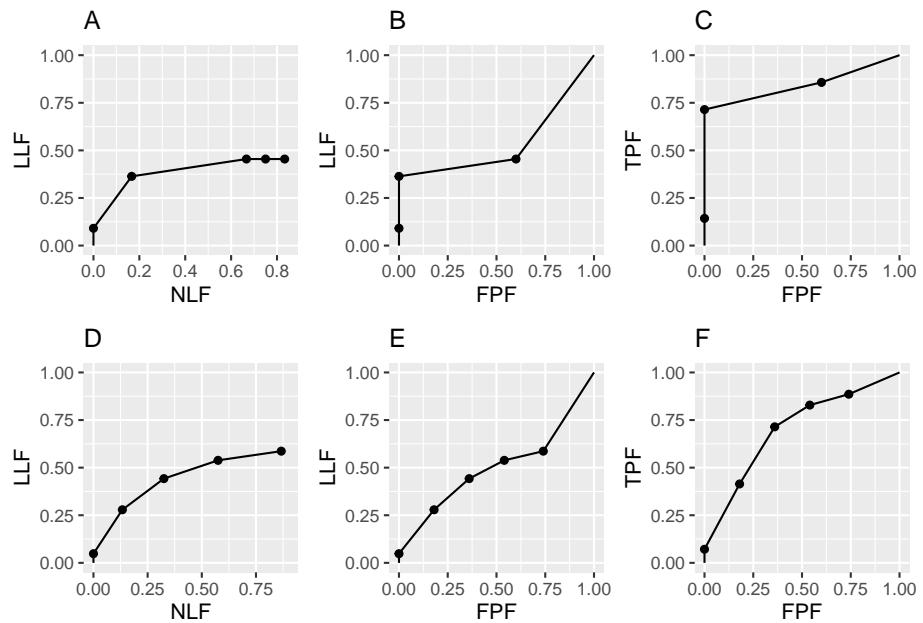


Figure 26.4: Binned FROC, AFROC and ROC plots with seed = 1: Plots A, B, C are for  $K_1 = 5$  and  $K_2 = 7$  cases while plots D, E, F are for  $K_1 = 50$  and  $K_2 = 70$  cases

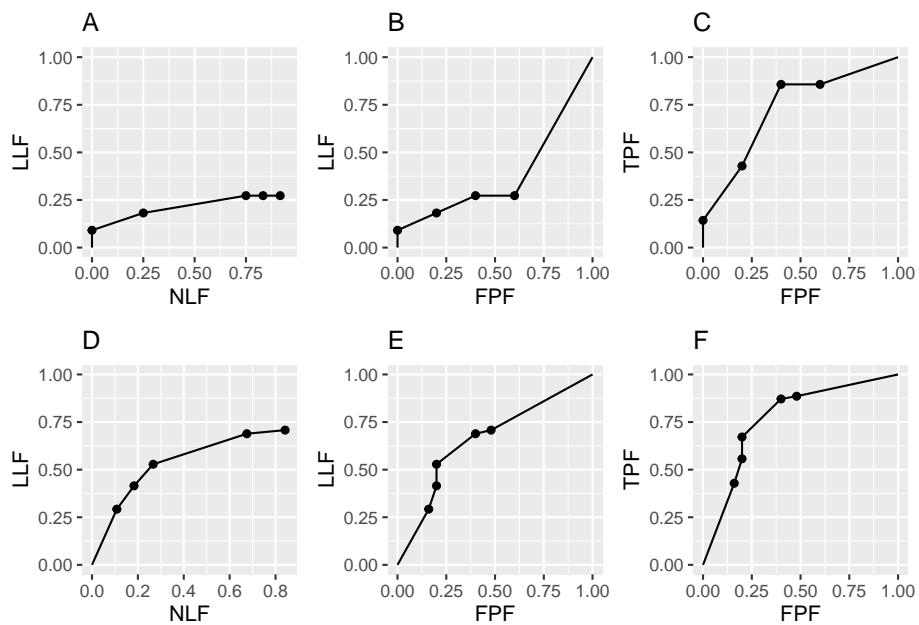


Figure 26.5: Binned FROC, AFROC and ROC plots with seed = 2: Plots A, B, C are for  $K_1 = 5$  and  $K_2 = 7$  cases while plots D, E, F are for  $K_1 = 50$  and  $K_2 = 70$  cases

- The `table()` function converts an array into a counts table.
- There are  $120 \times 4 = 480$  elements in the `NL` array to be “tabled”.
- From the output one sees that there are 378 entries in the `NL` array that equal  $-\infty$ , 50 that equal 1, 15 that equal 2, 12 that equal 3, and 25 that equal 4 (none of the NLs were binned into the rating “5” category). These sum to 480 (see code output above).
- Because the fourth dimension of the `NL` array is determined by cases with the *most* NLs, on the *unknown number* (to the data analyst) of cases with *fewer* NLs, this dimension is “padded” with negative-infinities.
- Because of the unknown number of negative-infinity paddings, one does not know how many of the 378 *observed* negative-infinities are *actually* latent NLs. The *actual* number of latent NLs could be considerably smaller - and the number of *marked* NLs even smaller - as this is determined by those latent NLs whose z-samples  $\geq \zeta_1$ . Notice that in the special case  $\zeta_1 = -\infty$  the observer marks all latent NL, in which case the observed count equal the actual count.

```
str(frocDataBinnedSeed1$ratings$LL)
#> num [1, 1, 1:70, 1:2] 3 4 4 4 3 ...
table(frocDataBinnedSeed1$ratings$LL)
#>
#> -Inf     1     2     3     4     5
#> 79      5    10    17    24     5
sum(as.numeric(table(frocDataBinnedSeed1$ratings$LL)))
#> [1] 140
sum(Lk2Seed1)
#> [1] 104
sum(Lk2Seed1) - sum(as.vector(table(frocDataBinnedSeed1$ratings$LL)))[2:6])
#> [1] 43
```

- The `LL` array contains  $70 \times 2 = 140$  values to be “tabled”.
- From the output one sees that there are 78 entries in the `LL` array that equal  $-\infty$ , 10 entries that equal 1, 5 entries that equal 2, 8 entries that equal 3, 35 entries that equal 4, and 4 entries that equal 5. These sum to 140, the product of the lengths of the third and fourth dimensions of the `LL` array.
- The number of negative-infinity counts is 78. This is smaller than 140 because, of the varying numbers of lesions, some of the location-holders are filled with negative infinities.
- The *known* total number of lesions – each of which contributes a latent `LL` – is 104, see 2nd last line of above code output.
- Summing the `LL` counts in bins 1 through 5 (corresponding to table

columns 2-6, since column 1 applies to the negative-infinities) and subtracting from the total number of lesions one gets:  $104 - (10+5+8+35+4) = 104 - 62 = 42$ , see last line of above code output.

- Therefore, the number of unmarked lesions is 42. The listed value (78) is an overestimate because it includes the  $-\infty$  counts from the fourth dimension negative-infinity “padding” of the LL array.

## 26.8 Summary

The preceding detailed example illustrates a key point: *The total number of latent NLs in the dataset is generally unknown to the data analyst, unlike the total number of latent LLs, which is known.* The only exception to this rule is if  $\zeta_1 = -\infty$ , in which case the observer marks all latent NL (and LL) sites.

## 26.9 Discussion

TBA

## 26.10 References



# Chapter 27

## FROC vs. wAFROC

### 27.1 TBA How much finished

50% Need to replace simulation values with analytical values

### 27.2 Introduction

In the medical imaging context the FROC curve, which was introduced in (Bunch et al., 1977b), has been widely used for evaluating performance in the free-response paradigm, particularly in CAD algorithm development. Typically CAD researchers report sensitivity at a stated value of false positives per image, i.e., they report a *pair* of values. (TBA) From basic ROC analysis, see Section 10.13, we know that a scalar FOM is preferable to reporting a pair of values. This chapter recommends adoption of the area under the wAFROC as the preferred scalar figure of merit in lieu of sensitivity / false positives per image pairs. operating characteristic in assessing performance in the free-response paradigm, and details simulation-based studies supporting this recommendation.

### 27.3 FROC vs. wAFROC

Recall, from Section 24.7, that the RSM is defined by parameters  $\mu, \lambda, \nu$  and  $\zeta_1$ . This section examines RSM-predicted TBA analytical FROC, wAFROC and ROC panels for two observers denoted R1 and R2. The former could be an algorithmic observer while the latter could be a radiologist. For typical threshold  $\zeta_1$  parameters, three types of situations are considered: R2 has moderately better performance than R1, R2 has much better performance than R1 and R2 has slightly better performance than R1. For each type of simulation pairs of

FROC, wAFROC and ROC curves are shown, one for each observer. Finally the simulations and panels are repeated for hypothetical R1 and R2 observers who report all suspicious regions, i.e.,  $\zeta_1 = -\infty$  for each observer. Both R1 and R2 observers share the same  $\lambda, \nu$  parameters, and the only difference between them is in the  $\mu$  and  $\zeta_1$  parameters.

### 27.3.1 Moderate difference in performance

```

1 source(here("R/CH13-CadVsRadPlots/CadVsRadPlots.R"))
2
3 nu <- 1
4 lambda <- 1
5 K1 <- 500
6 K2 <- 700
7 mu1 <- 1.0
8 mu2 <- 1.5
9 zeta1_1 <- -1
10 zeta1_2 <- 1.5
11 Lmax <- 2
12 seed <- 1
13
14 ret <- do_one_figure (
15   seed, Lmax, mu1,
16   mu2, lambda, nu, zeta1_1, zeta1_2, K1, K2)
17
18 froc_plot_1A <- ret$froc_plot_A
19 wafroc_plot_1B <- ret$wafroc_plot_B
20 roc_plot_1C <- ret$roc_plot_C
21 froc_plot_1D <- ret$froc_plot_D
22 wafroc_plot_1E <- ret$wafroc_plot_E
23 roc_plot_1F <- ret$roc_plot_F
24 wafroc_1_1B <- ret$wafroc_1_B
25 wafroc_2_1B <- ret$wafroc_2_B
26 roc_1_1C <- ret$roc_1_C
27 roc_2_1C <- ret$roc_2_C
28 wafroc_1_1E <- ret$wafroc_1_E
29 wafroc_2_1E <- ret$wafroc_2_E
30 roc_1_1F <- ret$roc_1_F
31 roc_2_1F <- ret$roc_2_F

```

The  $\lambda$  and  $\nu$  parameters are defined at lines 3 and 4 of the preceding code:  $\lambda = \nu = 1$ . The number of simulated cases is defined, lines 5-6, by  $K_1 = 500$  and  $K_2 = 700$ . The simulated R1 observer  $\mu$  parameter is defined at line 7 by  $\mu_1 = 1$  and that of the simulated R2 observer is defined at line 8 by  $\mu_2 = 1.5$ .

Based on these choices one expect R2 to be moderately better than R1. The corresponding threshold parameters are (lines 9 -10)  $\zeta_1 = -1$  for R1 and  $\zeta_1 = 1.5$  for R2. The maximum number of lesions per case is defined at line 11 by `Lmax = 2`. The actual number of lesions per case is determined determined by random sampling within the helper function `do_one_figure()` called at lines 14-16. This function returns a large list `ret`, whose contents are as follows:

- `ret$froc_plot_A`: a pair of FROC panels for the thresholds specified above, a red panel labeled “R: 1” corresponding to R1 and a blue panel labeled “R: 2” corresponding to R2. These are shown in panel A.
- `ret$wafroc_plot_B`: a pair of wAFROC panels, similarly labeled. These are shown in panel B.
- `ret$roc_plot_C`: a pair of ROC panels, similarly labeled. These are shown in panel C.
- `ret$froc_plot_D`: a pair of FROC panels for the both thresholds at  $-\infty$ . These are shown in panel D.
- `ret$froc_plot_E`: a pair of wAFROC panels for the both thresholds at  $-\infty$ . These are shown in panel E.
- `ret$froc_plot_F`: a pair of ROC panels for the both thresholds at  $-\infty$ . These are shown in panel F.
- `ret$wafroc_1_B`: the wAFROC AUC for R1, i.e., the area under the curve labeled “R: 1” in panel B.
- `ret$wafroc_2_B`: the wAFROC AUC for R2, i.e., the area under the curve labeled “R: 2” in panel B.
- `ret$roc_1_C`: the ROC AUC for R1, i.e., the area under the curve labeled “R: 1” in panel C.
- `ret$roc_2_C`: the ROC AUC for R2, i.e., the area under the curve labeled “R: 2” in panel C.
- `ret$wafroc_1_E`: the wAFROC AUC for R1, i.e., the area under the curve labeled “R: 1” in panel E.
- `ret$wafroc_2_E`: the wAFROC AUC for R2, i.e., the area under the curve labeled “R: 2” in panel E.
- `ret$roc_1_F`: the ROC AUC for R1, i.e., the area under the curve labeled “R: 1” in panel F.
- `ret$roc_2_F`: the ROC AUC for R2, i.e., the area under the curve labeled “R: 2” in panel F.

The coordinates of the end-point of the R1 FROC in panel A are (0.826, 0.590). Those of the R2 FROC curve in A are (0.049, 0.398). The FROC for the R1 observer extends to much larger NLF values while that for the R2 observer is relatively short and steep. One suspects the R2 observer is performing better than R1: he is better at finding lesions and producing fewer NLs, both of which are desirable characteristics, but he is adopting a too-strict reporting criterion. If he could be induced to relax the threshold and report more NLs, his LLF would exceed that of the R1 observer while still maintaining a lower

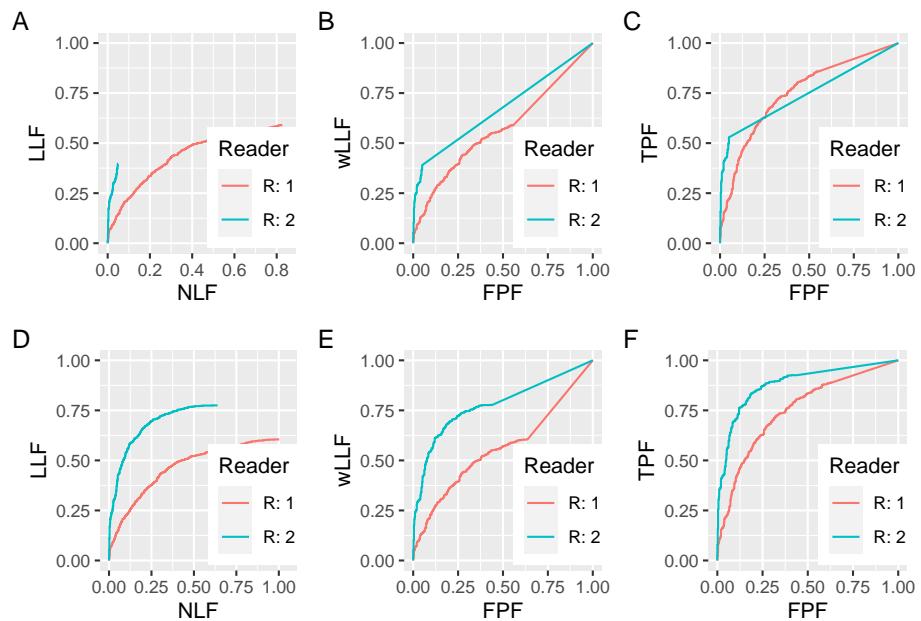


Figure 27.1: Plots A and D: FROC curves for the R1 and R2 observers; B and E are corresponding wAFROC curves and C and F are corresponding ROC curves. All curves in this plot are for  $\lambda = \nu = 1$ . All RAD\_1 curves are for  $\mu = 1$  and all RAD\_2 curves are for  $\mu = 1.5$ . For panels A, B and C,  $\zeta_1 = -1$  for R1 and  $\zeta_1 = 1.5$  for R2. For panels D, E and F,  $\zeta_1 = -\infty$  for R1 and R2.

NLF. However, as this involves a subjective extrapolation, it is not possible to objectively quantify this from the FROC curves. The basic issue is the lack of a common NLF range for the two panels. If a common NLF range is “forced”, for example defined as the common NLF range 0 to 0.0492, where both curves contribute, it would ignore most NLs from the R1 observer.

Algorithm developers typically quote LLF at a specified NLF. According to the two panels in A, the R2 observer is better if the NLF value is chosen to less than 0.0492 - this is the maximum NLF value for the R2 curve in A - but there is no basis for comparison for larger values of NLF (because the R2 observer does not provide any data beyond the observed end-point). A similar problem was encountered in ROC analysis when comparing a pair of sensitivity-specificity values, where, given differing choices of thresholds, ambiguous results can be obtained, see Section 10.13. Indeed, this was the rationale for using AUC under the ROC curve as an unambiguous measure of performance.

Plot B shows wAFROC curves for the same datasets whose FROC curves are shown in panel A. **The wAFROC is contained within the unit square, a highly desirable characteristic, which solves the lack of a common NLF range problem with the FROC.** The wAFROC AUC under the R2 observer is visibly greater than that for the R1 observer, even though – due to his higher threshold – his AUC estimate is actually biased downward (because the R2 observer is adopting a high threshold, his  $LLF_{max}$  is smaller than it would have been with a lower threshold, and consequently the area under the large straight line segment from the uppermost non-trivial operating point to (1,1) is smaller). AUCs under the two wAFROC panels in B are 0.5731 for R1 and 0.6737 for R2.

Plot C shows ROC curves. Since the curves cross, it is not clear which has the larger AUC. AUCs under the two curves in C are 0.7499 for R1 and 0.7453 for R2, which are close, but here is an example where the ordering given by the wAFROC is opposite to that given by the ROC.

Plots D, E and F correspond to A, B and C with this important difference: the two threshold parameters are set to  $-\infty$ . The coordinates of the end-point of the R1 FROC in panel D are (1.002, 0.605). Those of the R2 FROC in panel D are (0.639, 0.775). The R2 observer has higher LLF at lower NLF, and there can be no doubt that he is better. Panels E and F confirm that R2 is actually the better observer *over the entire FPF range*. AUCs under the two wAFROC curves in E are 0.5605 for R1 and 0.7780 for R2. AUCs under the two ROC curves in F are 0.7513 for R1 and 0.8826 for R2. These confirm the visual impressions of panels in panels E and F. Notice that each ROC AUC is larger than the corresponding wAFROC AUC. This is because the probability of a lesion localization (case is declared positive *and* a lesion is correctly localized) is smaller than the probability of a true positive (case is declared positive). In other words, the ROC is everywhere above the wAFROC.

### 27.3.2 Large difference in performance

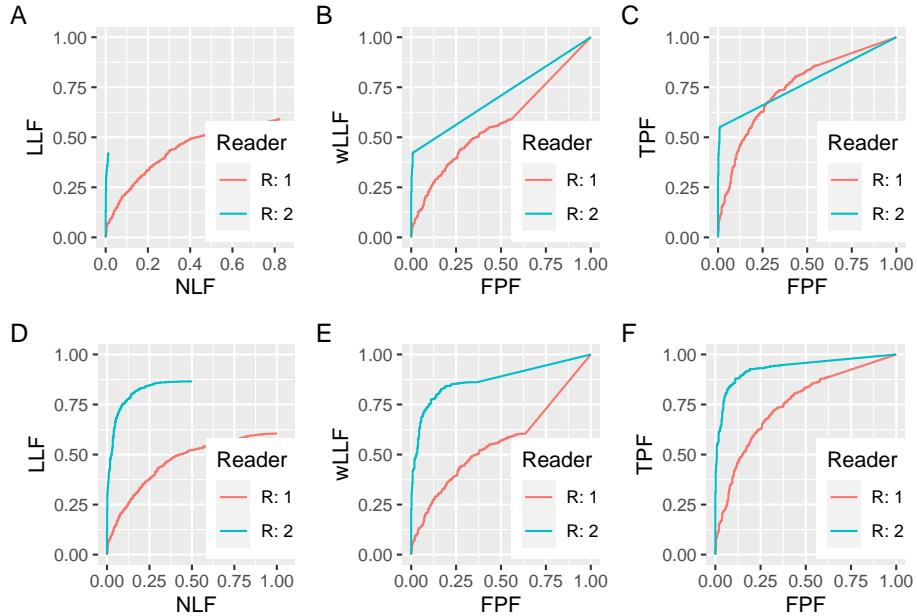


Figure 27.2: Similar to preceding figure but with the following changes. All RAD\_2 curves are for  $\mu = 2$  and for panels A, B and C  $\zeta_1 = 2$  for R2.

In Fig. 27.2 panel A, the R1 parameters are the same as in Fig. 27.1, but the R2 parameters are  $\mu_2 = 2$  and  $\zeta_1 = +2$ . Doubling the separation parameter over that of R1 ( $\mu_1 = 1$ ) has a huge effect on performance. The end-point coordinates of the FROC for R1 are (0.826, 0.590). The end-point coordinates of the FROC for R2 are (0.015, 0.421). The common NLF region defined by  $NLF = 0$  to  $NLF = 0.0150$  would exclude almost all of the marks made by R1. The wAFROC panels in panel B show the markedly greater performance of R2 over R1 (the AUCs are 0.5731 for R1 and 0.7075 for R2). The inter-reader difference is larger (compared to Fig. 27.1 panel B), despite the greater downward bias working against the R2 observer. Panel C shows ROC panels for the two observers. Although the curves cross, it is evident that R2 has the greater AUC. The AUCs are 0.7499 for R1 and 0.7722 for R2.

Plots D, E and F correspond to A, B and C with the difference that the two threshold parameters are set to  $-\infty$ . The coordinates of the end-point of the R1 FROC in panel D are OpPtStr(nlf\_1\_2D, llf\_1\_2D). Those of the R2 FROC in panel D are OpPtStr(nlf\_2\_2D, llf\_2\_2D). The R2 observer has higher LLF at lower NLF, and there can be no doubt that he is better. Panels E and F confirm that R2 is actually the better observer over the entire FPF

range. AUCs under the two wAFROC curves in E are 0.5605 for R1 and 0.8720 for R2. AUCs under the two ROC curves in F are 0.7513 for R1 and 0.9343 for R2. These confirm the visual impressions of panels in panels E and F. Notice that each ROC AUC is larger than the corresponding wAFROC AUC.

### 27.3.3 Small difference in performance and identical thresholds

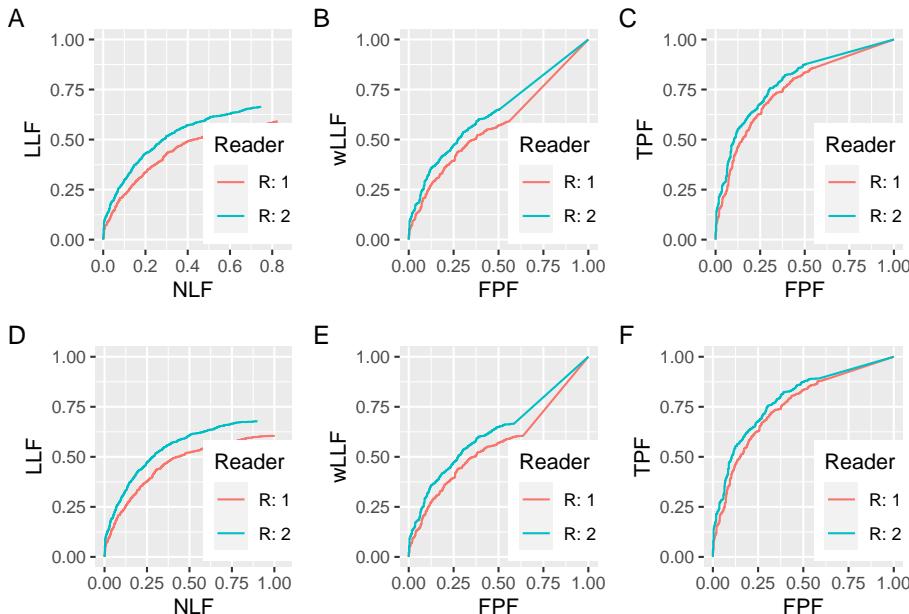


Figure 27.3: Similar to preceding figure but with the following changes. All RAD\_2 curves are for  $\mu = 1.1$  and for panels A, B and C,  $\zeta_1 = -1$  for R2.

The final example, Fig. 27.3 shows that *when there is a small difference in performance*, there is less ambiguity in using the FROC as a basis for measuring performance. The R1 parameters are the same as in Fig. 27.1 but the R2 parameters are  $\mu = 1.1$  and  $\zeta_1 = -1$ . In other words, the  $\mu$  parameter is 10% larger and the thresholds are identical. This time there is much more common NLF range overlap in panel A and one is counting most of the marks for the R1 reader. The end-point coordinates of the FROC for R1 are (0.826, 0.590). The end-point coordinates of the FROC for R2 are ((0.746, 0.664). The common NLF region defined by  $NLF = 0$  to  $NLF = 0.7458$  includes almost all of the marks made by R1. The wAFROC panels in panel B show the slight greater performance of R2 over R1 (the AUCs are 0.5731 for R1 and 0.6341 for R2).

Table 27.1: Summary of R1 simulations: A refers to panel A, B refers to panel B, etc.

wAFROC-B	wAFROC-E	ROC-C	ROC-F
0.5731	0.5605	0.7499	0.7513

Panel C shows ROC panels for the two observers. Although the curves cross, it is evident that R2 has the greater AUC. The AUCs are 0.7499 for R1 and 0.7722 for R2.

Plots D, E and F correspond to A, B and C with the difference that the two threshold parameters are set to  $-\infty$ . The coordinates of the end-point of the R1 FROC in panel D are ((1.002, 0.605). Those of the R2 FROC in panel D are ((0.901, 0.678). Panels E and F confirm that R2 is actually the better observer over the entire FPF range. AUCs under the two wAFROC curves in E are 0.5605 for R1 and 0.6238 for R2. AUCs under the two ROC curves in F are 0.7513 for R1 and 0.7857 for R2. These confirm the visual impressions of panels in panels E and F. Notice that each ROC AUC is larger than the corresponding wAFROC AUC.

## 27.4 Summary of simulations

The following tables summarize the numerical values from the plots in this chapter. Table 27.1 refers to the R1 observer, and Table 27.2 refers to the R2 observer.

### 27.4.1 Summary of R1 simulations

- The first column is labeled “wAFROC-B”, meaning the R1 wAFROC AUC in panel B, which are identical for the three figures (one may visually confirm that the red curves in panels A, B ad C in the three figures are identical; likewise for the red curves in panels D, E and F).
- The second column is labeled “wAFROC-E”, meaning the R1 wAFROC AUC in panel E, which are identical for the three figures.
- The third column is labeled “ROC-C”, meaning the R1 ROC AUC in panel C, which are identical for the three figures.
- The fourth column is labeled “ROC-F”, meaning the R1 ROC AUC in panel F, which are identical for the three figures.

Table 27.2: Summary of R2 simulations: Fig refers to the figure number in this chapter, A refers to panel A, B refers to panel B, etc.

Fig	wAFROC-B	wAFROC-E	ROC-C	ROC-F
1	0.6737	0.778	0.7453	0.8826
2	0.7075	0.872	0.7722	0.9343
3	0.6341	0.6238	0.7868	0.7857

### 27.4.2 Summary of R2 simulations

- The first column refers to the figure number, for example, “1” refers to Fig. 27.1, “2” refers to Fig. 27.2, and “3” refers to Fig. 27.3.
- The second column is labeled “wAFROC-B”, meaning the R2 wAFROC AUC corresponding to the blue curve in panel B.
- The third column is labeled “wAFROC-E”, meaning the R2 wAFROC AUC corresponding to the blue curve in panel E.
- The fourth column is labeled “ROC-C”, meaning the R2 ROC AUC corresponding to the blue curve in panel C.
- The fifth column is labeled “ROC-F”, meaning the R2 ROC AUC corresponding to the blue curve in panel F.

### 27.4.3 Comments

- For the same figure label the R1 panels are identical in the three figures. This is the reason why Table 27.1 has only one row. A *fixed* R1 dataset is being compared to *varying* R2 datasets.
- The first R2 dataset, Fig. 27.1 A, B or C, might be considered representative of an average radiologist, the second one, Fig. 27.2 A, B or C, is a super-expert and the third one, Fig. 27.3 A, B or C, is only nominally better than R1.
- Plots D, E and F are for hypothetical R1 and R2 observers that report *all* suspicious regions. The differences between A and D are minimal for the R1 observer, but marked for the R2 observer. Likewise for the differences between B and E.

## 27.5 Effect size comparison

- The effect size is defined as the AUC – calculated using either wAFROC or ROC – difference between RDR-2 and RDR-1 for the same figure. For example, for Fig. 27.2 and the wAFROC AUC effect size, one takes the difference between the AUCs under the R2 (blue) minus R1 (red) curves in panel B.

Table 27.3: Effect size comparisons for R1 simulations: Fig refers to the figure number in this chapter.

Fig	ES-wAFROC	ES-ROC
1	0.1006	-0.004654
2	0.1344	0.02222
3	0.061	0.03685

- In all three figures the wAFROC effect size (ES) is larger than the corresponding ROC effect size.
- For Fig. 27.1 panels B and C:
  - The wAFROC effect size is 0.1006,
  - The ROC effect size is -0.0047.
- For Fig. 27.2 panels B and C:
  - The wAFROC effect size is 0.1344,
  - The ROC effect size is 0.0222.
- For Fig. 27.3 panels B and C:
  - The wAFROC effect size is 0.0610,
  - The ROC effect size is 0.0369.

These results are summarized in Table 27.3.

Since effect size enters as the *square* in sample size formulas, wAFROC yields greater statistical power than ROC. The “small difference” example, corresponding to row number 2, is more typical of modality comparison studies where the modalities being compared are only slightly different. In this case the wAFROC effect size is about twice the corresponding ROC value - see chapter on FROC sample size TBA.

## 27.6 Performance depends on $\zeta_1$

Consider the wAFROC AUCs for the R2 curves in Fig. 27.2 panels B and E. The wAFROC AUC for R2 in panel B is 0.7075 while that for R2 in panel E is 0.8720. The only difference between the simulation parameters for the two curves are  $\zeta_1 = 2$  for panel B and  $\zeta_1 = -\infty$  for panel E. Clearly wAFROC AUC depends on the value of  $\zeta_1$ .

A similar result applies when considering the ROC curves in Fig. 27.2 panels C and F. The ROC AUC for R2 in panel C is 0.7722 while that for R2 in panel F is 0.9343. Clearly ROC AUC also depends on the value of  $\zeta_1$ .

The reason is that in panels B and C the respective AUCs are depressed due to high value of threshold parameter. The (very good) radiologist is seriously under-reporting and choosing to operate near the origin of a steep wAFROC/ROC curve. It is as if in an ROC study the reader is giving too much importance to specificity and therefore not achieving higher sensitivity.

*Since performance depends on threshold, this opens up the possibility of optimizing performance by finding the threshold that maximizes AUC. This is the subject of the next chapter.*

## 27.7 Discussion

## 27.8 References



# Chapter 28

## Meanings of FROC figures of merit

### 28.1 TBA How much finished

50%

### 28.2 Introduction

Chapter 25 focused on empirical plots possible with FROC data, for example, the FROC, AFROC, wAFROC and inferred ROC plots. Expressions were given for computing *operating points* for each plot from z-samples. Because of the ambiguity in ordering the two values associated with each operating points (e.g., sensitivity-specificity pairs in ROC plots), operating points should not be used as figures of merit. Rather one should use *area measures* derived from operating characteristics. This chapter is devoted to a number of such measures for FROC data.

A generic empirical area under a plot is denoted  $A_{oc}$ , where the “oc” subscript denotes the applicable operating characteristic. For example, the area under the empirical wAFROC is denoted  $A_{wAFROC}$ . Calculating areas from operating points using planimetry or geometry is tedious at best. *Needed are formulas for calculating them directly from ratings.* In this sense this chapter is analogous to Chapter 12 where it was shown that the area under the empirical ROC plot  $A_{ROC}$  equaled the Wilcoxon statistic calculated directly from the ratings, i.e., the Bamber theorem (Bamber, 1975).

I make a distinction between *empirical AUC under a plot*, i.e., an area measure, and a *FOM-statistic*, generically denoted  $\theta$ , that can be computed directly from

the ratings. While any function of the ratings is a possible FOM-statistic, whether it is useful depends upon whether it can be related to the area under an operating characteristic. This chapter derives formulas for FOM-statistics  $\theta_{oc}$ , which yield the same values as the areas  $A_{oc}$  under the corresponding empirical operating characteristics. The meanings of these FOM-statistics are discussed (Chakraborty and Zhai, 2016).

Here is the organization of the chapter.

- Expressions for the empirical AFROC FOM-statistic  $\theta_{AFROC}$  and the empirical weighted-AFROC FOM-statistic  $\theta_{wAFROC}$  are presented and their limiting values for chance-level and perfect performances are explored.
- Two important theorems are stated, whose proofs are in [TBA Online Appendix 14.A].
- The first theorem proves the equality between the empirical wAFROC FOM-statistic  $\theta_{wAFROC}$  and the area  $A_{wAFROC}$  under the empirical wAFROC plot. [A similar equality applies to the empirical AFROC FOM-statistic  $\theta_{AFROC}$  and the area  $A_{AFROC}$  under the empirical AFROC plot.]
- The second theorem derives an expression for the area under the straight-line extension of the wAFROC from the observed end-point to (1,1), and explains why it is essential to include this area.
- A small simulated-dataset is used to illustrate how NL and LL ratings and lesion weights determine the wAFROC empirical plot.
- It demonstrates that the wAFROC gives equal importance to all diseased cases, a desirable statistical characteristic.
- Corresponding results, but ignoring the weights, show that the AFROC gives excessive importance to cases with more lesions.
- A physical interpretation of the AUC or FOM-statistics is given. It shows explicitly how the ratings comparisons implied in FOM-statistic properly credit and penalize the observer for correct and incorrect decisions, respectively. The probabilistic meanings of the AFROC and wAFROC AUCs are given.
- Detailed derivations of FOM-statistics, applicable to the areas under the empirical FROC plot, the AFROC1 and wAFROC1 plots are not given. Instead, the results for all plots are summarized in [TBA Online Appendix 14.C], which shows that the definitions “work”, i.e., the FOM-statistics yield the correct areas as determined by numerical integration of the relevant curves.

## 28.3 Empirical AFROC FOM-statistic

$A_{\text{AFROC}}$  was defined in 25.8 as the area under the empirical AFROC. The corresponding FOM-statistic  $\theta_{\text{AFROC}}$  is defined as follows: one calculates the rating of the highest rated NL mark  $\text{FP}_{k_1 1}$  on each non-diseased case  $k_1 1$  (or  $-\infty$  if the case has no NL marks) and compares it to each LL rating using the kernel function  $\psi(x, y)$  defined in Eqn. (12.7)<sup>1</sup>. A summation is performed over all cases and all lesions.

The highest rating  $\text{FP}_{k_1 1}$  on non-diseased case  $k_1 1$  is defined as:

$$\left. \begin{aligned} \text{FP}_{k_1 1} &= \max_{l_1} (z_{k_1 1 l_1 1} \mid l_1 \neq \emptyset) \\ \text{FP}_{k_1 1} &= -\infty \mid l_1 = \emptyset \end{aligned} \right\} \quad (28.1)$$

If the case has at least one latent NL mark, then  $l_1 \neq \emptyset$ , where  $\emptyset$  is the null set, and the first definition applies. If the case has no marks, then  $l_1 = \emptyset$ , and the second definition applies.

The following equation sums over all cases and lesions:

$$\theta_{\text{AFROC}} = \frac{1}{K_1 L_T} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} \psi(\text{FP}_{k_1 1}, z_{k_2 2 l_2 2}) \quad (28.2)$$

Since every lesion is assigned a rating, albeit negative infinity for an unmarked lesion, the null set conditioning is not needed.

### 28.3.1 Upper limit for AFROC FOM-statistic

The FOM-statistic  $\theta_{\text{AFROC}}$  achieves its highest value, unity, if and only if every lesion is rated higher than any mark on non-diseased cases, for then the  $\psi$  function always yields unity, and the summations yield :

---

<sup>1</sup>The kernel function comparison yields 1 if the LL rating is higher, 0.5 if the ratings are identical and zero otherwise.

$$\begin{aligned}
 \theta_{\text{AFROC}} &= \frac{1}{K_1 \sum_{k_2=1}^{K_2} L_{k_2}} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} 1 \\
 &= \frac{1}{K_1 \sum_{k_2=1}^{K_2} L_{k_2}} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} L_{k_2} \\
 &= \frac{1}{K_1} \sum_{k_1=1}^{K_1} 1 \\
 &= 1
 \end{aligned} \tag{28.3}$$

### 28.3.2 Range of AFROC FOM-statistic

If, on the other hand, every lesion is rated lower than every mark on every non-diseased case, the  $\psi$  function always yields zero, and the FOM-statistic is zero. Therefore,

$$0 \leq \theta_{\text{AFROC}} \leq 1 \tag{28.4}$$

Eqn. (28.4) shows that  $\theta_{\text{AFROC}}$  behaves like a probability but its range is *twice* that of  $\theta_{\text{ROC}}$ ; recall that  $0.5 \leq \theta_{\text{ROC}} \leq 1$  (assuming the observer has equal or better than random performance and the observer does not have the direction of the rating scale accidentally reversed). This has the consequence that treatment related differences between  $\theta_{\text{AFROC}}$  (i.e., effect sizes) are larger relative to the corresponding ROC effect sizes (just as temperature differences in the Fahrenheit scale are larger than the same differences expressed in the Celsius scale). This has important implications for FROC sample size estimation, Chapter TBA.

Eqn. (28.4) is one reason why the “chance diagonal” of the AFROC, corresponding to  $AUC = 0.5$ , does not, in fact, reflect chance-level performance. An area under the AFROC equal to 0.5 is actually reasonable performance, being smack in the middle of the allowed range. An example of this was given in TBA §13.4.2.2 for the case of an expert radiologist who does not mark any cases.

## 28.4 Empirical weighted-AFROC FOM-statistic

The empirical weighted-AFROC plot and lesion weights were defined in Section 25.8. The empirical weighted-AFROC FOM-statistic (Chakraborty and Berbaum, 2004) is defined by including the lesion weights  $W_{k_2 l_2}$  inside the summations (but outside the kernel function):

$$\theta_{\text{wAFROC}} = \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} \psi(FP_{k_1 1}, z_{k_2 2 l_2 2}) \quad (28.5)$$

The weights obey the constraint:

$$\sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} = 1 \quad (28.6)$$

This ensures, as will be shown shortly, that each diseased case contributes equally to the FOM, regardless of how many lesions are in it. In the special case of one lesion per diseased case,  $\theta_{\text{AFROC}}$  and  $\theta_{\text{wAFROC}}$  are identical. For equally weighted lesions,

$$W_{k_2 l_2} = \frac{1}{L_{k_2}} \quad (28.7)$$

For example, for equally weighted lesions and a case with three lesions, each weight equals one-third ( $1/3$ )<sup>2</sup>.

## 28.5 Two Theorems

The area  $A_{\text{wAFROC}}$  under the wAFROC plot is obtained by summing the areas of individual trapezoids defined by drawing vertical lines from each pair of adjacent operating points to the x-axis. A sample plot is shown Fig. 28.1.

The operating point labeled  $i$  has coordinates  $(FPF_i, \text{wLLF}_i)$  given by Eqn. (25.14) and Eqn. (25.22), respectively, reproduced here for convenience:

$$FPF_i \equiv FPF(\zeta_i) = \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}(FP_{k_1 1} \geq \zeta_i) \quad (28.8)$$

$$\text{wLLF}_i \equiv \text{wLLF}_{\zeta_i} = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} \mathbb{I}(z_{k_2 2 l_2 2} \geq \zeta_i) \quad (28.9)$$

TBA Online Appendix 14.A proves the following theorems:

---

<sup>2</sup>The `RJafroc` function `DfReadDataFile()` checks that the weights sum to unity to a precision of about 5 decimal places. The easy way to assign equal weights to all lesions on a diseased case is to set the corresponding `lesionWeights` field in the Excel file `Truth` worksheet to zeroes.

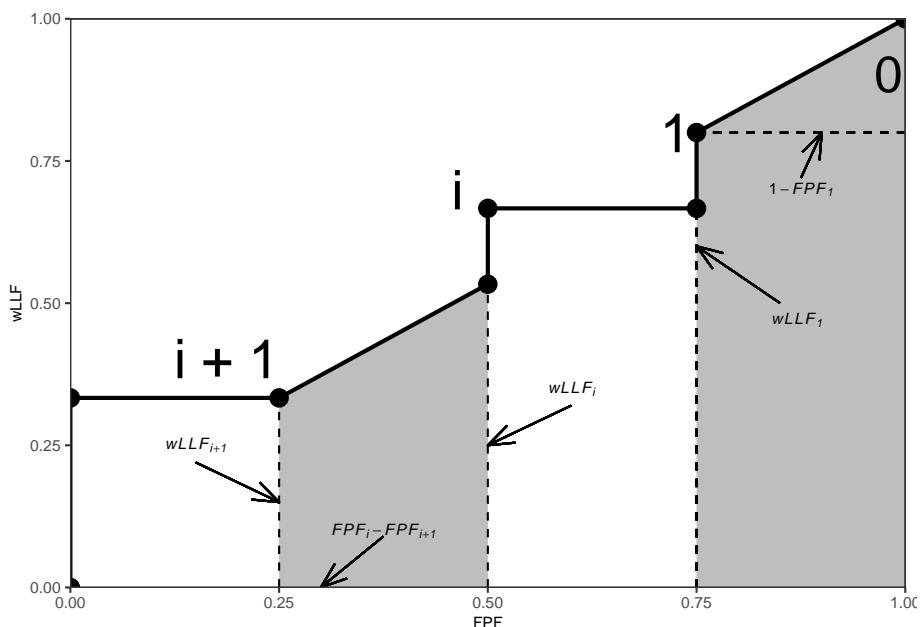


Figure 28.1: An example wAFROC plot; from left to right, the two shaded areas correspond to  $A_i$  and  $A_0$ , respectively, defined below.

### 28.5.1 Theorem 1

The area  $A_{\text{wAFROC}}$  under the empirical wAFROC plot equals the weighted-AFROC FOM-statistic  $\theta_{\text{wAFROC}}$  defined by Eqn. (28.5):

$$\theta_{\text{wAFROC}} = A_{\text{wAFROC}} \quad (28.10)$$

This is the FROC counterpart of Bamber's Wilcoxon vs. empirical ROC area equivalence theorem (Bamber, 1975), derived in Section 12.7.

### 28.5.2 Theorem 2

The area  $A_0$  under the straight-line extension of the wAFROC from the observed end-point  $(\text{FPF}_1, \text{wLLF}_1)$  to  $(1,1)$  is given by:

$$A_0 = \frac{(1 - \text{FPF}_1)(1 + \text{wLLF}_1)}{2} \quad (28.11)$$

According to Eqn. (28.11),  $A_0$  increases as  $\text{FPF}_1$  decreases, i.e., as more non-diseased cases are *not marked* and as  $\text{wLLF}_1$  increases, i.e., as more lesions, especially those with greater weights, *are marked*. Both observations are in keeping with the behavior of a valid FOM.

- Failure to include the area under the straight-line extension results in not counting the full positive contribution to the FOM of unmarked non-diseased cases and marked lesions.
- Each unmarked non-diseased case represents a perfect decision.
- For a perfect observer whose operating characteristic is the vertical line from  $(0,0)$  to  $(0,1)$  followed by the horizontal line from  $(0,1)$  to  $(1,1)$ , *the area under the straight-line extension comprises the entire AUC*. Excluding it would yield zero AUC for a perfect observer, which is obviously incorrect.
- Stated equivalently, for the perfect observer  $\text{FPF}_1 = 0$  and  $\text{wLLF}_1 = 1$  and then, according to Eqn. (28.11), the area under the straight line extension is  $A_0 = 1$ .

## 28.6 Numerical illustrations

The wAFROC and AFROC concepts are perhaps best illustrated with a numerical simulation-based illustration with very few cases.

Parameters of the simulation are  $\mu = 2$ ,  $\lambda = 1$ ,  $\nu = 1$ ,  $\zeta_1 = -1$  and  $L_{\max} = 2$ . One simulates a dataset consisting of  $K_1 = 4$  non-diseased cases and  $K_2 =$

4 diseased cases. The first two diseased cases have one lesion each, and the remaining two have two lesions each.

```
#> AFROC AUC = 0.7708333
#> wAFROC AUC = 0.7875
```

Shown in Fig. 28.2 are the AFROC and wAFROC plots with operating points.

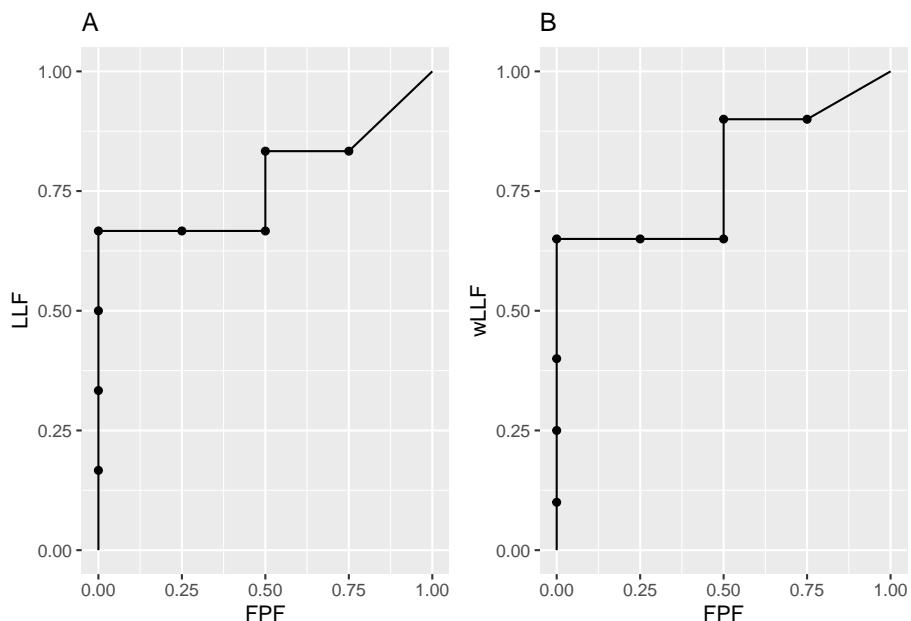


Figure 28.2: Left: AFROC plot; Right: corresponding wAFROC plot.

The number of lesions for diseased cases is shown next. Notice that the first two cases have one lesion each and the next two have two lesions each.

```
Lk2
#> [1] 1 1 2 2
```

The ratings are shown next.

```
x1 <- as.data.frame(frocData$ratings$NL[1,1,,])
colnames(x1) <- c("location1", "location2")
x2 <- as.data.frame(frocData$ratings$LL[1,1,,])
colnames(x2) <- c("location1", "location2")
x1
```

```
#>      location1 location2
#> 1      -Inf      -Inf
#> 2  0.4874291      -Inf
#> 3  0.7383247 0.5757814
#> 4 -0.3053884      -Inf
#> 5  1.5117812      -Inf
#> 6      -Inf      -Inf
#> 7      -Inf      -Inf
#> 8      -Inf      -Inf
x2
#>      location1 location2
#> 1  0.8523430      -Inf
#> 2 -0.2146999      -Inf
#> 3  1.5884892      -Inf
#> 4  2.9438362  1.98381
```

- The length of the third dimension of the NL array is eight (4 non-diseased + 4 diseased cases).
- The fifth sequential case corresponds to NLs on the first diseased case, etc.
- The first non-diseased case has no latent marks.
- The second non-diseased case has one latent mark rated 0.4874291.
- The third non-diseased case has two latent marks rated 0.7383247 and 0.5757814.
- The fourth non-diseased case has one latent mark rated -0.3053884.
- The first diseased case has one latent NL mark rated 1.5117812.
- The remaining diseased case have no latent NL marks.

## 28.7 Summary tables of ratings

Table 28.1 shows the layout of mark-rating pairs on *non-diseased* cases, illustrating FP ratings, corresponding to the green circles in Fig. 28.3. [UM denotes an unmarked location and blank cells denote unrealized z-samples.]

- Because non-diseased cases have no lesions, all z-samples listed in this table are for NLs.
- The first column lists the case numbers.
- The column labeled  $k_t l_s s$  lists the case-location indexing subscripts.
- The column labeled  $z_{k_t l_s s}$  lists the corresponding z-samples, when realized, and otherwise the cells are blank.
- The column labeled  $FP_{k_t t}$  lists the FP rating for each non-diseased case, which is the highest of all realized z-samples on the case or  $-\infty$  if none are realized.

Table 28.1: Layout of mark-rating pairs on non-diseased cases; UM denotes an unmarked non-diseased case.

	$k_t tl_s s$	$z_{k_t tl_s s}$	$FP_{k_t t}$	Label
	1111			
1	1121		$-\infty$	UM
	2111	0.487		
2	2121		0.487	F
	3111	0.738		
3	3121	0.576	0.738	E
	4111	-0.305		
4	4121		-0.305	H

Table 28.2: Layout of mark-rating pairs on diseased cases; UM denotes an unmarked lesion.

	$L_{k_2}$	$k_t tl_s s$	$z_{k_t tl_s s}$	$k_t tl_s s$	$z_{k_t tl_s s}$	weights	Label
		1211	1.512	1212	0.852	1	D
1	1	1221		1222			
		2211		2212	-0.215	1	G
2	1	2221		2222			
		3211		3212	1.588	0.6	C
3	2	3221		3222		0.4	UM
		4211		4212	2.944	0.4	A
4	2	4221		4222	1.984	0.6	B

- Column 5: the labels **A** - **H** correspond to the operating points shown in Fig. 28.3 and Fig. 28.4.

Table 28.2 shows the layout of mark-rating pairs on *diseased* cases, illustrating LL ratings, corresponding to the red circles in Fig. 28.3. [UM denotes an unmarked location and blank cells denote unrealized z-samples.]

- Because diseased cases can have NLs and LLs, both are shown in this table.
- The first column lists the case numbers.
- The second column lists the number of lesions present.
- Columns 3 and 4 illustrate NL indexing and z-samples.
- Columns 5 and 6 illustrate LL indexing and z-samples.
- Column 7 lists the lesion weights.
- Column 8: the labels **A** - **H** correspond to the operating points shown in Fig. 28.3 and Fig. 28.4.

## 28.8 AFROC plot from first principles

The following example is based on the same data involving 8 cases that were used to generate Table 28.1 and Table 28.2. It involves use of a linear or “one-dimensional” depiction of the ratings described next.

In Fig. 28.3, plot A, FPs and LLs, represented by green and red circles, respectively, are shown ordered, from left to right, with higher z-samples to the right, henceforth referred to as a *linear plot*. Each circle is labeled using the  $k_t l_s$  notation. For example, the right-most red circle corresponds to the LL z-sample originating from the first lesion in the fourth diseased case, i.e.,  $z_{4212}$ . Consistent with the three unique values in the fourth column of Table 28.1, there are three green circles (FPs)<sup>3</sup>. Likewise, consistent with the five unique values in the sixth column of Table 28.2, there are five red circles (LLs)<sup>4</sup>.

Starting from  $\infty$ , moving a virtual threshold continuously to the left generates the AFROC plot, see plot A in Fig. 28.3. As each FP is crossed, the operating point moves to the right by:

$$\frac{1}{K_1} = 0.025$$

As each LL is crossed, the operating point moves up by:

$$\sum_{k_2=1}^{K_2} L_{k_2} = \frac{1}{6}$$

Since it has one lesion, crossing the z-sample for the first case would result in an upward movement of  $1/6$ , and likewise for the second case. Since the third case contains two lesions, crossing the corresponding z-samples would result in a net upward movement of the operating point by  $1/3$ . *This behavior shows explicitly that the non-weighted method gives greater importance to diseased cases with more lesions, i.e., such cases make a greater contribution to AUC.* The jumps from lesions in the same case need not be contiguous – they could be distributed, with intervening jumps from lesions on other cases, but eventually the jumps will occur and contribute to the net upward movement. As an example, the jumps due to the two lesions on the fourth diseased case are contiguous: see points A and B, in Fig. 28.3. However, the jumps due to the two lesions on the third diseased case are not contiguous: the first lesion gives the point C, but the unmarked lesion on this case, indicated by “UM” in Table 28.2, eventually contributes when the operating point moves diagonally from point H to (1,1).

---

<sup>3</sup>Not counting  $FP_{11}$ , which occurs at  $z = -\infty$ , representing the first non-diseased case with no marks.

<sup>4</sup>Not counting  $z_{3222} = -\infty$  representing the unmarked second lesion on the third diseased case.

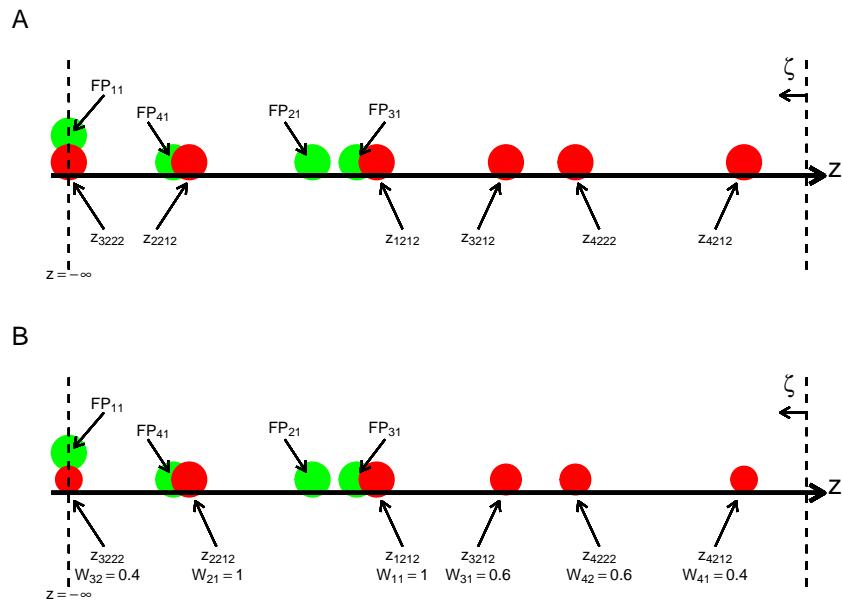


Figure 28.3: Plot A (illustrating generation of the AFROC): a one-dimensional depiction of the data in Table 28.1 and Table 28.2, showing  $z$ -samples used for plotting the AFROC; the red circles correspond to lesion localizations (LLs) and the green to false positives (FPs). Plot B (illustrating generation of the wAFROC): Data in same tables but this time including the weights, for plotting the weighted-AFROC plot; the sizes of the red circles code the lesions weights; the weights are shown below each  $z$ -sample.

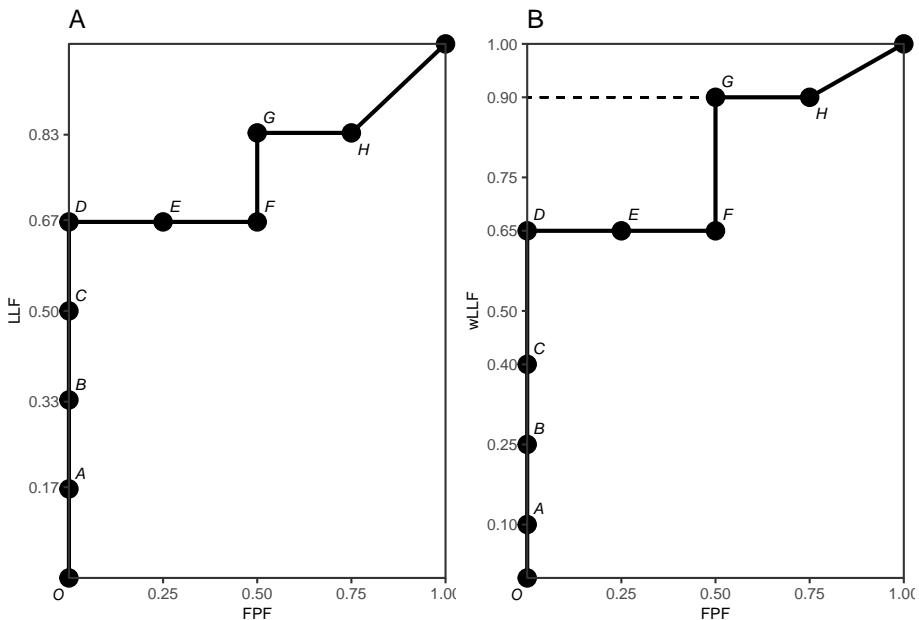


Figure 28.4: Plot A: The empirical AFROC plot for the data shown in Table 28.1 and Table 28.2. The labels correspond to the last columns of the tables. The corresponding one-dimensional depiction is plot A in Fig. 28.3. The area under the empirical plot is 0.7708. Plot B: The empirical weighted-AFROC (wAFROC) plot for the data shown in Table 28.1 and Table 28.2. The corresponding one-dimensional plot is plot B in Fig. 28.3. The area under the wAFROC is 0.7875.

## 28.9 wAFROC plot from first principles

Plot B in Fig. 28.3, which is the wAFROC analog of plot A in the same figure, is a one-dimensional depiction of the data in Table 28.1 and Table 28.2, but this time the lesion weights, shown in Table 28.2, are incorporated, as indicated by varying the size of each red circle (in Fig. 28.3 plot A, all red circles were of the same size). In addition, each lesion is labeled with its rating and weight.

Moving a virtual threshold continuously to the left generates the wAFROC plot, Fig. 28.4 plot B. The movement of the operating point in response to crossing FPs is the same as before. However, as each LL is crossed the operating point moves up by an amount that depends on the lesion weight:

$$\frac{W_{k_2 l_2}}{K_2} = \frac{W_{k_2 l_2}}{4}$$

Since the first two diseased cases have one lesion each (i.e., unit weights), crossing the corresponding z-samples results in upward jumps of 0.25, Fig. 28.4 plot B – compare the jumps C to D and from F to G. According to the weights in Fig. 14.4, crossing the z-sample of the first lesion in the third diseased case, results in an upward jump of 0.6/4. That from the second lesion in the same case results in an upward jump of 0.4/4, for a net upward jump of the third case of 0.25, the same as for each of the first two diseased cases. Likewise crossing the z-samples of the two lesions in the 4th disease case results in upward jump of 0.4/4 = 0.1 (compare the jump from O to A), for the 1st lesion and 0.6/4 = 0.15 (compare the jump from B to C), for the 2nd lesion, for a net upward jump of 1/4, which is the same as for each of the first three diseased cases. *This shows explicitly that the weighting method gives each diseased case the same importance, regardless of the number of lesions in it, a property not shared by the area under the AFROC.*

## 28.10 Physical interpretations

From the preceding sections, it is seen that the AFROC-based trapezoidal plots consist of upward and rightward jumps, starting from the origin (0,0) and ending at (1,1). This is true regardless of whether the z-samples are binned or not: i.e., at the “microscopic” level the jumps always exist. Each upward jump is associated with a LL rating exceeding a virtual threshold. Each rightward jump is associated with a FP rating exceeding the threshold. Upward jumps tend to increase the area under the AFROC-based plots and rightward jumps tend to decrease it. This makes physical sense in terms of correct decisions being rewarded and incorrect ones being penalized, and can be seen from two extreme-case examples. If there are only upward jumps, then the trapezoidal plot rises from the origin to (0,1), where all lesions are correctly localized without any

generating FPs and performance is perfect – the straight-line extension to (1,1) ensures that the net area is unity. If there are only horizontal jumps, that takes the operating point from the origin to (1,0), where none of the lesions are localized and every non-diseased image has at least one NL mark, representing worst possible performance. Here, despite the straight line extension to (1,1), the net area is zero.

### 28.10.1 Physical interpretation of area under AFROC

The area under the AFROC has the following physical interpretation: it is the fraction of LL vs. FP z-sample comparisons where the LL sample is equal (counting as half a comparison) or greater (counting as a full comparison) than the FP z-sample. From Tables 1 and 2, there are four FPs and six LLs for 24 possible comparisons. Inspection of the tables reveals that there are  $4 \times 4 = 16$  comparisons contributing ones, two comparisons (from the 2nd diseased case) contributing ones, and one comparison (from the 2nd lesion on the 3rd diseased case) contributing 0.5, which sum to 18.5. Dividing by 24 yields  $18.5/24 = 0.7708$ , the empirical TBA AFROC-AUC, §14.5.1. In probabilistic terms:

*The area under the AFROC is the probability that a lesion is rated higher than any mark on a non-diseased case.*

### 28.10.2 Physical interpretation of area under wAFROC

The area under the wAFROC has the following physical interpretation: it is the lesion-weight adjusted fraction of diseased cases vs. non-diseased case comparisons where LL z-samples are equal (counting as half a comparison times the weight of the lesion in question) or greater (counting as a full comparison times the weight of the lesion) than FP z-samples. Note that there are still 24 LL vs. FP comparisons but the counting proceeds differently. The fourth diseased case contributes  $0.4 \times 4 + 0.6 \times 4$ , i.e., 4 (compared to 8 in the preceding example). The third diseased case contributes  $0.6 \times 4 + 0.4 \times 0.5$ , i.e., 2.6 (compared to 4.5 in the preceding example). The second diseased case contributes  $1 \times 2 = 2$  (compared to 2 in the preceding example), and the first diseased case contributes  $1 \times 4 = 4$  (compared to 4 in the preceding example). Summing these values and dividing by 16 (the total number of diseased cases vs. non-diseased cases comparisons) one gets  $12.6/16 = 0.7875$ , which is the area under the wAFROC, §14.5.1. In probabilistic terms:

*The area under the weighted-AFROC is the lesion-weight adjusted probability that a lesion is rated higher than any mark on a non-diseased case.*

## 28.11 Discussion

TBA TODOLAST

The primary aim of this chapter was to develop expressions for FOMs (i.e., functions of ratings) and show their equivalences to the empirical AUCs under corresponding operating characteristics. Unlike the ROC, the AFROC and wAFROC figures of merit are represented by quasi-Wilcoxon like constructs, not the well-known Wilcoxon statistic<sup>5</sup>.

I am aware from users of my software that their manuscript submissions have sometimes been held up with the critique that the meaning of the AFROC FOM-statistic is “not intuitively clear”<sup>6</sup> TBA. Any critique based on intuitive clarity or lack thereof suffers from a fundamental flaw: it is un-falsifiable. What is “intuitively not clear” to one could be “intuitively very clear” to another, and there is no way of testing either viewpoint. Un-falsifiable claims have no place in science.

An example was given in a previous chapter. This is one reason I have tried to make the meaning clear, perhaps at the risk of making it painfully clear. Clinical interpretations do not always fit into convenient easy to analyze paradigms. Not understanding something is not a reason for preferring a simpler method. Use of the simpler ROC paradigm to analyze location specific tasks results in loss of statistical power and sacrifices better understanding of what is limiting performance. It is unethical to analyze a study with a method with lower statistical power when one with greater power is available<sup>7-9</sup>. The title of the paper by Halpern et al is “The continuing unethical conduct of under-powered clinical trials”. The AFROC FOM-statistic was proposed in 1989 and it has been used, at the time of writing, in over 107 publications .

The subject material of this chapter is not that difficult. However, it does require the researcher to be receptive an unbiased. Dirac addressed an analogous then-existing concern about quantum mechanics, namely it did not provide a satisfying “picture” of what is going on, as did classical mechanics . To paraphrase him, the purpose of science (quantum physics in his case) is not to provide satisfying pictures but to explain data. FROC data is inherently more complex than the ROC paradigm and one should not expect a simple FOM-statistic. The detailed explanations given in this chapter should allow one to understand the wAFROC and AFROC FOMs.

A misconception regarding the wAFROC FOM-statistic is that the weighting may sacrifice statistical power and render the method equivalent to ROC analysis in terms of statistical power. Analysis of clinical datasets and simulation studies suggests that this is not the case; loss of power is minimal. As noted earlier, the highest rating carries more information than a randomly selected rating.

Bamber’s equivalence theorem led to much progress in non-parametric analysis of ROC data. The proofs of the equivalences between the areas under the

AFROC and wAFROC and the corresponding quasi-Wilcoxon statistics provide a starting point. To realize the full potential of these proofs, similar work like that conducted by DeLong et al<sup>10</sup> is needed for the FROC paradigm. This work is not going to be easy; one reason being the relative dearth of researchers working in this area, but it is possible. Indeed work has been published by Popescu<sup>11</sup> on non-parametric analysis of the exponentially transformed FROC (EFROC) plot which, like the AFROC and wAFROC, is completely contained within the unit square. This work should be extended to the wAFROC. For reasons stated in Chapter 13, non-parametric analysis of FROC curves<sup>12-14</sup> is not expected to be fruitful.

Current terminology prefixes each of the AFROC-based FOMs with the letter “J” for Jackknife. The author recommends dropping this prefix, which has to do with significance testing procedure rather than the actual definition of the FOM-statistic. For example, the correct way is to refer to the AFROC figure of merit, not the JAFROC figure of merit. For continuity, the software packages implementing the methods are still referred to as JAFROC (Windows) or RJAfroc (cross-platform, open-source).

To gain deeper insight into the FROC paradigm, it is necessary to look at methods used to measure visual search, the subject of the next chapter.

## 28.12 References



# Chapter 29

## Visual Search

### 29.1 TBA How much finished

10%

### 29.2 Introduction

To understand free-response data, specifically how radiologists interpret images, one must come to grips with visual search. Casual usage of everyday terms like “search”, “recognition” and “detection” in specific scientific contexts can lead to confusion, so in this chapter I will attempt to carefully define them. Visual search, including the medical imaging search task, is defined in a broad sense as grouping and labeling parts of an image. Two experimental methods for studying search are described. The more common method, which is widely used in the non-medical imaging context, consists of showing observers known targets (i.e., known shape and size but unknown location) and known distractors (again, known shapes and sizes but unknown locations). One measures how rapidly the observer can perceive the presence of the target (reaction time) and their accuracy (fraction of correct decisions on target present vs. target absent presentations). In the medical imaging paradigm, one does not know the sizes, shapes and locations of the targets and distractors. Instead, one relies on eye-tracking measurements to determine where the observer is looking and for how long. A clustering algorithm is applied to determine regions that received prolonged examination (dwell time) and presumably are the sites where decisions were made. The focus in this chapter is on the second paradigm, which closely parallels the FROC task. A schema of how radiologists find lesions, termed the Kundel-Nodine search model is described. The importance of this

major conceptual model is not widely appreciated by researchers. It is a two-stage model, where the first stage identifies suspicious regions. The second stage analyzes the suspicious regions and if the level of suspicion is high enough, the region is marked. The Kundel-Nodine search model is the basis of the radiological search model (RSM) described in the next chapter. A section is devoted to a recently developed method for analyzing simultaneously acquired eye tracking and FROC data.

The starting point is the definition of recognition/finding. The following sections draw heavily on work by Nodine and Kundel<sup>1-5</sup>. The author also acknowledges critical insights gained through conversations with Dr. Claudia Mello-Thoms.

### 29.3 Grouping and labeling ROIs

Looking at and understanding an image involves grouping and assigning labels to different regions of interest (ROIs) in the image, where the labels correspond to entities that exist (or have existed in the examples to follow) in the real world. As an example, if one looks at Fig. 15.1, one would label them (from left to right and top to bottom, in raster fashion): Franklin Roosevelt, Harry Truman, Lyndon Johnson, Richard Nixon, Jimmy Carter, Ronald Reagan, George H. W. Bush, and the presidential seal. The accuracy of the labeling depends on prior-knowledge, i.e., expertise, of the observer. If one were ignorant about US presidents, one would be unable to correctly label them.

Image interpretation in radiology is not fundamentally different. It involves assigning labels to an image by grouping and recognizing areas of the image that have correspondences to the radiologist's knowledge of the underlying anatomy, and, most importantly, deviations from the underlying anatomy. Most doctors, who need not be radiologists, can look at a chest x-ray and say, "this is the heart", "this is a rib", "this is a clavicle", "this is the aortic arch", etc., Fig. 15.2 (A). This is because they know the underlying anatomy, Fig. 15.2 (B) and have a basic understanding of the x-ray image formation physics that relates the anatomy to the image.

### 29.4 Recognition vs. detection

The process of grouping and labeling parts of an image is termed recognition. This was illustrated with the pictures of the US presidents, Fig. 15.1. Recognition is distinct from detection, which is deciding about the presence of something that is unexpected or the absence of something that is expected, in other words, a deviation, in either direction, from what is expected. An example of detecting the presence of something that is unexpected would be a lung nodule and an example of detecting the absence of something that is expected would be an



Figure 29.1: This image consists of 8 sub-images or ROIs. Understanding an image involves grouping and assigning labels to different ROIs, where the labels correspond to entities that exist in the real world. One familiar with US history would label them, from left to right and top to bottom, in raster fashion, Franklin Roosevelt, Harry Truman, Lyndon Johnson, Richard Nixon, Jimmy Carter, Ronald Reagan, George H. Bush and the presidential seal. Labeling accuracy depends on expertise of the observer. The row and column index of each ROI identifies its location.

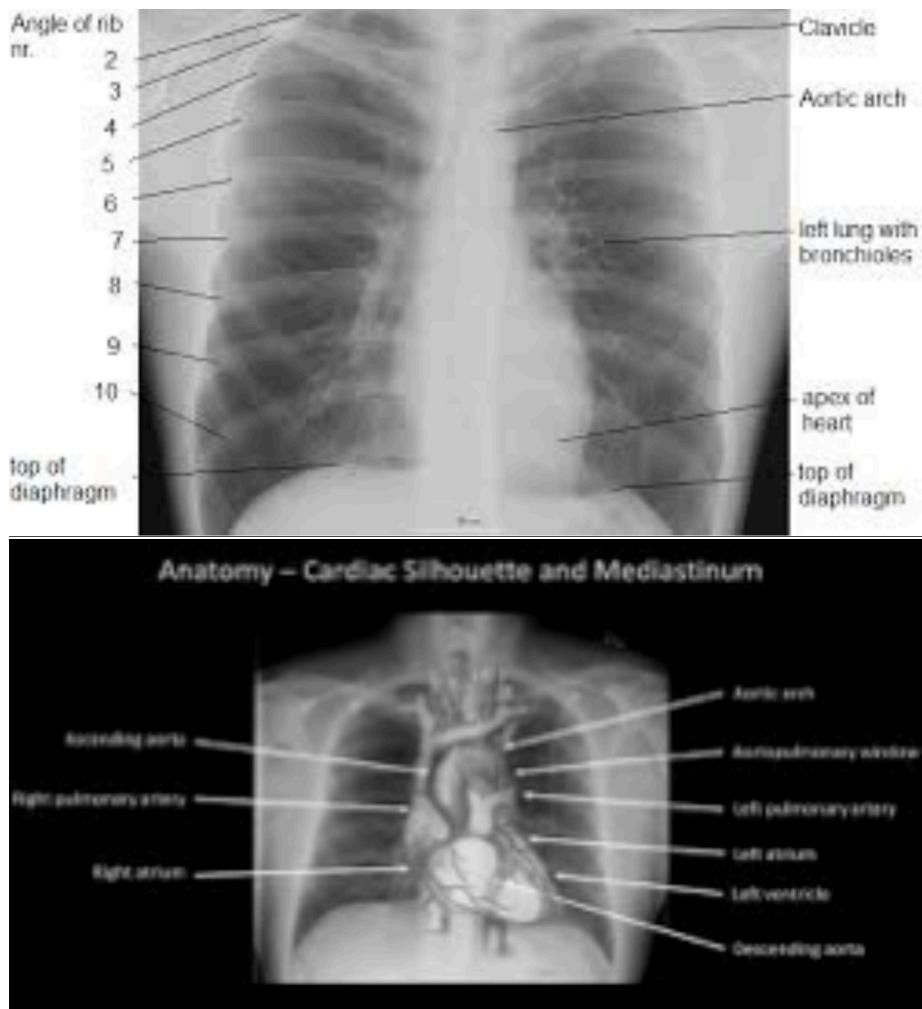


Figure 29.2: Image interpretation in radiology also involves assigning labels to an image by grouping and recognizing areas of the image that have correspondences to the radiologist's knowledge of the underlying anatomy. (A) Most doctors can look at a chest x-ray and say, "this is the heart", "this is a rib", "this is the clavicle", "this is the aortic arch", etc. (B) This is because they know the underlying anatomy and have a basic understanding of x-ray image formation physics that relates anatomy to the image.

image of a patient with a missing rib (yes, it does occur, even excluding the biblical Adam).

The terms “expected” and “unexpected” are important: they imply expertise dependent expectations regarding the true structure of the non-diseased image, which I term a non-diseased template, and therefore an ability to recognize clinically relevant deviations or perturbations, in either direction, from this template; e.g., a lung nodule that could be cancer. By “clinically relevant” I mean perturbations related to the patient’s health outcome – recognizing scratches, dead pixels, artifacts of known origin, and lead patient ID markers, do not count. There is a location associated with recognition , but not with detection. Detection is the presence or absence of something: the perturbation could be anywhere. For example, in Fig. 15.1, recognizing a face is equivalent to assigning a row and column index in the image. Specifically, recognizing of George H.W. Bush implies pointing to row = two and column = three. Detecting George H.W. Bush implies stating that George H.W. Bush is present in the image, but the location could be in any of the eight locations. Recognition is an FROC paradigm task, while detection is an ROC paradigm task. Instead of recognition, I prefer the more clinical term “finding”, as in “finding” a lesion: in the clinic radiologists report “findings”.

## 29.5 Search vs. classification

Since template perturbations can occur at different locations in the images, the ability to selectively recognize them is related to search expertise. The term “selectively” is important: a non-expert can trivially recognize all perturbations by claiming all regions in the image are perturbed. Search expertise is the selective ability to find clinically relevant perturbations that are actually present while minimizing finding what appear to be clinically relevant perturbations that are actually not present. In FROC terminology, search expertise is the ability to find latent LLs while minimizing the numbers of found latent NLs. Lesion-classification expertise is the ability to correctly classify a found suspicious region as malignant or benign.

The skills required to recognize a nodule in a chest x-ray are different from that required to recognize a low-contrast circular or Gaussian shaped artificial nodule against a background of random noise. In the former instance the skills of the radiologist are relevant: e.g., the skilled radiologist knows not to confuse a blood vessel viewed “end on” for a nodule, especially since the radiologist knows where to expect these vessels, e.g., the aorta. In the latter instance, (i.e., viewing artificial nodules superposed on random noise) there are no expected anatomic structures, so the skills possessed by the radiologist are rendered irrelevant. This is the reason why having radiologists interpret random noise images and pretending that this makes it “clinically relevant” is a waste of reader resources and bad science. One might as well used undergraduates with good eyesight,

motivation and training. To quote Nodine and Kundel<sup>1</sup> “Detecting an object that is hidden in a natural scene is not the same as detecting an object displayed against a background of random noise.” This paragraph also argues against usage of phantoms as stand-ins for clinical images for “clinical” performance assessment. Phantoms are fine in the QC context, as in Chapter 01, but they do not allow radiologists the opportunity to exercise their professional skills.

## 29.6 Two visual search paradigms

There are two visual search paradigms: what I term the conventional paradigm and the medical imaging paradigm.

### 29.6.1 The conventional paradigm

In the conventional paradigm, one measures reaction time and percent correct in the following task. Images are shown briefly and followed, after an intersimulus interval, by a mask image (e.g., random noise, to “wipe-out” memory of the preceding image). Each image may contain a defined target in a set of defined distractors. Defined targets and defined distractors mean that their presence and numbers are under the control of the experimenter and the observer, via training images, knows their characteristics (e.g., shapes, sizes, etc.). For example<sup>6</sup> a target could be the letter “T” and distractors could be the letter “L”. The observer’s task is to discriminate between two conditions: (i) target and distractors present and (ii) only distractors present, by pressing a “yes” (target is present) or “no” key. Also measured is the time it takes, from image onset, to make the target-present target-absent decision, termed reaction time. This is by far the most widely used paradigm<sup>7,8</sup> (see for example Ref. 7 and the literature cited therein; the paper, cited 2908 times as of 8/18/2016, is an excellent review of this paradigm). Typically measured is the dependence of percent correct and reaction time on set size (defined as the number of distractors). The following example, adapted from Ref. 8, describes an actual study using this paradigm:

Stimuli (stimuli = items = distractors plus a possible target) were red and green Xs and Os on a black background. Individual items could be placed at any of 36 locations within a pre-defined square field. On each trial, items were presented at 8, 16, or 32 randomly chosen positions within the square field (thereby varying the set size, i.e., the number of distractors). The target was a red O (in target present images) and distractors were green Os and red Xs. On target present trials, one of the locations contained a target item. Targets were present on 50% of trials. Set size, positions of target and distractors, and presence or absence of a target was random across trials. Subjects responded by pressing one of two keys: a yes

key if a target was detected and a no key if it was not. Reaction times were measured from stimulus onset. The stimuli remained visible until the subject responded and feedback was given on each trial.

The results are used to test different models of visual search. In particular, there has been interest in determining if the items are processed in parallel or sequentially. There is, as stated above, a very large literature on this paradigm, or variants of it, and this brief account is given simply to distinguish it from the medical imaging paradigm that follows.

### 29.6.2 The medical imaging visual search paradigm

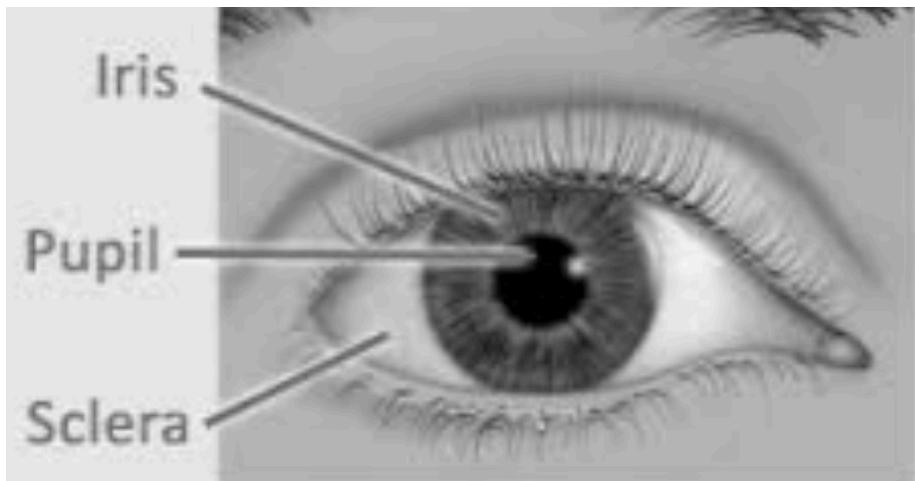


Figure 29.3: Eye position data were recorded using a limbus reflection technique. [Limbus is defined as the border between the cornea (the transparent layer making up the outermost front part of the eye, covering the iris and pupil) and sclera (opaque white of the eye).] Eye movements are measured by having the observer wear a specially designed spectacle frame (newer machines do not require this) containing infrared emitters and sensors that measure changes in light reflected from the border between the iris and sclera.

The key difference is the dependence on eye-tracker technology<sup>1,9,10</sup>. This is not to imply that users of the conventional paradigm have not used eye-tracker technology. They have, but the medical imaging paradigm is crucially dependent on this technology whereas the conventional paradigm is not. Eye-tracker technology determines the location and duration of the axis of gaze (i.e., where and for how long the radiologist looks at different locations in the image. The difference between the two paradigms is necessitated by several factors:

1. Unlike the conventional paradigm, one does not know the numbers and precise shapes, sizes, contrasts, etc., of true lesions, the “targets” in conventional terminology. These are camouflaged in anatomic noise and are more complex than the “Ts” and “Ls”, or “Xs” and “Os”, of the conventional method. 2. One does not know the numbers and precise shapes, sizes, contrasts, etc., of the “distractors”. In fact, the radiologist perceives these and what constitutes a latent NL to one radiologist may not be a latent NL to another. Two radiologists may not even agree on the number of latent NLs on a specific image. Unlike the conventional paradigm, the number of NLs in the medical imaging paradigm must be treated as a radiologist dependent random number. 3. The medical image paradigm allows for zero latent marks (i.e., no distractors), which has no counterpart in the conventional method. These images are the “unambiguous” non-diseased cases that do not generate any marks. 4. In medical imaging, one is more interested in objective performance measurements (does a radiologist find the lesion at high confidence?) than in reaction time. 5. In addition to eye-tracking data one may acquire ratings data as in the ROC paradigm<sup>11</sup>, or more recently, mark-rating data, as in the FROC paradigm<sup>12,13</sup>. If using the ROC method the performance measure (e.g., AUC under ROC curve) is comparable to the percent correct obtained in the conventional paradigm, except that the ROC-ratings method is more efficient<sup>14</sup>. However, since location of the perceived target or lesion is ignored, the scoring is ambiguous in the sense originally noted by Bunch et al<sup>15</sup>: i.e., the observer may have not seen the target and mistaken a distractor for the target on a target present image, and that event-combination would be scored as a correct decision. The FROC paradigm accounts for location, thereby ruling out this ambiguity.

Compared to the many papers using the conventional visual search paradigm, research in medical imaging visual search is relatively limited. Prof. Kundel, Prof. Nodine, Prof. Krupinski and Dr. Claudia Mello-Thoms have made major contributions to this field. The following is an example of how data is collected in the medical imaging visual search paradigm<sup>16</sup>.

Eye position data were recorded using a limbus reflection technique. [Limbus, Fig. 15.3, is defined as the border between the cornea (the transparent layer making up the outermost front part of the eye, covering the iris and pupil) and sclera (opaque white of the eye).] Eye movements are measured by having the observer wear a specially designed spectacle frame (newer machines do not require this) containing infrared emitters and sensors that measure changes in light reflected from the border between the iris and sclera, Fig. 15.3. The viewers were told they had 15 seconds to search the lung fields of each image for the presence of a nodule and additionally to remember the locations of regions suspected of containing a nodule but considered negative. Following the 15-second presentation, the viewers rated each image for presence of disease.

Eye-position data is collected only during the initial 15 seconds while the radiologist searches the image. One issue with this way of collecting data is that during the reporting phase, a radiologist may discover something new and pro-

ceed to investigate this finding, but because eye-position recording has been terminated, that information is not captured. In the data collection methodology used in a recent study<sup>13</sup> searching and reporting occur simultaneously with eye-position collection . The newer paradigm more closely resembles clinical practice, and potentially allows one to follow the perceptual and interpretative process entailed in case reading from beginning to end, without researcher-initiated interruptions.

## 29.7 Determining where the radiologist looks

The eye-tracking (ET) device I am familiar with TBA 17 (ASL Model H6, Applied Sciences Laboratory, Bedford, MA) uses a magnetic head tracker to monitor head position, and this allows the radiologists to freely move their head from side to side as well as towards the displays. The ET system integrates eye position calculated from limbus-reflection, and head position, to calculate the intersection of the line of gaze and the display plane. The data stream (raw-data) provided by the eye-tracker consists of several bytes of data at 60Hz, containing the (x,y) coordinates of where the observer is looking plus various flag bits (e.g., indicating blinks). The eye moves in rapid jumps (saccades) with intervening longer pauses (fixations). The eye-movement induced reflectance changes are converted to display coordinates, which indicate the locations and durations of fixations. Fixations occurring in clusters indicate where attention is being directed and decisions are made. The raw-data needs to be processed, Appendix 15A, to determine regions where decisions were made; the processing, which is guided by models of human perception, does depend on the researcher.

## 29.8 The Kundel - Nodine search model

The Kundel-Nodine model<sup>1-5</sup> is a schema of events that occur from the radiologist's first glance to the decision about the image. The model is similar to the guided search model<sup>7,8,18</sup> proposed by Prof. Jeremy Wolfe in the non-medical imaging context.

Assuming the task has been defined prior to viewing, based on eye-tracking recordings obtained on radiologists while they interpreted images, Kundel and Nodine proposed the following schema for the diagnostic interpretation process, consisting of two major components: (1) glancing or global impression and (2) scanning or feature analysis, Fig. 15.4.

### 29.8.1 Glancing / Global impression

The colloquial term "glancing" is meant literally . The glance is brief, typically lasting about 100 - 300 ms, too short for detailed foveal examination and

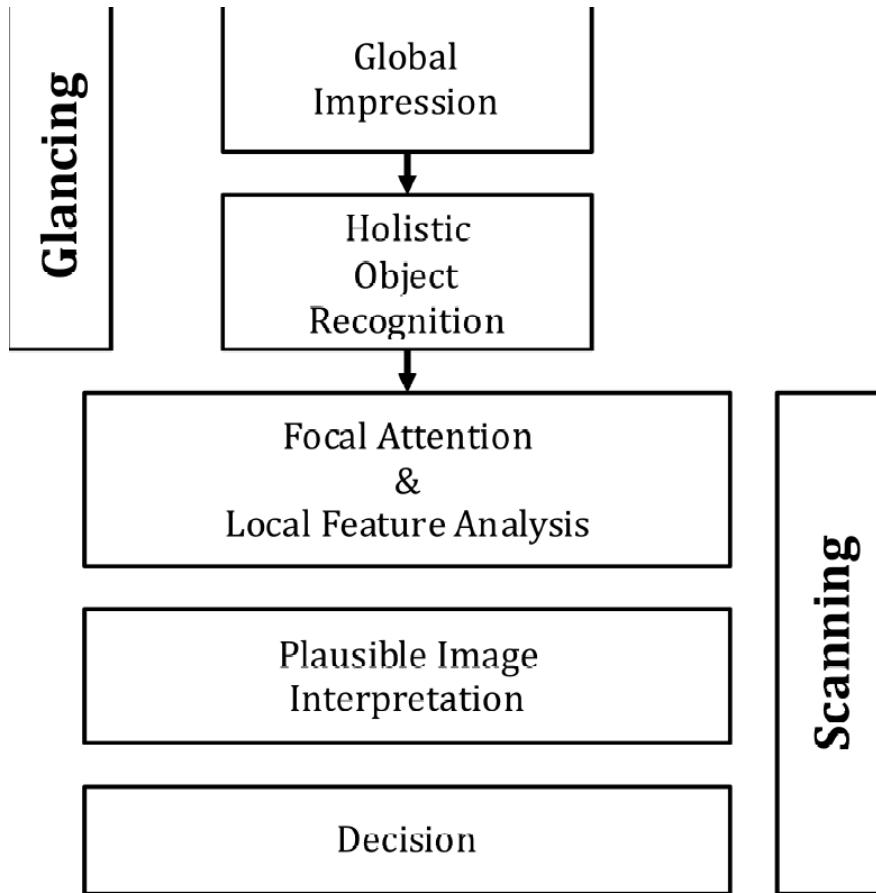


Figure 29.4: The Kundel-Nodine 2-stage model of radiological search. The glancing/global stage identifies perturbations from the template of a generic non-diseased case. The scanning stage performs detailed analysis of the identified perturbations and calculates the probability that the perturbation is a true lesion. Only perturbations with sufficiently high probability are marked/reported.

interpretation. Instead, during this brief interval peripheral vision and reader expertise are the primary mechanisms responsible for the identification of latent marks. The glance results in a global impression, or gestalt, that identifies perturbations from the template defined earlier. Object recognition occurs at a holistic level, i.e., in the context of the whole image, as there is insufficient time for detailed viewing and all of this is going on using peripheral vision. It is remarkable that radiologists can make reasonably accurate interpretations from information obtained in a brief glance (Fig. 6 in Ref. 1). Suspicious regions, which are perturbations from the template, are flagged for subsequent detailed viewing, i.e., the initial glance tells the visual system where to look more closely. See Chapter 12, section on “solar” analogy, for further background on this important aspect of vision. Since eye-tracking technology does not measure peripheral vision, the locations of the perturbations need to be inferred from the scanning stage described next.

### 29.8.2 Scanning / local feature analysis

The global impression identifies suspicious regions for detailed foveal viewing by the central vision<sup>19</sup>. During this process - termed scanning or feature analysis - the observer scrutinizes and analyzes the suspicious regions for evidence of possible disease. In principle, they calculate the probability of malignancy. For those readers more familiar with how CAD works, this corresponds to the feature analysis stage of CAD where regions found by the global search, termed initial detections in CAD, are analyzed for probability of malignancy. The scrutiny is conducted via clusters of closely spaced fixations. In the absence of closely spaced fixations or retinal jitter, in a laboratory condition known as retinal stabilization, perceptions tend to rapidly fade away<sup>20</sup>. Perception is sensitive to temporal changes; it there is a high-pass temporal filter that suppresses stationary features, so that changes from it are quickly perceived, no doubt a result of evolution. The evidence is used to decide whether to report the region. The corresponding locations are the “big-clusters” in Fig. 15.A.1 (C) in the online appendix. After places identified during the global impression have been scrutinized, the viewer may follow the same scanning pattern aimed at discovering something that was missed, or, may simply scan at random while thinking about the image.

The fixations that cluster at perturbations are collecting data necessary to test for the presence of a lesion. If testing yields a sufficiently high probability of lesion, a decision is made to report the lesion. If testing is negative or inconclusive, search continues. Thus, the report “normal chest” is an overall impression based on a series of local decisions that are needed because the relevant anatomic features can only be resolved by foveal vision. The viewer is not aware of all of the decisions, positive and negative, made during scanning . The eye-tracking record however, reveals where the eye lingered, providing indirect evidence about where covert decisions were made. However, as noted earlier,

the eye-tracking record does not include perturbations perceived by peripheral vision. It is believed that prolonged or multiple fixations that cluster on image detail signal the testing and decision-making activity associated with the interpretation of anatomical perturbations that have potential as tumor targets. This is the reason for the use, in Appendix 15A, of a total dwell time of 800 ms to determine where decisions occurred. The value is somewhat arbitrary and investigator dependent.

The essential point that emerges is that decisions are made at a finite, relatively small, number of regions. Attention units are not uniformly distributed through the image, in raster-scan fashion; rather the global impression identifies a smaller set of regions that require detailed scanning.

Eye-tracker recordings for a two-view digital mammogram for two observers are shown in Fig. 15.3, for an inexperienced observer (upper two panels) and an expert mammographer (lower two panels). The small circles indicate individual fixations (dwell time  $\sim 100$  ms). The larger high-contrast circles indicate clustered fixations (cumulative dwell time  $\sim 1$  s). The larger low-contrast circles indicate a mass visible on both views. The inexperienced observer finds many more suspicious regions than does the expert mammographer but misses the lesion in the MLO view. In other words, the inexperienced observer generates many latent NLs but only one latent LL. The mammographer finds the lesion in the MLO view, which qualifies as a latent LL, without finding suspicious regions in the non-diseased parenchyma, i.e., the expert generated zero latent NLs on this case and one latent LL. It is possible the observer was so confident in the malignancy found in the MLO view that there was no need to fixate the visible lesion in the other view - the decision had already been made to recall the patient for further imaging.

**Details:** Eye-tracking recordings for a two-view digital mammogram display for two observers, an inexperienced observer (upper two panels) and an expert mammographer (lower two panels). The small circles indicate individual fixations (dwell time  $\sim 100$  ms). The larger high-contrast circles indicate clustered fixations (cumulative dwell time  $\sim 1$  sec). The latter correspond to the latent marks in the search-model. The larger low-contrast circles indicate a mass visible on both views. The inexperienced observer finds many more suspicious regions than does the expert mammographer but misses the lesion in the MLO view. In other words the inexperienced observer generates many latent NLs but only one latent LL. The mammographer finds the lesion in the MLO view, which qualifies as a latent LL, without finding suspicious regions in the non-diseased parenchyma, i.e., the expert generated zero latent NLs on this case and one latent LL. It is possible the observer was so confident in the malignancy found in the MLO view that there was no need to fixate the visible lesion in the other view - the decision had already been made to recall the patient for further imaging, which confirmed the finding.

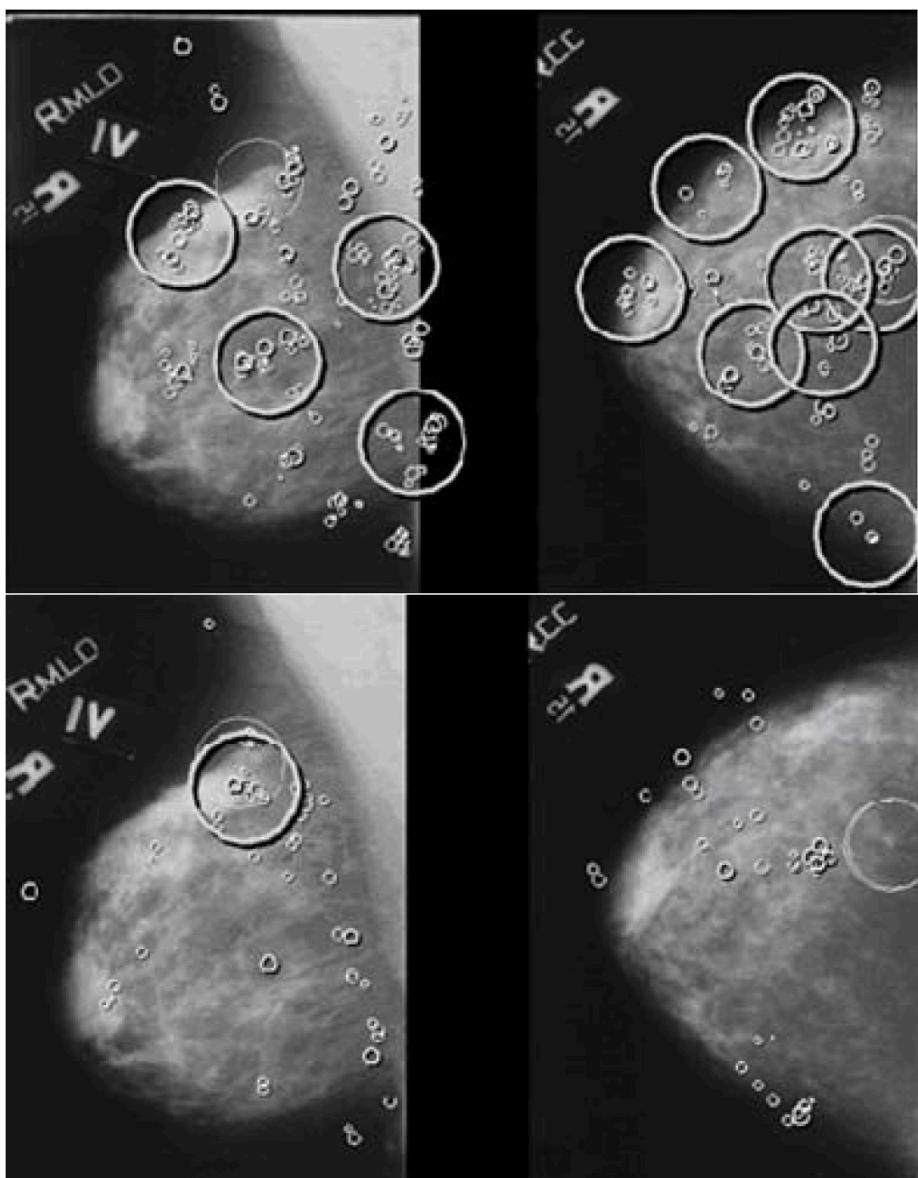


Figure 29.5: Eye-tracking recordings for a two-view digital mammogram: see details.

## 29.9 Kundel-Nodine model and CAD algorithms

It turns out that the designers of CAD algorithms independently arrived at a two-stage process remarkably similar to that described by Kundel-Nodine for radiologist observers. CAD algorithms are designed to emulate expert radiologists, and while this goal is not yet met, these algorithms are reasonable approximations to radiologists, and include the critical elements of search and localization that are central to clinical tasks. CAD algorithms involve two steps analogous to the holistic and cognitive stages of the Kundel-Nodine visual search model<sup>1,3,4</sup>. In other words, CAD has a perceptual correspondence to human observers that to my knowledge is not shared by other method of predicting what radiologists will call on clinical images.

In the first stage of CAD, termed initial detections<sup>21</sup>, the algorithm finds “all reasonable” regions that could possibly be a malignancy. The term “all reasonable” is used because an unreasonable observer could trivially “find” every malignancy by marking all regions of the image. A reasonable observer preferably marks lesions while minimizing marking other regions. Therefore, the idea of CAD’s initial detection stage is to find many of the malignancies as possible while not finding too many non-diseased regions. This corresponds to the search stage of the Kundel-Nodine model and the RSM. Unfortunately, CAD is rather poor at this task compared to expert radiologists. Progress in this area has been stymied by lack of understanding of search and how to measure performance in the FROC task. Indeed a widely held misconception is that CAD is perfect (!) at search, because it “looks at” everything (Dr. Ron Summers, NIH, private communication, Dublin, ca. 2010). In giving equal attention units to all parts of the image, CAD will trivially find all cancers, but it will also find a large number of NLs. Expert radiologists do not give equal attention units to all parts of the image. They are particularly good at giving more attention units to cancers than the surround, especially for the mass detection task, Fig. 15.3. Measuring search performance is addressed in Chapter 17.

CAD researchers are, in my opinion, at the forefront of those presuming to understand how radiologists interpret cases. They work with real images and real lesions and the manufacturer’s reputation is on the line, just like a radiologist’s, and Medicare even reimburses CAD interpretations. While their current track record is not that good for breast masses compared to expert radiologists, with proper understanding of what is limiting CAD, namely the search process, there is no doubt in my opinion, that future generations CAD algorithms will approach and even surpass expert radiologists.

## 29.10 Simultaneously acquired eye-tracking and FROC data

Studies of medical image interpretation have focused on either assessing radiologists' performance using, for example, the receiver operating characteristic (ROC) paradigm, or assessing the interpretive process by analyzing eye-tracking (ET) data. Analysis of ET data has not benefited from threshold-bias independent figures-of-merit (FOMs) analogous to the area under the ROC curve. In essence, research in this field is restricted to sensitivity/specificity analysis, and ignoring the benefits of accounting for their anti-correlation (recall the study by Beam et al that showed large decrease in inter-reader variability when AUC was used as a figure of merit instead of sensitivity or specificity, Fig. 3.6 and Table 3.3). A recent study<sup>13</sup> demonstrated the feasibility of such FOMs and measured agreement between figures-of-merit derived from free-response ROC (FROC) and ET data. A pre-publication copy, Analysis of simultaneously acquired ET-FROC data.pdf, is included in the online supplemental material. This section summarizes the salient points.

### 29.10.1 FROC and Eye-Tracking Data Collection

The data collection is shown schematically in Fig. 15.7. A head-mounted eye-position tracking system was worn that used an infrared beam to calculate line-of-gaze by monitoring the pupil and the corneal reflection. A magnetic head tracker was used to monitor head position, and this allows the radiologists to freely move their head. The eye-tracker integrates eye-position and head position to calculate the intersection of the line of gaze and the display plane.

The computer automatically captured the following information: i. The (x,y) location of marks made by the radiologists. Each mark was compared to the locations of the actual lesion and classified as FROC lesion localization (FROC-LL) if it fell within 2.5° (the proximity criterion) of visual angle (roughly 200 pixels). Otherwise, it was classified as FROC non-lesion localization (FROC-NL). ii. The confidence level (rating) for each mark. iii. Time-stamped, raw eye-position data collected during the entire time that the radiologists were examining the case. This data, which was acquired at 60 frames per second, included flags to indicate when image manipulation activities (such as marking, rating or window/level adjustments) and blinks occurred. The flagged data frames were excluded from analysis.

Eight expert breast radiologists interpreted a case set of 120 two-view mammograms while eye-tracking (ET) data and FROC data were continuously collected during the interpretation interval. Regions that attract prolonged ( $>800\text{ms}$ ) visual attention, using the algorithm in Appendix 15A, were considered to be eye-tracking marks. Based on the dwell and approach-rate (inverse of time-to-hit) eye-tracking ratings were assigned to each ET-mark. The ET- ratings were

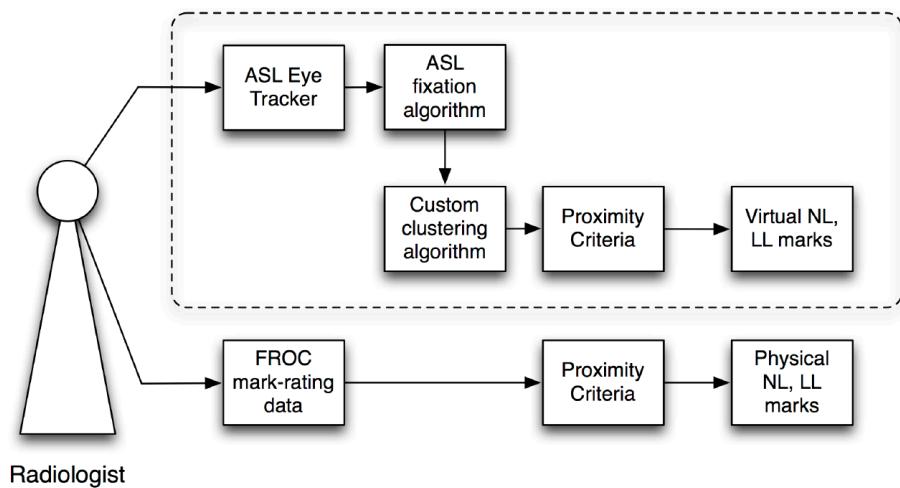


Figure 29.6: Schematic of the data collection and processing to obtain real and eye-tracking marks: the radiologists interpreted the images using a two-monitor workstation. Concurrently, and for the duration of the interpretation, an ASL eye-position tracking system determined the line-of-gaze. The ASL fixation and clustering algorithms are described in the text. The proximity criterion, defined as  $2.5^\circ$  of visual angle, is the maximum distance between a lesion center and a mark for the mark to be considered a LL (correct localization). Non-lesion localizations are all other marks. ASL = Applied Sciences Laboratory; NL = non-lesion localization; LL = lesion localization. Reproduced, with permission, from Ref. 13.

used to define threshold-bias independent FOMs in a manner analogous to the area under the trapezoidal alternative FROC (AFROC) curve (0 = worst, 1 = best). Agreement between ET FOM and FROC FOM was measured (0.5 = chance, 1 = perfect) using the jackknife and 95% confidence intervals (CI) for the FOMs and agreement were estimated using the bootstrap.

### 29.10.2 Measures of Visual Attention

At each big cluster location the following eye-position quantities were calculated:

- Dwell time (D): this was defined as the cumulative gaze in seconds of all fixations that comprised the big-cluster with total dwell exceeding 800 ms.
- Approach-rate (A): this was defined as the reciprocal (s-1) of shortest time-to-hit a big-cluster with total dwell exceeding 800 ms, i.e., approach the center of the big-cluster to within 2.5°. The reciprocal is taken to maintain a common directionality. In most cases greater perceptual attention is expected to be accompanied by greater approach-rate and larger values of dwell (the exception to this occurs for very large lesions, which “pop-out” from the surrounding background but do not need much cognitive processing to be resolved – in this case, dwell is not expected to be long). Dwell time has been linked to the amount of cognitive processing at a given location, and a dwell threshold has been proposed to separate the different types of errors<sup>5</sup>. Approach-rate can be thought of as a perceptual measure of how much a perceived area “pops-out” from the background, and it has been shown to be significantly related to the likelihood that a given breast cancer will be reported by radiologists<sup>22</sup>, with greater approach-rates being related to correct decisions<sup>23</sup>.

### 29.10.3 Generalized ratings

The eye tracking paradigm is conceptually similar to the FROC paradigm in the sense that both yield decisions at locations found by the observer. In effect, the big-clusters can be regarded as eye-tracking marks. In the FROC paradigm the observer marks regions that are considered sufficiently suspicious for presence of a lesion, and the degree of suspicion is recorded as a conscious rating. Analogously, eye-tracking yields the locations of regions that attracted visual attention long enough to allow a decision to be made at the location (the big-clusters), and for each region, there is a dwell time and an approach-rate. Dwell and approach-rate can be regarded as generalized (unconscious) ratings. Just as a figure-of-merit can be defined from FROC mark-rating data, likewise figures-of-merit can be defined from the eye tracking marks and generalized ratings. Details are in Appendix 15B, where three figures-of-merit are defined, and , where R stands for ratings, D for dwell and A for approach-rate and j is the reader index. These are analogous to the AFROC AUC (since the dataset contained only one lesion per diseased case, these are the same as the wAFROC AUC). The range of each figure-of-merit is from zero to unity.

A jackknife-based method for measuring individual case-level agreement between any pair of figures-of-merit is described in Appendix 15C. Defined there are , and which measure agreement between ratings and dwell, dwell and approach-rate and ratings and approach-rate, respectively. Each agreement measure ranges from 0.5 (chance level agreement) to one (perfect agreement). A bootstrap-based method for obtaining confidence intervals for figures-of-merit and agreements is described in Appendix 15D. The two-sided Wilcoxon signed rank test was used to measure the significance of differences between matched pairs of variables, one pair per reader, such as numbers of marks, ratings, figures-of-merit and agreements.

The AFROC mark-ratings FOM was largest 0.734, CI = (0.65, 0.81) followed by the dwell 0.460 (0.34, 0.59) and then by the approach-rate FOM 0.336 (0.25, 0.46). The differences between the FROC mark-ratings FOM and the perceptual FOMs were significant ( $p < 0.05$ ). All pairwise agreements were significantly better than chance: ratings vs. dwell 0.707 (0.63, 0.88), dwell vs. approach-rate 0.703 (0.60, 0.79) and rating vs. approach-rate 0.606 (0.53, 0.68). The agreement between ratings vs. approach-rate was significantly smaller than that between dwell vs. approach-rate ( $p = 0.008$ ).

This brief description shows how methods developed for analyzing observer performance data could be leveraged to complement current ways of analyzing ET data and lead to new insights.

## 29.11 Discussion / Summary

This chapter has introduced the terminology associated with a search task: recognition/finding, classification, and detection. Search involves finding lesions and correctly classifying them, so two types of expertise are relevant: search expertise is the ability to find (true) lesions without finding non-lesions, while classification accuracy is concerned with correct classification (benign vs. malignant) of a suspicious region that has already been found. Quantification of these abilities is described in the next chapter. Two paradigms are used to measure search, one in the non-medical context and the other, the focus of this book, in the medical context. The second method is based on the eye tracking measurements performed while radiologists perform quasi-clinical tasks (performing eye-tracking measurements in a true clinical setting is difficult). A method for analyzing eye-tracking data using methods developed for FROC analysis has been described. It has the advantage of taking into account information present in eye-tracking data, such as dwell time and approach rate, in a quantitative manner, essentially by treating them as eye-tracking ratings to which modern FROC methods can be applied. The Kundel-Nodine model of visual search in diagnostic imaging was described. The next chapter describes a statistical parameterization of this model, termed the radiological search model (RSM).

## 29.12 References

- Nodine CF, Kundel HL. Using eye movements to study visual search and to improve tumor detection. *RadioGraphics*. 1987;7(2):1241-1250.
2. Kundel HL, Nodine CF, Conant EF, Weinstein SP. Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study. *Radiology*. 2007;242(2):396-402.
3. Kundel HL, Nodine CF. Modeling visual search during mammogram viewing. *Proc SPIE*. 2004;5372:110-115.
4. Kundel HL, Nodine CF. A visual concept shapes image perception. *Radiology*. 1983;146:363-368.
5. Kundel HL, Nodine CF, Carmody D. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Invest Radiol*. 1978;13:175-181.
6. Horowitz TS, Wolfe JM. Visual search has no memory. *Nature*. 1998;394(6693):575-577.
7. Wolfe JM. Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*. 1994;1(2):202-238.
8. Wolfe JM, Cave KR, Franzel SL. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance*. 1989;15(3):419.
9. Carmody DP, Kundel HL, Nodine CF. Performance of a computer system for recording eye fixations using limbus reflection. *Behavior Research Methods & Instrumentation*. 1980;12(1):63-66.
10. Duchowski AT. Eye Tracking Methodology: Theory and Practice. Clemson, SC: Clemson University; 2002.
11. Nodine C, Mello-Thoms C, Kundel H, Weinstein S. Time course of perception and decision making during mammographic interpretation. *AJR*. 2002;179:917-923.
12. Nodine CF, Kundel HL, Mello-Thoms C, et al. How experience and training influence mammography expertise. *Acad Radiol*. 1999;6(10):575-585.
13. Chakraborty DP, Yoon H-J, Mello-Thoms C. Application of threshold-bias independent analysis to eye-tracking and FROC data. *Academic radiology*. 2012;19(12):1474-1483.
14. Burgess AE. Comparison of receiver operating characteristic and forced choice observer performance measurement methods. *Med Phys*. 1995;22(5):643-655.
15. Bunch PC, Hamilton JF, Sanderson GK, Simmons AH. A Free-Response Approach to the Measurement and Characterization of Radiographic-Observer Performance. *J of Appl Photogr Eng*. 1978;4:166-171.
16. Kundel HL, Nodine CF, Krupinski EA. Searching for lung nodules: visual dwell indicates locations of false-positive and false-negative decisions. *Investigative Radiology*. 1989;24:472-478.
17. Chakraborty DP, Yoon H-J, Mello-Thoms C. Application of threshold-bias independent analysis to eye-tracking and FROC data. *Academic Radiology*. 2012;In press.
18. Wolfe JM. Visual Search. In: Pashler H, ed. *Attention*. London, UK: University College London Press; 1998.
19. Larson AM, Loschky LC. The contributions of central versus peripheral vision to scene gist recognition. *Journal of Vision*. 2009;9(10):6-6.
20. Pritchard RM, Heron W, Hebb DO. Visual perception approached by the method of stabilized images. *Canadian Journal of Psychology/Revue canadienne de psychologie*. 1960;14(2):67.
21. Edwards DC, Kupinski MA, Metz CE, Nishikawa RM. Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. *Med Phys*. 2002;29(12):2861-2870.
22. Kundel HL, Nodine CF, Krupinski EA,

Mello-Thoms C. Using Gaze-tracking Data and Mixture Distribution Analysis to Support a Holistic Model for the Detection of Cancers on Mammograms. Academic Radiology. 2008;15(7):881-886. 23. Mello-Thoms C, Hardesty LA, Sumkin JH, et al. Effects of lesion conspicuity on visual search in mammogram reading. Acad Radiol. 2005;12:830-840.

# Chapter 30

## The radiological search model

### 30.1 TBA How much finished

10%

### 30.2 Introduction

Brief accounts of the radiological search model (RSM) were presented earlier in connection with the simulator used to generate FROC data. This chapter describes the statistical model in more detail. It embodies the essential ideas of the Nodine-Kundel model of visual search described in the previous chapter. *It turns out that all that is needed to model seemingly as complex a process as visual search, at least to first order, is one additional parameter.* All models of ROC data involve two parameters (not counting thresholds). For example, the unequal variance binormal model in Chapter 13 requires the (a,b) parameters. Alternative ROC models described in Chapter 20 also require two parameters. The model described below contains three parameters:  $\mu$ ,  $\lambda$  and  $\nu$ . The  $\mu$  parameter is the simplest to understand: it is the perceptual signal-to-noise ratio  $pSNR$  of latent LL ratings relative to latent NL ratings. The parameters  $\lambda$  and  $\nu$  describe the search process, i.e., the first stage of the Nodine-Kundel model (glancing or global impression). They describe the ability of the observer to find latent LLs while not finding latent NLs. It turns out that it is easier to understand the search process via intermediate primed parameters,  $\lambda'$  and  $\nu'$ ; however, unlike  $\lambda$  and  $\nu$  the primed parameters depend on  $\mu$ , i.e., *they are not intrinsic*. So in what follows I will introduce, in order,  $\mu$ ,  $\lambda'$  and  $\nu'$  and explain their meanings via software examples, as well as how one might measure

them via eye-tracking measurements. Finally, a model re-parameterization is proposed, which takes into account that  $\lambda'$  and  $\nu'$  must depend on  $\mu$ , and this is where the un-primed parameters  $\lambda$  and  $\nu$  are introduced, *which are expected to be intrinsic*, i.e., independent of  $\mu$ .

TBA [The online appendices explain Poisson and binomial sampling at a simple level. It is my experience that users of my software are generally not trained in statistics.]

### 30.3 The radiological search model

The Radiological Search Model (RSM) for the free-response paradigm is a statistical parameterization of the Nodine-Kundel model. It consists of:

- A *search stage* corresponding to the initial glance in the Nodine-Kundel sense, in which suspicious regions, i.e., the latent marks, are flagged for subsequent scanning. The total number of latent marks on a case is  $\geq 0$ , so some cases may have zero latent marks, a fact that will turn out to have important consequences for the shapes of all RSM predicted operating characteristics.
- A *decision stage* during which each latent mark is scanned, features are extracted and analyzed and the observer obtains a decision variable (i.e., a z-sample) at each latent mark. Typically radiologists spend  $\sim 1$  s per site and high-resolution foveal inspection is necessary to extract relevant details of the region being examined and make a decision whether or not to mark it. The number of realized z-samples equals the number of latent marks on the case.
- Latent marks can be either latent NLs (corresponding to non-diseased regions) or latent LLs (corresponding to diseased regions). The number of latent NLs on a case is denoted  $l_1$ . The number of latent LLs on a diseased case is denoted  $l_2$ . Latent NLs can occur on non-diseased and diseased cases, but latent LLs can only occur on diseased cases. Assume for now that every diseased case has  $L$  actual lesions. Later this is extended to arbitrary number of lesions per diseased case. Since the number of latent LLs cannot exceed the number of lesions,  $0 \leq l_2 \leq L$ . The symbol  $l_s$  denotes a location with site-level truth state  $s$ , where  $s = 1$  for a NL and  $s = 2$  for a LL<sup>1</sup>.

---

<sup>1</sup>In this chapter distributional assumptions are made for the numbers of latent NLs and LLs and for the associated z-samples. Since one is dealing with a parametric model one does not need to show explicitly case and location dependence as in the empirical description in Chapter @ref(#froc-empirical). This allows for a simpler notation, as the reader may have noticed, unencumbered by the plethora of subscripts in Table #ref(#froc-empirical-notaion).

## 30.4 RSM assumptions

The number of latent NLs,  $l_1 \geq 0$ , is an integer random variable sampled from the Poisson distribution with mean  $\lambda'$ :

$$l_1 \sim \text{Poi}(\lambda') \quad (30.1)$$

The probability mass function *pmf* of the Poisson distribution is defined by:

$$\text{pmf}_{\text{Poi}}(l_1, \lambda') = \exp(-\lambda') \frac{(\lambda')^{l_1}}{(l_1')!} \quad (30.2)$$

The number of latent LLs,  $l_2$ , where  $0 \leq l_2 \leq L$ , is an integer random variable sampled from the binomial distribution  $B$  with success probability  $\nu'$  and trial size  $L$ :

$$l_2 \sim \text{Bin}(L, \nu') \quad (30.3)$$

The *pmf* of the binomial distribution is defined by:

$$\text{pmf}_{\text{Bin}}(l_2, \nu') = \binom{L}{l_2} (\nu')^{l_2} (1 - \nu')^{L-l_2} \quad (30.4)$$

Each latent mark is associated with a z-sample. That for a latent NL is denoted  $z_{l_1}$  while that for a latent LL is denoted  $z_{l_2}$ . Latent NLs can occur on non-diseased and diseased cases while latent LLs can only occur on diseased cases.

1. For latent NLs, the z-samples are obtained by sampling  $N(0, 1)$ :

$$z_{l_1} \sim N(0, 1) \quad (30.5)$$

2. For latent LLs, the z-samples are obtained by sampling  $N(\mu, 1)$ :

$$z_{l_2} \sim N(\mu, 1) \quad (30.6)$$

3. In an FROC study with  $R$  ratings, the observer adopts  $R$  ordered cutoffs  $\zeta_r$ , where ( $r = 1, 2, \dots, R$ ). Defining  $\zeta_0 = -\infty$  and  $\zeta_{R+1} = \infty$ , then if  $\zeta_r \leq z_{l_s} < \zeta_{r+1}$  the corresponding latent site is marked and rated in bin  $r$ , and if  $z_{l_s} \leq \zeta_1$  the site is not marked.

4. The location of the mark is at the center of the latent site that exceeded a cutoff and an infinitely precise proximity criterion is adopted. Consequently, there is no confusing a mark made because of a latent LL z-sample exceeding the cutoff with one made because of a latent NL z-sample exceeding the cutoff, and vice-versa. Therefore, any mark made because of a latent NL z-sample that satisfies  $\zeta_r \leq z_{l_1} < \zeta_{r+1}$  will be scored as a non-lesion localization (NL) and rated  $r$ . Likewise, any mark made because of a latent LL z-sample that satisfies  $\zeta_r \leq z_{l_2} < \zeta_{r+1}$  will be scored as a lesion-localization (LL) and rated  $r$ .
5. In addition unmarked LLs (latent or not) are assigned the zero rating. By “latent or not” I mean that even lesions that were not flagged by the search stage, and therefore do not qualify as latent LLs, are assigned the zero rating. This is because they represent observable events.
6. By choosing  $R$  large enough, the above discrete rating model is applicable to continuous z-samples.

### 30.5 Summary of RSM

- First stage: initial glance, observer identifies latent NLs and latent LLs:
  - Number of NLs  $\sim$  Poisson with mean  $\lambda'$ ,
  - Number of LLs  $\sim$  binomial with success probability  $\nu'$  and trial size  $L$ .
- Second stage: detailed scrutiny, observer calculates z-sample at each latent mark:
  - z-sample for latent NL  $\sim N(0, 1)$ ,
  - z-sample for latent LL  $\sim N(\mu, 1)$ .
- Latent mark is actually marked if  $z \geq \zeta_1$ .
- The rating assigned to a mark is the index of the nearest threshold that was just equaled or exceeded by the z-sample.
- Unmarked latent NLs are unobservable events, but unmarked LLs, latent or not, are assigned the zero rating.

### 30.6 Physical interpretation of RSM parameters

The parameters  $\mu$ ,  $\lambda'$  and  $\nu'$  have the following meanings:

### 30.6.1 The $\mu$ parameter

The  $\mu$  parameter is the lesion contrast-to-noise-ratio, or more accurately, the perceptual signal to noise ratio  $pSNR$  introduced in TBA Chapter 12, between latent NLs and latent LLs. It is not the pSNR of the latent LL relative to its immediate surround. For structured backgrounds - as opposed to homogeneous backgrounds - pSNR is determined by the competition for latent marks from other regions, outside the immediate surround, that could be mistaken for lesions.

The  $\mu$  parameter is similar to detectability index  $d'$ , which is the separation parameter of two unit normal distributions required to achieve the observed probability of correct choice (PC) in a two alternative forced choice (2AFC) task between cued (i.e., pointed to by toggleable arrows) NLs and cued LLs. One measures the locations of the latent marks using eye-tracking apparatus TBA3 and clusters the data, then runs a 2AFC study as follows. Pairs of images are shown, each with a cued location, one a latent NL and the other a latent LL, where all locations were recorded in prior eye-tracking sessions for the specific radiologist. The radiologist's task is to pick the image with the latent LL. The probability correct PC in this task is related to the  $d'$  parameter by:

$$\mu = \sqrt{2}\Phi^{-1}(\text{PC}) \quad (30.7)$$

The radiologist on whom the eye-tracking measurements were performed and the one who performs the two alternative forced choice tasks must be the same, as two radiologists may not agree on latent NL marks. A complication in conducting such a study is that because of memory effects, a lesion can only be shown once; this could result in a limited number of comparisons and a consequential imprecise estimate of  $\mu$ .

### 30.6.2 The $\lambda'$ parameter

The  $\lambda'$  parameter determines the tendency of the observer to generate latent NLs. The mean number of latent NLs per case is an estimate of  $\lambda'$ . This can also be measured via eye-tracking apparatus. This time it is only necessary to cluster the marks and classify each mark as a latent NL or latent LL according to the adopted acceptance radius. An eye-tracking based estimate would be the total number of latent NLs in the dataset divided by the total number of cases.

Consider two observers, one with  $\lambda' = 1$  and the other with  $\lambda' = 2$ . While one cannot predict the exact number of latent NLs on any specific case, one can predict the average number of latent NLs on a given case set.

```
seed <- 1; set.seed(seed)
samples1 <- rpois(100, 1)
mean(samples1)
```

```

## [1] 1.01

samples1[1:10]

## [1] 0 1 1 2 0 2 3 1 1 0

seed <- 1; set.seed(seed)
samples2 <- rpois(100,2)
mean(samples2)

## [1] 2.02

samples2[1:10]

## [1] 1 1 2 4 1 4 4 2 2 0

```

In this example, the number of samples has been set to 100 (the first argument to `rpois()`).

- For the first observer,  $\lambda' = 1$  (the second argument to `rpois()`), the first case generated zero latent NLs, the 2nd and 3rd cases generated one NL each, the third case generated 2 NLs, etc.
- For the second observer, the first and second case generated one latent NL each, the third generated two, etc. While one cannot predict what will happen on any specific case, one can predict that the average or `mean()` number of latent NL marks per case for the 1st observer will be close to 1 (the observed values is 1.01) and that for the 2nd one will be close to 2 (the observed values is 2.02).

Estimates should be accompanied by confidence intervals. The following code illustrates Poisson sampling and estimation of an exact confidence interval for the mean for 100 samples from two Poisson distributions.

```

K <- 100
lambdaP <- c(1,2)
cat ("K = ", K, ", lambdaP 1st reader = ", lambdaP[1], ", lambdaP 2nd reader = ", lambdaP[2], ", lambdaP 3rd reader = ", lambdaP[3], ", lambdaP 4th reader = ", lambdaP[4], ", lambdaP 5th reader = ", lambdaP[5], ", lambdaP 6th reader = ", lambdaP[6], ", lambdaP 7th reader = ", lambdaP[7], ", lambdaP 8th reader = ", lambdaP[8], ", lambdaP 9th reader = ", lambdaP[9], ", lambdaP 10th reader = ", lambdaP[10]
## K = 100 , lambdaP 1st reader = 1 , lambdaP 2nd reader = 2

```

```

seed <- 1;set.seed(seed);samples1 <- rpois(K,lambda = lambdaP[1]);cat("obs. mean, reader 1 = ", m

## obs. mean, reader 1 = 1.01

seed <- 1;set.seed(seed);samples2 <- rpois(K,lambda = lambdaP[2]);cat("obs. mean, reader 2 = ", m

## obs. mean, reader 2 = 2.02

ret11 <- poisson.exact(sum(samples1),K)
ret21 <- poisson.exact(sum(samples2),K)

cat ("Rdr. 1: 95% CI = ", ret11$conf.int[1:2], "\n")

## Rdr. 1: 95% CI = 0.8226616 1.227242

cat ("Rdr. 2: 95% CI = ", ret21$conf.int[1:2], "\n")

## Rdr. 2: 95% CI = 1.751026 2.318599

```

For reader 1 the estimate of the Poisson parameter (the mean parameter of the Poisson distribution is frequently referred to as the Poisson parameter) is 1.01 with 95% confidence interval (0.823, 1.227); for reader 2 the corresponding estimates are 2.02 with 95% confidence interval (1.751, 2.319). As the number of cases increases, the confidence interval shrinks. For example, with 10000 cases, i.e., 100 times the value in the previous example:

```

K <- 10000
lambdaP <- c(1,2)
cat ("K = ", K, ", lambdaP 1st reader = ", lambdaP[1], ", lambdaP 2nd reader = ", lambdaP[2], "\n")

## K = 10000 , lambdaP 1st reader = 1 , lambdaP 2nd reader = 2

seed <- 1;set.seed(seed);samples1 <- rpois(K,lambda = lambdaP[1]);cat("obs. mean, reader 1 = ", m

## obs. mean, reader 1 = 1.0055

seed <- 1;set.seed(seed);samples2 <- rpois(K,lambda = lambdaP[2]);cat("obs. mean, reader 2 = ", m

## obs. mean, reader 2 = 2.006

```

```

ret12 <- poisson.exact(sum(samples1),K)
ret22 <- poisson.exact(sum(samples2),K)

cat ("Rdr. 1: 95% CI = ", ret12$conf.int[1:2], "\n")

## Rdr. 1: 95% CI = 0.9859414 1.025349

cat ("Rdr. 2: 95% CI = ", ret22$conf.int[1:2], "\n")

## Rdr. 2: 95% CI = 1.978335 2.033955

```

This time for reader 1, the estimate of the Poisson parameter is 1.01 with 95% confidence interval (0.986, 1.025); for reader 2 the corresponding estimate is 2.01 with 95% confidence interval (1.978, 2.034). The width of the confidence interval is inversely proportional to the square root of the number of cases (the example below is for reader 1):

```
ret11$conf.int[2] - ret11$conf.int[1]
```

```
## [1] 0.40458
```

```
ret12$conf.int[2] - ret12$conf.int[1]
```

```
## [1] 0.03940756
```

Since the number of cases was increased by a factor of 100, the width decreased by a factor of 10, the square-root of the ratio of the numbers of cases.

### 30.6.3 The $\nu'$ parameter

The  $\nu'$  parameter determines the ability of the observer to find lesions. Assuming the same number of lesions per diseased case, the mean fraction of latent LLs per diseased case is an estimate of  $\nu'$ . It too can be measured via eye-tracking apparatus performed on a radiologist. An eye-tracking based estimate would be the total number of latent LLs in the dataset divided by the total number of lesions. Consider two observers, one with  $\nu' = 0.5$  and the other with  $\nu' = 0.9$ . Again, while one cannot predict the precise number of latent LLs on any specific diseased case, or which specific lesions will be correctly localized, one can predict the average number of latent LLs. The code follows:

```

K2 <- 100;L <- 1;nuP1 <- 0.5;nuP2 <- 0.9;
cat ("K2 = ", K2,", nuP 1st reader = ", 0.5,", nuP 2nd reader = ", 0.9,"\\n")

## K2 = 100 , nuP 1st reader = 0.5 , nuP 2nd reader = 0.9

seed <- 1;set.seed(seed);samples1 <- rbinom(K2,L,nuP1);cat("mean, reader 1 = ", mean(samples1)/L,

## mean, reader 1 = 0.48

seed <- 1;set.seed(seed);samples2 <- rbinom(K2,L,nuP2);cat("mean, reader 2 = ", mean(samples2)/L,

## mean, reader 2 = 0.94

ret1 <- binom.exact(sum(samples1),K2*L)
ret2 <- binom.exact(sum(samples2),K2*L)

cat ("Rdr. 1: 95% CI = ", ret1$conf.int[1:2],"\\n")

## Rdr. 1: 95% CI = 0.3790055 0.5822102

cat ("Rdr. 2: 95% CI = ", ret2$conf.int[1:2],"\\n")

## Rdr. 2: 95% CI = 0.8739701 0.9776651

```

This code also uses 100 samples (K2). The result shows that for reader 1 the estimate of the binomial success rate parameter is 0.48 with 95% confidence interval (0.38, 0.58). For reader 2 the corresponding estimates are 0.94 with 95% confidence interval (0.87, 0.98). As the number of diseased cases increases, the confidence interval shrinks in inverse proportion to the square root of cases.

As a more complicated but clinically realistic example, consider a dataset with 100 cases in all where 97 have one lesion per case, two have two lesions per case and one has three lesions per case (these are typical lesion distributions observed in screening mammography). The code follows:

```

K2 <- c(97,2,1);Lk <- c(1,2,3);nuP1 <- 0.5;nuP2 <- 0.9;
samples1 <- array(dim = c(sum(K2),length(K2)))
cat("K2[1] =", K2[1],", K2[2] =", K2[2],", K2[3] =", K2[3], ", nuP1 =", nuP1, ", nuP2 =", nuP2, "

## K2[1] = 97 , K2[2] = 2 , K2[3] = 1 , nuP1 = 0.5 , nuP2 = 0.9

```

```

seed <- 1; set.seed(seed)
for (l in 1:length(K2)) {
  samples1[1:K2[1],1] <- rbinom(K2[1],Lk[1],nuP1)
}
cat("obsvd. mean, reader 1 = ", sum(samples1[!is.na(samples1)])/sum(K2*Lk), "\n")

## obsvd. mean, reader 1 =  0.4903846

samples2 <- array(dim = c(sum(K2),length(K2)))
seed <- 1; set.seed(seed)
for (l in 1:length(K2)) {
  samples2[1:K2[1],1] <- rbinom(K2[1],Lk[1],nuP2)
}
cat("obsvd. mean, reader 2 = ", sum(samples2[!is.na(samples2)])/sum(K2*Lk), "\n")

## obsvd. mean, reader 2 =  0.9326923

ret1 <- binom.exact(sum(samples1[!is.na(samples1)]),sum(K2*Lk))
ret2 <- binom.exact(sum(samples2[!is.na(samples2)]),sum(K2*Lk))

cat ("Rdr. 1: 95% CI = ", ret1$conf.int[1:2], "\n")

## Rdr. 1: 95% CI =  0.3910217 0.5903092

cat ("Rdr. 2: 95% CI = ", ret2$conf.int[1:2], "\n")

## Rdr. 2: 95% CI =  0.8662286 0.9725125

```

## 30.7 Model re-parameterization

While the parameters  $\mu$ ,  $\lambda'$  and  $\nu'$  are physically meaningful, and can be estimated from eye-tracking measurements, a little thought reveals that they cannot be varied independently of each other. Rather,  $\mu$  is an intrinsic parameter whose value, together with two other intrinsic parameters  $\lambda$  and  $\nu$ , determine the physically more meaningful parameters  $\lambda'$  and  $\nu'$ , respectively. The following is a convenient re-parameterization:

$$\nu' = 1 - \exp(-\mu\nu) \quad (30.8)$$

$$\lambda' = \frac{\lambda}{\mu} \quad (30.9)$$

The parameterization is not unique, but is relatively simple. The need for the first re-parameterization (involving  $\nu'$ ) was foreseen (using different notation) in the original search model TBA papers<sup>4,5</sup> but the need for the second re-parameterization (involving  $\lambda'$ ) was discovered more recently. Since it determines  $\nu'$ , the  $\nu$  parameter can be considered as the intrinsic (i.e.,  $\mu$ -independent) ability to find lesions; specifically, it is the rate of increase of  $\nu'$  with  $\mu$  at small  $\mu$ :

$$\nu' = \left( \frac{\partial \nu'}{\partial \mu} \right)_{\mu=0} \quad (30.10)$$

The dependence of  $\nu'$  on  $\mu$  is consistent with the fact that higher contrast lesions are easier to find. Any observer, even one without special expertise, can find a high contrast lesion. This is why  $\nu'$  is not an intrinsic property. Conversely, lower contrast lesions will be more difficult to find even by expert observers. The colloquial term *find* is used as shorthand for *flagged for further inspection by the holistic 1st stage of the search mechanism, thus qualifying as a latent site*. In other words, *finding* a lesion means the lesion was perceived as a suspicious region, which makes it a latent site, independent of whether or not the region was actually marked. Finding refers to the search stage. Marking refers to the decision stage, where the region's z-sample is determined and compared to a marking threshold.

According to Eqn. (30.8), as  $\mu \rightarrow \infty$ ,  $\nu' \rightarrow 1$ , and in the opposite limit as  $\mu \rightarrow 0$ ,  $\nu' \rightarrow 0$ . Recall the analogy to finding the sun made in TBA Chapter 12: objects with very high perceptual SNR are certain to be found and conversely, objects with zero perceptual SNR are found only by chance.

According to Eqn. (30.9) the value of  $\mu$  also determines  $\lambda'$ : as  $\mu \rightarrow \infty$ ,  $\lambda' \rightarrow 0$ , and conversely, as  $\mu \rightarrow 0$ ,  $\lambda' \rightarrow \infty$ . This too is clear from the sun analogy of TBA Chapter 12. Since the sun has very high contrast, there is no reason for the observer to find other suspicious regions, which have no possibility of resembling the sun. On the other hand, attempting to locate a faint star hidden by clouds is guaranteed to generate several latent NLs (because the expected small SNR from the faint real star is comparable to that from a number of regions in the background).

The re-parameterization used here is not unique, but is simple and has the right limiting behaviors.

## 30.8 Discussion / Summary

This chapter has described a statistical parameterization of the Nodine-Kundel model. The 3-parameter model of search in the context in the medical imaging accommodates key aspects of the process: search, the ability to find lesions while

minimizing finding non-lesions, is described by two parameters, specifically,  $\lambda'$  and  $\nu'$ . The ability to correctly mark a found lesion (while not marking found non-lesions) is characterized by the third parameter of the model,  $\mu$ . While the primed parameters have relatively simple physical meaning, they depend on  $\mu$ . Consequently, it is necessary to define them in terms of intrinsic parameters.

The next chapter explores the predictions of the radiological search model.

### 30.9 References

1. Chakraborty DP. Computer analysis of mammography phantom images (CAMPI): An application to the measurement of microcalcification image quality of directly acquired digital images. *Medical Physics*. 1997;24(8):1269-1277.
2. Chakraborty DP, Eckert MP. Quantitative versus subjective evaluation of mammography accreditation phantom images. *Medical Physics*. 1995;22(2):133-143.
3. Chakraborty DP, Yoon H-J, Mello-Thoms C. Application of threshold-bias independent analysis to eye-tracking and FROC data. *Academic Radiology*. 2012;In press.
4. Chakraborty DP. ROC Curves predicted by a model of visual search. *Phys Med Biol*. 2006;51:3463–3482.
5. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. *Phys Med Biol*. 2006;51:3449–3462.

# Chapter 31

## Radiological search model predictions

### 31.1 TBA How much finished

10%

### 31.2 Introduction

The preceding chapter described the radiological search model (RSM) for FROC data. This chapter describes predictions of the RSM and how they compare with evidence. The starting point is the inferred ROC curve. While mathematically rather complicated, the results are important because they are needed to derive the ROC-likelihood function, which is used to estimate RSM parameters from ROC data in TBA Chapter 19. The preceding sentence should lead the inquisitive reader to the question: *since the ROC paradigm ignores search, how is it possible to derive parameters of a model of search from the ROC curve?* The answer is that the *shape* of the ROC curve contains information about the RSM parameters. It is fundamentally different from predictions of all conventional ROC models: binormal (Dorfman and Alf, 1969), contaminated binormal model (Dorfman and Berbaum, 2000), bigamma (Dorfman et al., 1997) and proper ROC (Metz and Pan, 1999), namely it has a *constrained end-point property*, while all other models predict that the *end-point*, namely the uppermost non-trivial point on the ROC, reached at infinitely low reporting threshold, is (1,1), while the RSM predicts it does not reach (1,1). The nature of search is such that the limiting end-point is constrained to be below and to the left of (1,1). This key difference, allows one to estimate search parameters from ROC data.

Next, the RSM is used to predict FROC and AFROC curves. Two following sections show how search performance and lesion-classification performance can be quantified from the location of the ROC end-point. Search performance is the ability to find lesions while avoiding finding non-lesions, and lesion-classification performance is the ability, having found a suspicious region, to correctly classify it; if classified as a NL it would not be marked (in the mind of the observer every mark is a potential LL, albeit at different confidence levels). Note that lesion-classification is different from classification between diseased and non-diseased cases, which is measured by the ROC-AUC. Based on the ROC/FROC/AFROC curve predictions of the RSM, a comparison is presented between area measures that can be calculated from FROC data, and this leads to an important conclusion, namely the FROC curve is a poor descriptor of search performance and that the AFROC/wAFROC are preferred. This will come as a surprise (shock?) to most researchers somewhat familiar with this field, since the overwhelming majority of users of FROC methods, particularly in CAD, have relied on the FROC curve. Finally, evidence for the validity of the RSM is presented.

### 31.3 Inferred integer ROC ratings

Consider a  $R_{\text{FROC}} \geq 1$  rating FROC study with allowed ratings  $r = 1, 2, \dots, R_{\text{FROC}}$ . In Chapter TBA 13, the inferred-ROC rating of a case was defined as the rating of the highest rated mark on a case or  $-\infty$ , if the case has no marks. Since a  $-\infty$  rating is inconvenient notation at best and the ratings are ordered labels, the corresponding integer rating is defined to be ROC:1. No ordering information is lost provided every other rating is also “bumped up” by unity. Therefore, the integer inferred ROC scale extends from one to  $R_{\text{FROC}} + 1$ . Thus, a  $R_{\text{FROC}}$  rating FROC study formally corresponds to a  $R_{\text{FROC}} + 1$  rating ROC study.

Henceforth the word “inferred” will be implicit when referring to an RSM-predicted ROC curve.

In addition, instead of referring to FP and TP ratings, in this chapter it will be more convenient to use the symbol  $h_{k_t t}$  to denote the rating of the highest rated z-sample on case  $k_t t$  with truth state  $t$ . Thus  $h_{k_1 1}$  refers to the highest rating on a non-diseased case  $k_1 1$  and  $h_{k_2 2}$  refers to the highest rating on diseased case  $k_2 2$ . For non-diseased cases, the maximum is over all latent NLs on the case. For diseased cases, the maximum is over all latent NLs *and* latent LLs on the case.

Reiterating, the integer ROC rating is the one-incremented highest FROC rating of the case, or ROC:1 if the case has no marks.

As before, when there is a possibility of confusion, one precedes the rating with the applicable paradigm. Formally, consider a set of ordered thresholds  $\zeta_r < \zeta_{r+1}$  and dummy thresholds defined by  $\zeta_0 = -\infty, \zeta_{R_{\text{FROC}}+1} = \infty$ , then, if

$\zeta_r \leq h_{k_t t} < \zeta_{r+1}$  the case is rated  $r$ . As an example, if  $h_{k_t t} < \zeta_1$  the case is rated ROC:1.

### 31.3.1 Comments

- Since  $r = 1, 2, \dots, R_{\text{FROC}}$  the lowest allowed ROC rating on a case with at least one mark is ROC:2.
- On a case with no marks or the highest rated latent site satisfies  $h_{k_t t} < \zeta_1$  the observer gives the ROC:1 rating. From the analyst's point of view, one cannot distinguish between whether the ROC:1 rating was the result of the case not having any marks or the case had at least one latent site, but none of the z-samples exceeded  $\zeta_1$ .

A consequence of the possibility that some cases have no marks is that all RSM-predicted operating characteristics share a *constrained end-point property*, which is the next topic.

## 31.4 Constrained end-point property

The full range of ROC space, i.e.,  $0 \leq FPF(\zeta) \leq 1$  and  $0 \leq TPF(\zeta) \leq 1$ , is not continuously accessible to the observer. In fact,  $0 \leq FPF(\zeta) \leq FPF_{\max}$  and  $0 \leq TPF(\zeta) \leq TPF_{\max}$  where  $FPF_{\max}$  and  $TPF_{\max}$  are each generally less than (or, in special cases, equal to) unity. Therefore,  $(FPF_{\max}, TPF_{\max})$  represents a constraint on the end-point; the abscissa of the end-point has to be smaller than or equal to  $FPF_{\max}$  and the ordinate has to be smaller than or equal to  $TPF_{\max}$ .

*Starting from a finite value, as  $\zeta_1$  is lowered to  $-\infty$ , some of the previously ROC:1 rated cases that had at least one latent site will be marked and “bumped-up” to the ROC:2 bin, until eventually only cases with no marks remain in the ROC:1 bin: these cases will never be rated ROC:2. A rational observer who finds no suspicious regions, literally nothing to report, will assign the lowest available bin to the case, which happens to be ROC:1. The finite number of cases in the ROC:1 bin at infinitely low threshold has the consequence that the uppermost non-trivial continuously accessible operating point – that obtained by cumulating ratings ROC:2 and above, is below and to the left of (1,1). The (1,1) point is reached “trivially” when the researcher cumulates the counts in all bins, i.e., ROC:1 and above. This behavior is distinct from traditional ROC models where the entire curve extending from (0, 0) to (1, 1) is continuously accessible to the observer. This is because in conventional models every case yields a finite decision variable, no matter how small. Lowering the lowest threshold to  $-\infty$  eventually moves all cases in the previously ROC:1 bin to the ROC:2 bin, and one is eventually left with zero counts in the ROC:1 bin, and the operating point, obtained by cumulating bins ROC:2 and above, is (1,1).*

In the RSM, starting with an infinitely high threshold, as the observer is encouraged to be more “aggressive in reporting lesions”, the ROC point moves continuously upwards and to the right from  $(0, 0)$  to  $(FPP_{max}, TPF_{max})$  and no further. The ROC curve cannot just “hang there” since cumulating all cases yields the “trivial”  $(1,1)$  operating point. Therefore, the complete ROC curve is obtained by extending the end-point using a dashed line that connects it to  $(1,1)$ . The observer cannot operate along the dashed line. See further elaboration of this point, in particular why guessing to operate along the dashed line, is not an option, in TBA §17.12.1.

How closely the observer approaches the limiting point  $(FPP_{max}, TPF_{max})$  is unrelated to the number of bins; rather, it depends on the position of the lowest threshold, i.e.,  $\zeta_1$ . As the latter is lowered the observed end-point approaches  $(FPP_{max}, TPF_{max})$  from below. How closely  $(FPP_{max}, TPF_{max})$  approaches  $(1,1)$  depends on  $\lambda'$  and  $\nu'$ . As  $\lambda'$  and  $\nu'$  increase, the limiting point approaches  $(1,1)$  from below, see TBA Eqn. (17.1) and Eqn. TBA (17.2). These parameters determine the probability that a case has one or more marks, and depending on the truth-state of case, non-diseased or diseased, these probabilities equal  $FPP_{max}$  or  $TPF_{max}$ , respectively, both of which increase to unity as  $\lambda'$  and  $\nu'$  increase.

### 31.4.1 The abscissa of the ROC end-point

One needs the probability that a non-diseased case has at least one latent NL. Such a case will generate a finite value of  $h_{k_1 1}$  and with an appropriately low  $\zeta_1$  the case will be rated ROC:2 or higher. The probability of zero latent NLs, see TBA Eqn. 16.2, is:

$$\text{pmf}_{Poi}(0, \lambda')$$

Therefore the probability that the case has at least one latent NL, which is the maximum continuously accessible abscissa of the ROC, is:

$$FPP_{max} = 1 - \exp(-\lambda') = 1 - \exp\left(\frac{-\lambda}{\mu}\right) \quad (31.1)$$

The second form on the right hand side of Eqn. (31.1) expresses the result in terms of *intrinsic* RSM parameters (the  $\mu$  independent  $\lambda, \nu$  parameters; not the physical ones, see section on model re-parameterization in TBA §16.4). As  $\mu$  increases  $FPP_{max}$  moves to the left, reaching zero in the limit  $\mu = \infty$ . Recall the by now familiar “solar” analogy in TBA Chapter 12. For fixed  $\mu > 0$  increasing  $\lambda$  causes  $FPP_{max}$  to move to the right approaching one in the limit  $\lambda = \infty$ , because in this limit every case will have a latent NL.

### 31.4.2 The ordinate of the ROC end-point

A diseased case has no marks, even for very low  $\zeta_1$ , if it has zero latent NLs, the probability of which is  $\exp(-\lambda')$ , and it has zero latent LLs, the probability of which is, TBA Eqn. 16.4,

$$\text{pmf}_{Bin}(0, L, \nu')$$

Here  $L$  is the number of lesions in each diseased case, assumed constant.

- Assumption 1: occurrences of latent LLs are independent of each other, i.e. the probability that a lesion is found is independent of whether other lesions were found on the same case.
- Assumption 2: occurrences of latent NLs are independent of each other; i.e., the probability that a non-diseased region is found is independent of whether other non-diseased regions were found on the same case.
- Assumption 3: occurrence of a latent NL is independent of the occurrence of a latent LL on the same case.

By the independence assumptions, the probability of zero latent NLs and zero latent LLs on a diseased case is the product of the two probabilities, namely

$$\exp(-\lambda')(1 - \nu')^L$$

Therefore, the probability that there exists at least one latent site is:

$$\left. \begin{aligned} \text{TPF}_{max} &= 1 - \exp(-\lambda')(1 - \nu')^L \\ &= 1 - \exp\left(-\frac{\lambda}{\mu}\right) \exp(-\nu\mu L) \end{aligned} \right\} \quad (31.2)$$

The second expression on the right hand side, in terms of intrinsic parameters, follows from Eqn. 16.8 and Eqn. 16.9. As  $\lambda \rightarrow \infty$ ,  $(\text{FPF}_{max}, \text{TPF}_{max})$  approaches  $(1,1)$ , because in this limit every case is assured to have a latent NL, thereby yielding a finite z-sample, and will be marked at sufficiently low  $\zeta_1$ . Conversely, as  $\lambda \rightarrow 0$ ,  $(\text{FPF}_{max}, \text{TPF}_{max})$  approaches:

$$(0, 1 - \exp(-\nu\mu L))$$

as in this limit there are no cases with latent NLs, so all non-diseased cases are unmarked and lesions are marked to the extent determined by the product

$\nu\mu L$ . As this product increases,  $TPF_{max}$  approaches unity, in other words, in the simultaneous limits  $\lambda \rightarrow 0$  and  $\mu\nu \rightarrow \infty$ , the ROC plot approaches perfect performance, i.e., a vertical line from the origin to  $(0,1)$  - the continuously accessible section - followed by the continuously inaccessible horizontal line connecting  $(0,1)$  to  $(1,1)$ . Since  $\nu$  is positive,  $\mu\nu \rightarrow \infty$  is equivalent to  $\mu \rightarrow \infty$ .

### 31.4.3 Variable number of lesions per case

Define  $f_L$  as the fraction of diseased cases with  $L$  lesions, and  $L_{max}$  the maximum number of lesions per diseased case in the dataset, then:

$$\sum_{L=1}^{L_{max}} f_L = 1 \quad (31.3)$$

By restricting attention to the set of diseased cases with  $L$  lesions each, Eqn. (31.2) for  $TPF_{max}$  applies, and since TPF is a probability, and probabilities of independent processes add, it follows that

$$TPF_{max}(\mu, \lambda', \nu', \vec{f}_L) = \sum_{L=1}^{L_{max}} TPF_{max}(\mu, \lambda', \nu', f_L) \quad (31.4)$$

In other words, the ordinate of the uppermost point is a weighted sum over the lesion distribution. It is seen that the ordinate is less than or equal to unity (as each term  $TPF_{max}(\mu, \lambda', \nu', f_L)$  in the weighted summation is less than or equal to unity). The expression for  $FTP_{max}$  is unaffected.

## 31.5 The RSM-predicted ROC curve

To predict the continuous ROC curve, one dispenses with binning and assumes the observer indicates the actual value of the z-sample (or some fixed monotonic increasing function of it) for each latent site. The ROC decision variable is the rating of the highest rated mark  $h_{k_t t}$  for the case. Since one is on the continuous section of the curve, each case must have at least one site and one does not have to worry about cases with no marks. Therefore, false positive fraction (FPF) is the probability that  $h_{k_t t}$  on a non-diseased case exceeds the virtual threshold  $\zeta$  and true positive fraction (TPF) is the probability that  $h_{k_t t}$  on a diseased case exceeds  $\zeta$ :

$$\begin{aligned} FPF(\zeta) &= P(h_{k_1 t} \geq \zeta) \\ TPF(\zeta) &= P(h_{k_2 t} \geq \zeta) \end{aligned} \quad (31.5)$$

Varying the threshold parameter  $\zeta$  from  $\infty$  to  $-\infty$  sweeps out the continuous section of the predicted RSM-predicted ROC curve, extending from  $(0,0)$  to  $(FPF_{max}, TPF_{max})$ .

### 31.5.1 Derivation of FPF

- Assumption 4: the z-samples of NLs on the same case are independent of each other.

Consider the set of non-diseased cases with  $n$  latent NLs each, where  $n > 0$ . According to the RSM, each latent NL yields a z sample from  $N(0, 1)$ . The probability that a z-sample from a latent NL is smaller than  $\zeta$  is  $\Phi(\zeta)$ . By the independence assumption the probability that all  $n$  samples are smaller than  $\zeta$  is  $(\Phi(\zeta))^n$ . If all z-samples are smaller than  $\zeta$ , then the highest z-sample  $h_{k_1 1}$  must be smaller than  $\zeta$ . Therefore, the probability that  $h_{k_1 1}$  exceeds  $\zeta$  is:

$$\left. \begin{aligned} FPF(\zeta | n) &= P(h_{k_1 1} > \zeta | n) \\ &= 1 - [\Phi(\zeta)]^n \end{aligned} \right\} \quad (31.6)$$

The conditioning notation in Eqn. (31.6) reflects the fact that this expression applies specifically to non-diseased cases with  $n$  latent NLs. To obtain  $(FPF_{max})$  one performs a Poisson-weighted summation of  $FPF(\zeta | n)$  over  $n$  from 1 to  $\infty$  (in other words, one computes the expectation of  $FPF(\zeta | n)$  by averaging over the Poisson distribution of the random variable  $n$  from zero to infinity, as the zero term makes a zero contribution to the above equation):

$$FPF(\zeta) = \sum_{n=0}^{\infty} pmf_{Poi}(n, \lambda') FPF(\zeta | n) \quad (31.7)$$

The infinite summations, see below, are easier performed using symbolic algebra software such as Maple<sup>TM</sup>. Inclusion, in the summation, of  $n = 0$ , which term evaluates to zero, is done to make it easier for Maple to evaluate the summation in closed form. Otherwise one would need to simplify the Maple-generated result. The result is shown below (Maple 17, Waterloo Maple Inc.), where `lambda` and `nu` refer to the primed quantities.

```
restart; phi := proc (t, mu) exp(-(1/2)(t-mu)^2)/sqrt(2Pi) end: PHI := proc (c, mu) local t; int(phi(t, mu), t = -infinity .. c) end: l) end: FPF := proc(zeta,lambda) sum(Poisson(n,lambda)(1 - PHI(zeta,0)^n), n=0..infinity);end: FPF(zeta, lambda);
```

The Maple code yields the following result (the second line uses the physical (primed) to intrinsic transformation):

$$\left. \begin{aligned} FPF(\zeta, \lambda') &= 1 - \exp\left(-\frac{\lambda'}{2}\left[1 - \operatorname{erf}\left(\frac{\zeta}{\sqrt{2}}\right)\right]\right) \\ FPF(\zeta, \lambda) &= 1 - \exp\left(-\frac{\lambda}{2\mu}\left[1 - \operatorname{erf}\left(\frac{\zeta}{\sqrt{2}}\right)\right]\right) \end{aligned} \right\} \quad (31.8)$$

The error function in Eqn. (31.8) is defined by:

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt \quad (31.9)$$

It is related, TBA Online Appendix 17.A, to the normal CDF function  $\Phi(x)$  by:

$$\operatorname{erf}(x) = 2\Phi(\sqrt{2}x) - 1 \quad (31.10)$$

The error function ranges from -1 to +1 as its argument ranges from  $\infty$  to  $-\infty$ . For  $\zeta = -\infty$ , Eqn. (31.8) yields the same expression for  $(FPF_{max}$  as does Eqn. (31.1):

$$1 - \exp\left(-\frac{\lambda}{2\mu}\left[1 - \operatorname{erf}\left(\frac{\zeta \rightarrow -\infty}{\sqrt{2}}\right)\right]\right) = 1 - \exp\left(-\frac{\lambda}{2\mu}[1 + 1]\right) = 1 - \exp\left(-\frac{\lambda}{\mu}\right)$$

Therefore, FPF ranges from zero to  $(FPF_{max}$  as  $\zeta$  ranges from  $\infty$  to  $-\infty$ , showing once again the constrained property of the abscissa of the predicted end-point.

### 31.5.2 Derivation of TPF

The derivation of the true positive fraction  $FPF(\zeta)$  follows a similar line of reasoning except this time one needs to consider the highest of the latent NLs and latent LL z-samples. Consider a diseased case with  $n$  latent NLs and  $l$  latent LLs. Each latent NL yields a decision variable sample from  $N(0, 1)$  and each latent LL yields a sample from  $N(\mu, 1)$ . The probability that all  $n$  latent NLs have z-samples less than  $\zeta$  is  $[\Phi(\zeta)]^n$ . The probability that all  $l$  latent LLs have z-samples less than  $\zeta$  is  $[\Phi(\zeta - \mu)]^l$ . Using the independence assumptions, the probability that all latent marks have z-samples less than  $\zeta$  is the product of these two probabilities. The probability that  $h_{k_2 2}$  (the highest z-sample on diseased case  $k_2 2$ ) is larger than  $\zeta$  is the complement of the product probabilities, i.e.,

$$\operatorname{TPF}_{n,l}(\zeta, \mu, n, l) = P(h_{k_2 2} > \zeta | \mu, n, l) = 1 - [\Phi(\zeta)]^n [\Phi(\zeta - \mu)]^l$$

One averages over the distributions of  $n$  and  $l$  to obtain the desired ROC-ordinate,

$$\text{TPF}(\zeta, \mu, \lambda', \nu') = \sum_{n=0}^{\infty} \text{pmf}_{Poi}(n, \lambda') \sum_{l=0}^L \text{Bin}_{pmf}(l, \nu', L) \text{TPF}_{n,l}(\zeta, \mu, n, l)$$

This can be evaluated using Maple, TBA Online Appendix 17.B, yielding:

$$\text{TPF}(\zeta, \mu, \lambda', \nu', L) = 1 - \left( 1 - \frac{\nu'}{2} + \frac{\nu'}{2} \text{erf}\left(\frac{\zeta - \mu}{\sqrt{2}}\right) \right)^L \exp\left(-\frac{\lambda'}{2} + \frac{\lambda'}{2} \left( \text{erf}\left(\frac{\zeta}{\sqrt{2}}\right) \right)\right) \quad (31.11)$$

It can be confirmed that for  $\zeta = -\infty$  Eqn. (31.11) yields the same expression for  $\text{TPF}_{max}$  as Eqn. (31.2):

$$\text{TPF}(-\infty, \mu, \lambda', \nu', L) = 1 - (1 - \nu')^L \exp(-\lambda') = 1 - \exp\left(-\frac{\lambda}{\mu}\right) \exp(-\mu\nu L)$$

### 31.5.3 Extension to varying numbers of lesions

To extend the results to varying numbers of lesions per diseased case, one averages the right hand side of Eqn. (17.14) over the fraction of diseased cases with  $L$  lesions:

$$\text{TPF}(\zeta, \mu, \lambda', \nu', \overrightarrow{f_L}) = \sum_{L=1}^{L_{max}} f_L \left( 1 - \left( 1 - \frac{\nu'}{2} + \frac{\nu'}{2} \text{erf}\left(\frac{\zeta - \mu}{\sqrt{2}}\right) \right)^L \exp\left(-\frac{\lambda'}{2} + \frac{\lambda'}{2} \left( \text{erf}\left(\frac{\zeta}{\sqrt{2}}\right) \right)\right) \right) \quad (31.12)$$

The right hand side can be expressed in terms of intrinsic parameters, but the resulting expression is cumbersome and is not shown. The expression for FPF is, of course, unaffected.

### 31.5.4 “Proper” property of the RSM-predicted ROC curve

A “proper” ROC curve has the property that it never crosses the chance line and its slope decreases monotonically as the operating point moves up the ROC curve TBA10. It is shown next that the continuously accessible portion of the ROC curve is “proper”. For convenience one abbreviates FPF and TPF to  $x$

and  $y$ , respectively, and suppresses the dependence on model parameters. From Eqn. (31.8) and Eqn. (31.11) one can express the ROC coordinates as:

$$\left. \begin{aligned} x(\zeta) &= 1 - G(\zeta) \\ y(\zeta) &= 1 - F(\zeta)G(\zeta) \end{aligned} \right\} \quad (31.13)$$

where:

$$\left. \begin{aligned} G(\zeta) &= \exp\left(-\frac{\lambda'}{2} + \frac{\lambda'}{2}\operatorname{erf}\left(\frac{\zeta}{\sqrt{2}}\right)\right) \\ F(\zeta) &= 1 - \left(1 - \frac{\nu'}{2} + \frac{\nu'}{2}\operatorname{erf}\left(\frac{\zeta-\mu}{\sqrt{2}}\right)\right)^L \end{aligned} \right\} \quad (31.14)$$

These equations have exactly the same structure as Swensson's TBA 11 Eqns. 1 and 2 and the logic he used to demonstrate that ROC curves predicted by his model were "proper" also applies to the present situation. Specifically, since the error function ranges between -1 and 1 and  $0 \leq \nu' \leq 1$ , it follows that  $F(\zeta) \leq 1$ . Therefore  $y(\zeta) \geq x(\zeta)$  and the ROC curve is constrained to the upper half of the ROC space, namely the portion above the chance diagonal. Additionally the more general constraint shown by Swensson applies, namely the slope of the ROC curve at any operating point  $(x, y)$  cannot be less than the slope of the straight line connecting  $(x, y)$  and  $x_{max}, y_{max}$ , the coordinates of the predicted end-point. This implies that the slope decreases monotonically and also rules out curves with "hooks".

The results in this section are also valid for arbitrary numbers of lesions per case. Proving this is left as an exercise for the reader.

*It is important to note that the proper ROC prediction applies to the continuous section of the ROC only. It is possible for the slope to change abruptly, even increasing, as one crosses over from the continuous section to the inaccessible part of the plot. Recall that the inaccessible part corresponds to cases that do not provide z-samples, and are therefore not subject to the decision rule that generates the continuous section of the plot.*

### 31.5.5 The pdfs for the ROC decision variable

In TBA Chapter 06 the pdf functions for non-diseased and diseased cases for the unequal variance binormal ROC model1 were derived. The procedure was to take the derivative of the appropriate cumulative distribution function (CDF) with respect to  $\zeta$ . An identical procedure is used for the RSM. The CDF for non-diseased cases is the complement of FPF, Eqn. (31.8). The pdf corresponding to non-diseased cases is given by:

$$\text{pdf}_N(\zeta) = \frac{\partial}{\partial \zeta} \exp \left( -\frac{\lambda'}{2} \left[ 1 - \text{erf} \left( \frac{\zeta}{\sqrt{2}} \right) \right] \right) \quad (31.15)$$

Using Maple, this evaluates to:

$$\text{pdf}_N(\zeta) = \frac{\lambda' \exp(-\frac{1}{2}\zeta^2) \exp(-\frac{-\lambda'}{2} [1 - \text{erf}(\frac{\zeta}{\sqrt{2}})])}{\sqrt{2\pi}} \quad (31.16)$$

Similarly, for the diseased cases,

$$\text{pdf}_D(\zeta) = \frac{\partial}{\partial \zeta} \left( 1 - \frac{\nu'}{2} + \frac{\nu'}{2} \text{erf} \left( \frac{\zeta - \mu}{\sqrt{2}} \right) \right)^L \exp \left( -\frac{\lambda'}{2} + \frac{\lambda'}{2} \left( \text{erf} \left( \frac{\zeta}{\sqrt{2}} \right) \right) \right) \quad (31.17)$$

Maple does evaluate the derivative, but it is cumbersome to display. The formulas for the RSM-predicted ROC and the pdfs are coded in `RJafroc` in the embedded function `PlotRsmOperatingCharacteristics()`.

It is seen that the integrals of the pdfs are given by (non-diseased followed by diseased):

$$\begin{aligned} \int_{-\infty}^{\infty} \text{pdf}_N(\zeta) d\zeta &= \exp \left( -\frac{\lambda'}{2} \left[ 1 - \text{erf} \left( \frac{\zeta}{\sqrt{2}} \right) \right] \right) \Big|_{-\infty}^{\infty} \\ &= 1 - \exp(-\lambda') \\ &= \text{FPF}_{max} \end{aligned} \quad (31.18)$$

$$\begin{aligned} \int_{-\infty}^{\infty} \text{pdf}_D(\zeta) d\zeta &= \left( 1 - \frac{\nu'}{2} + \frac{\nu'}{2} \text{erf} \left( \frac{\zeta - \mu}{\sqrt{2}} \right) \right)^L \exp \left( -\frac{\lambda'}{2} + \frac{\lambda'}{2} \left( \text{erf} \left( \frac{\zeta}{\sqrt{2}} \right) \right) \right) \Big|_{-\infty}^{\infty} \\ &= 1 - \exp(-\lambda') (1 - \nu')^L \\ &= \text{TPF}_{max} \end{aligned} \quad (31.19)$$

In other words, they evaluate to the coordinates of the predicted end-point, *each of which is less than unity*. The reason is that the integration is along the *continuous* section of the ROC curve and does not include the contribution from the area under the straight line extension from  $(\text{FPF}_{max}, \text{TPF}_{max})$  to  $(1,1)$ . This contribution is the probability of no marks,  $1 - \text{FPF}_{max}$  for non-diseased cases and  $1 - \text{TPF}_{max}$  for diseased cases. Adding these contributions to the integral under the continuous section yields unity for both types of cases. I am aware that pdfs that do not integrate to unity present conceptual problems and the original RSM TBA publications 12,13 unnecessarily introduced Dirac delta functions to force the integrals to be unity. I trust the explanation given here clarifies the issue.

### 31.5.6 RSM-predicted ROC-AUC and AFROC-AUC

It is possible to numerically perform the integration under the RSM-ROC curve to get RSM-ROC-AUC,  $AUC_{RSM}^{ROC}(\mu, \lambda, \nu, \bar{f}_L)$ :

$$AUC_{RSM}^{ROC}(\mu, \lambda, \nu, \bar{f}_L) = \sum_{L=0}^{L_{max}} f_L \int_0^1 TPF(\mu, \lambda, \nu, f_L) d(FPF(\zeta, \lambda)) \quad (31.20)$$

Since the RSM predicts other operating characteristics, the superscript *ROC* is needed to keep track of it.

The right hand side of Eqn. (31.20) can be evaluated using a numerical integration function implemented in R, which is used in the **RJafroc** function **UtilAnalyticalAucsRSM()** whose help page follows:

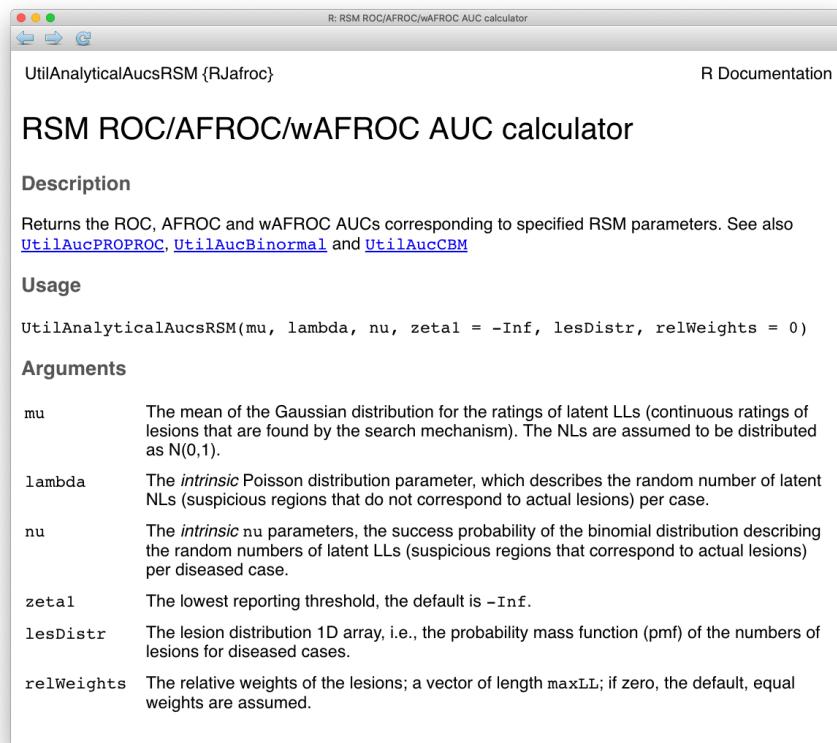


Figure 31.1: Help file for function UtilAnalyticalAucsRSM.

The arguments to the function are the intrinsic RSM parameters  $\mu$ ,  $\lambda$  and  $\nu$ . The

default value of  $\zeta_1$  is used, namely  $-\infty$ .<sup>1</sup> The function also needs to know the lesion distribution `lesDistr`. In the following it is supplied as the vector `c(0.5, 0.3, 0.2)`, meaning fraction 0.5 of diseased cases have one lesion, fraction 0.3 have two lesions and fraction 0.2 have three lesions. Finally, the function needs to know the lesion weights. In the following the default value, zero, is used, which gives equal weights to all lesions. The function returns a list containing the AUCs under the ROC, the AFROC and the weighted AFROC. TBA!! With equal weights the AFROC and the wAFROC AUCs are identical.

```

mu <- 1; lambda <- 1; nu <- 1
lesDistr <- c(0.5, 0.3, 0.2)
aucs <- UtilAnalyticalAucsRSM(mu = mu, lambda = lambda, nu = nu, lesDistr = lesDistr, relWeights
cat("mu = ", mu,
", lambda = ", lambda,
", nu = ", nu,
", AUC/ROC = ", aucs$aucROC,
", AUC/AFROC = ", aucs$aucAFROC, "\n")

## mu = 1 , lambda = 1 , nu = 1 , AUC/ROC = 0.7802109 , AUC/AFROC = 0.577889

```

This code was used with varying RSM parameters to generate the following Table.

---

<sup>1</sup>The AUCs depend on the value of  $\zeta_1$ ; this is used in `RJafrocBook` to optimize the performance of a CAD algorithm, i.e., determine the optimal reporting threshold.

Table 31.1: The last two columns list the RSM-predicted AUCs under the ROC and AFROC, respectively. The corresponding RSM parameters are listed in the first four columns.

$\mu$	$\lambda$	$\nu$	$L_{max}$	AUC-ROC	AUC-AFROC
1				0.7148	0.5779
2	1			0.9027	0.874
3				0.97	0.9628
	0.5	1		0.7551	0.676
	0.1		1	0.8016	0.7838
		2		0.7939	0.7228
		0.5		0.6337	0.4292
1		0.25		0.5752	0.3218
	1		2	0.8224	0.5779
		1		0.8803	0.5779
3			3	0.9995	0.9628

Examination of Table 31.1 reveals the following points.

- Both AUCs are increasing functions of  $\mu$ . Increasing perceptual signal-to-noise-ratio (pSNR) always leads to improved performance: for background on this dependence the reader is referred to the “solar analogy” in TBA Chapter 12. Increasing  $\mu$  increases the separation between the two pdfs defining the ROC curve, which increases ROC AUC. Furthermore, the number of NLs decreases because  $\lambda' = \lambda/\mu$  decreases, which increases performance as fewer FPs are generated. Finally,  $\nu'$  increases and approaches unity, which leads to more LL events and increased performance. Because all three effects reinforce each other, a change in  $\mu$  results in a large effect on performance.
- Both AUCs are increasing functions of decreasing  $\lambda$ . This is because decreasing  $\lambda$  results in fewer latent NLs per case and thereby increases performance as fewer NLs are marked. This is a relatively weak effect.

- Both AUCs are increasing functions of  $\nu$ . Increasing  $\nu$  results in more LLs being marked, which increases performance. This is a relatively strong effect.
- ROC-AUC increases with  $L_{max}$ , since with more lesions per case, there is increased probability that at least one of them will be found, i.e., will be a latent LL and therefore more likely to be marked, and the diseased distribution pdf moves to the right. However, AFROC-AUC is independent of  $L_{max}$ , because the y-axis is LLF, which is independent of the number of lesions in the dataset or their distribution.
- ROC-AUC values are constrained to the range 0.5 to 1 while the AFROC-AUC values are constrained to the larger range 0 to 1. The reader should confirm that the difference in ROC-AUC between any two rows in Table 31.1 is smaller than the corresponding difference in AFROC-AUC: the AFROC effect-size is always larger than the corresponding ROC effect-size.

### 31.5.7 RSM-predicted ROC and pdf curves

Fig. 31.2 displays ROC curves for indicated values of  $\mu$ . The remaining intrinsic RSM model parameters are  $\lambda = 1$ ,  $\nu = 1$ ,  $\zeta_1 = -\infty$  and one lesion per diseased case.

The following are evident from these figures:

1. As  $\mu$  increases the ROC curve more closely approaches the upper-left corner of the ROC plot. This signifies increasing performance and the area under the ROC and AFROC curves approach unity, which is the best possible performance. The end-point abscissa decreases, denoting greater numbers of unmarked non-diseased cases, i.e., more good decisions on non-diseased cases. The end-point ordinate increases, denoting smaller numbers of unmarked lesions, i.e., more good decisions on diseased cases.
2. For  $\mu$  close to zero the operating characteristic approaches the chance diagonal and the area under the ROC curve approaches 0.5, which is the worst possible ROC performance.
3. The area under the ROC increases monotonically from 0.5 to 1 as  $\mu$  increases from zero to infinity.
4. For large  $\mu$  the accessible portion of the operating characteristic approaches the vertical line connecting (0,0) to (0,1), the area under which is zero. The complete ROC curve is obtained by connecting this point to (1,1) by the dashed line and in this limit the area under the complete ROC curve approaches unity. Omitting the area under the dashed portion of the curve will result in a severe underestimate of true performance.
5. As  $\mu$  is increased (allowed values are 1, 2, 3, etc.) the area under the ROC curve increases, approaching unity and approaches unity while remains

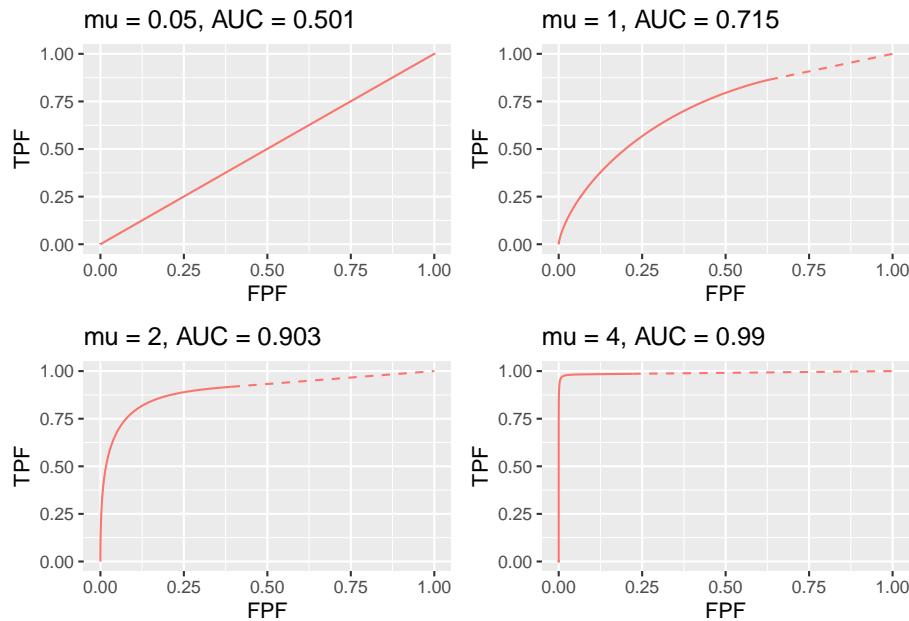


Figure 31.2: RSM-predicted ROC curves for indicated values of the  $\mu$  parameter. The solid curve is the continuous section and the dashed part is the inaccessible part. Notice the transition, as  $\mu$  increases, from near chance level performance to almost perfect performance, and the end-point moves from near  $(1,1)$  to near  $(0,1)$ . The area under the ROC curve includes that under the red dashed line, which credits unmarked non-diseased cases. If this area is not included, a severe underestimate of performance can occur, especially for large  $\mu$ .

constant at a value determined by , Eqn. (17.1). With more lesions per diseased case, the chances are higher that at least one of them will be found and marked.

6. As decreases decreases to zero and decreases approaching , Eqn. (17.2). The decrease in is in line with the fact that there is less chance of a NL being rated higher than a LL, and one is completely dependent on at least one lesion being found.
7. As increases stays constant at the value determined by and , Eqn. (17.1), while approaches unity. The corresponding physical parameter increases approaching unity, guaranteeing every lesion will be found.
8. As long as each parameter is in the range  $> 0$ , the ROC curve is always proper.

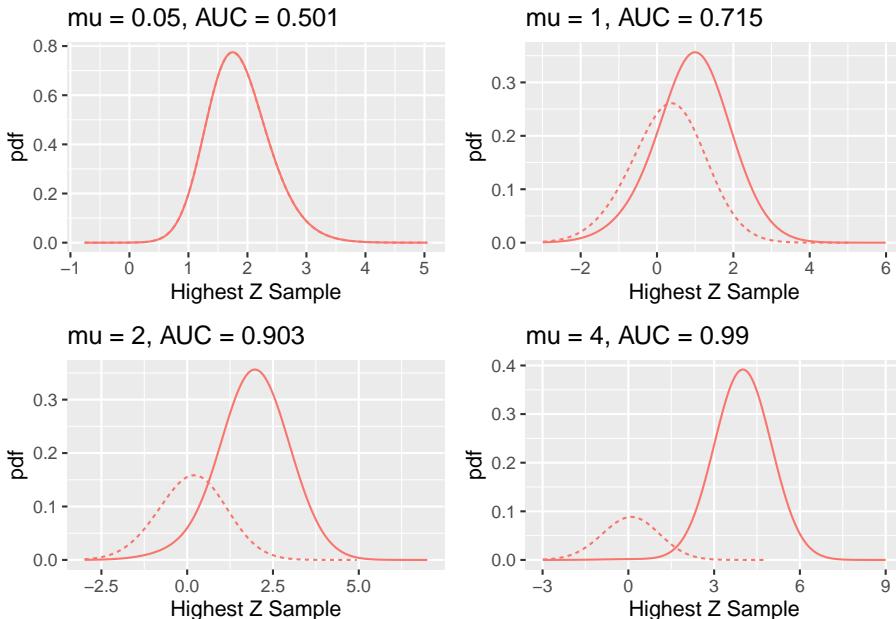


Figure 31.3: RSM-predicted pdf curves for indicated values of the  $\mu$  parameter. The solid curve corresponds to diseased cases and the dotted curve corresponds to non-diseased cases.

Fig. 31.3 shows pdf plots for the same values of parameters as in Fig. 31.2.

Consider the plot of the pdfs for  $\mu = 1$ . Since the integral of a pdf function over an interval amounts to counting the fraction of events occurring in the interval, it should be evident that the area under the non-diseased pdf equals  $F_{PF_{max}}$  and that under the diseased pdf equals  $TPF_{max}$ . For the chosen value  $\lambda = 1$  one has  $F_{PF_{max}} = 1 - e^{-\lambda} = 0.632$ . The reason for this has already

been given. To repeat, the area under the non-diseased pdf is less than unity because it is missing the contribution of non-diseased cases with no marks, the probability of which is  $e^{-\lambda} = e^{-1} = 0.368$ . Equivalently, it is missing the area under the straight line segment. Likewise, the area under the diseased pdf equals  $TPF_{max}$ , Eqn. (31.2), which is also less than unity. For the chosen values of  $\mu = \lambda = \nu = L = 1$  it equals  $TPF_{max} = 1 - e^{-\lambda}e^{-\nu} = 0.865$ . This area is somewhat larger than that under the non-diseased pdf, as is evident from visual examination of the plot. A greater fraction of diseased cases generate marks than do non-diseased cases, which is consistent with the presence of lesions in diseased cases. The complement of 0.865 is due to diseased cases with no marks, which account for a fraction 0.135 of diseased cases. To summarize, the pdf's do not integrate to unity for the reason that the integrals account only for the continuous section of the ROC curve and do not include cases with zero latent marks that do not generate z-samples. The effect becomes more exaggerated for higher values of  $\mu$  as this causes  $FPF_{max} = 1 - e^{-\lambda/\mu}$  to further decrease.

The plot in Fig. 17.2 labeled  $\mu = 0.05$  may be surprising. Since it corresponds to a small value of  $\mu$ , one may expect both pdfs to overlap and be centered at zero. Instead, while they do overlap, the shape is non-Gaussian and centered at approximately 1.8. This is because the small value of  $\mu$  results in a large value of the  $\lambda'$  parameter, since  $\lambda' = \lambda/\mu = 20$ . The highest of a large number of samples from the unit normal distribution is not normal and is peaked at a value above zero (Fisher and Tippett, 1928).

## 31.6 The RSM-predicted FROC curve

The derivation of the FROC curve is much simpler. From the property of the Poisson distribution, namely, its mean is the  $\lambda'$  parameter of the distribution, it follows that the expected number of latent NLs per case is  $\lambda'$ . One multiplies this by  $P(Z > \zeta | Z \sim N(0, 1))$ , i.e.,  $\Phi(-\zeta)$ , to obtain the expected number of latent NLs per case that is actually marked, i.e., NLF:

$$NLF(\zeta, \lambda') = \frac{\lambda}{\mu} \Phi(-\zeta) \quad (31.21)$$

Diseased cases are separated into groups, each with a fixed number of  $L$  lesions per case, where  $L$  varies from one to  $L_{max}$ . For each group characterized by  $L$ , one seeks the fraction of the expected number of latent LLs per case divided by the total number of lesions in each case ( $L$ ). Since  $\nu'$  is the probability that a lesion is found, it must equal the desired fraction. Next, one multiplies by  $P(Z > \zeta | Z \sim N(\mu, 1))$  i.e.,  $\Phi(\mu - \zeta)$ , to obtain the fraction that is actually marked. Finally, one performs a weighted summation over the different groups with  $f_L$  as the weighting fraction. Therefore,

$$\left. \begin{aligned} LLF(\zeta, \mu, \nu', \vec{f_L}) &= \sum_{L=1}^{L_{max}} f_L \nu' \Phi(\mu - \zeta) \\ &= \nu' \Phi(\mu - \zeta) \\ &= (1 - \exp(-\nu\mu)) \Phi(\mu - \zeta) \end{aligned} \right\} \quad (31.22)$$

Note that  $LLF(\zeta, \mu, \nu', \vec{f_L})$  is independent of  $\vec{f_L}$ . Summarizing, the coordinates of the RSM-predicted point on the FROC curve are given by Eqn. (31.21) and Eqn. (31.22). The FROC curve starts at (0,0) and ends at  $(\lambda', \nu')$ . The x-coordinate does not extend to arbitrarily large values and the y-coordinate does not approach unity (unless  $\nu'$ ). The constrained end-point property, demonstrated before for the ROC curve, also applies to the FROC curve:

$$\left. \begin{aligned} NLF_{max} &= \lambda/\mu \\ LLF_{max} &= 1 - \exp(-\nu\mu) \end{aligned} \right\} \quad (31.23)$$

### 31.7 The RSM-predicted AFROC curve

The AFROC x-coordinate is the same as the ROC x-coordinate and Eqn. (31.8) applies. The AFROC y-coordinate is identical to the FROC y-coordinate and the second Eqn. (31.22) applies. The second expression on the right hand side uses the intrinsic RSM parameters. Note that the expression is independent of the number of lesions in the dataset or their distribution (unlike the AFROC, the weighted AFROC does depend on the lesion distribution ) TBA!. The limiting coordinates of the AFROC are:

$$\left. \begin{aligned} FPF_{max} &= 1 - \exp(\lambda/\mu) \\ LLF_{max} &= 1 - \exp(-\nu\mu) \end{aligned} \right\} \quad (31.24)$$

It too has the constrained end-point property. In terms of the intrinsic RSM parameters, As  $\mu$  increases starting from  $0+$ ,  $FPF_{max}$  decreases starting from  $1-$ , approaching  $0+$  as  $\mu$  approaches  $\infty$ . In the same limit,  $TPF_{max}$  increases starting from 0, approaching  $1-$ . [The notation “1-” denotes a number just less than one.]

Source the file mainRsmAFROC.R. Note the change at line 7 with type = “AFROC”, which generates AFROC plots shown in Fig. 17.4 (A-F) for the following values of : 0.001, 1, 2, 3, 4 and 5.

As  $\mu$  increases, the area under the AFROC increases monotonically from 0 to 1. The reader should check that this is true regardless of the choices of the other parameters in the model. This is expected of a well-behaved area measure that can be used as a figure of merit.

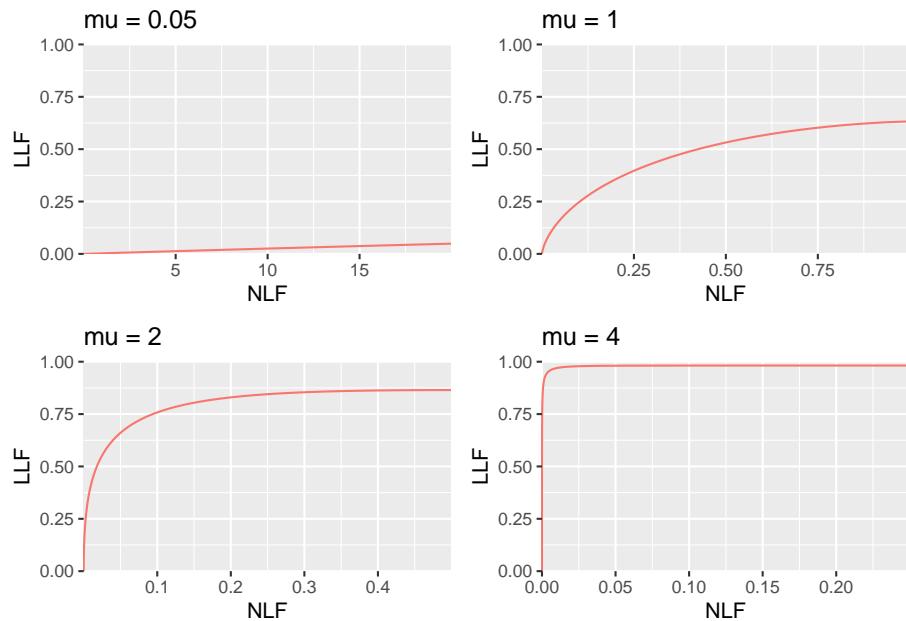


Figure 31.4: RSM-predicted FROC curves for indicated values of the  $\mu$  parameter. As  $\mu$  increases the curve approaches the top left corner, in the limit it is the vertical line connecting the origin to  $(0,1)$ . Notice the wide range of variation of the x-axis scaling. In top left it ranges from 0 to 20 while in bottom right it ranges from 0 to 0.2. The total area under the FROC curve actually decreases as  $\mu$  increases. Because it is not contained within the unit square, the FROC cannot be used as the basis of a meaningful figure of merit.

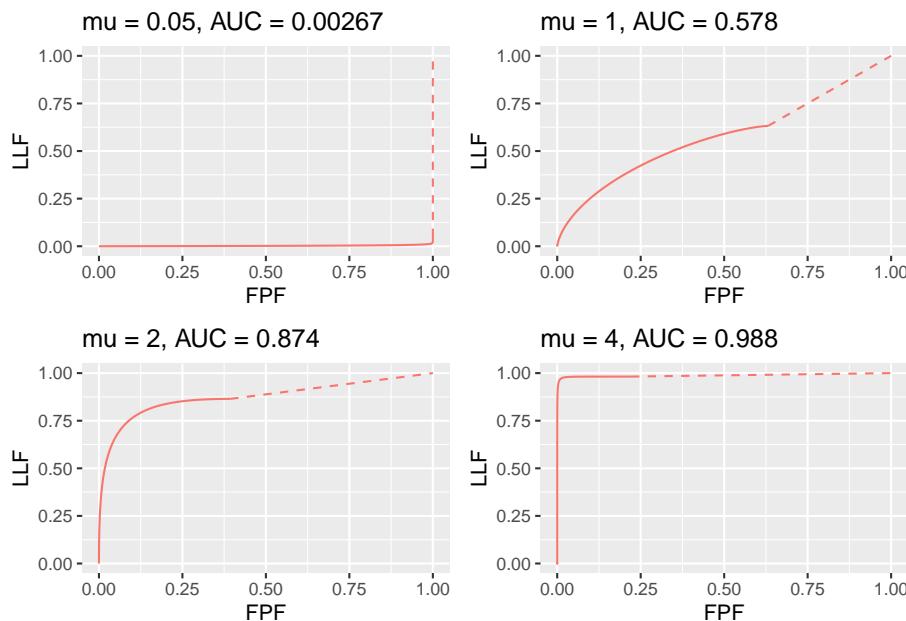


Figure 31.5: RSM-predicted AFROC curves for indicated values of the  $\mu$  parameter. As  $\mu$  increases, AFROC-AUC increases; the curve increasingly approaches the top-left corner, followed by an inaccessible dashed linear extension to  $(1,1)$ . Each plot is completely contained within the unit square, which makes it easy to define a figure of merit.

Experiment with different values for the parameters to confirm that the following statements are true:

1. The AFROC plot is independent of the number of lesions per case. Area under AFROC is independent of . These statements are not true for the wAFROC TBA!!.. In contrast, the ROC ordinate increases with increasing numbers of lesions per case.
2. From Eqn. (31.2) and Eqn. (31.23) it follows that  $TPF_{max} \geq LLF_{max}$  with the equality holding in the limit  $\mu \rightarrow \infty$ .

The physical reason for  $TPF_{max} \geq LLF_{max}$  is that the ROC gives credit for incorrect localizations on diseased cases, while the AFROC does not. This is the well-known “right for wrong reason” argument (Bunch et al., 1977b) originally advanced in 1977.

3. As increases the AFROC curve more closely approaches the upper-left corner of the plot, denoting increasing performance and the area under the AFROC curve approaches 1, which is the best possible performance: (a) decreases and (b) increases, denoting decreasing numbers of incorrect decisions on non-diseased and diseased cases, respectively.
4. For and non-zero the operating characteristic approaches the horizontal line extending from the origin to (1,0), which is the continuous section of the curve, followed by the vertical dashed line connecting (1,0) to (1,1) and AFROC-AUC approaches zero. In this limit, none of the lesions is localized and every case has at least one NL mark, which implies worst possible performance.
5. For the accessible portion of the operating characteristic approaches the vertical line connecting (0,0) to (0,1), the area under which is zero. The complete AFROC curve is obtained by connecting this point to (1,1) by the dashed line and in this limit the area under the complete ROC curve approaches 1. As with the ROC, omitting the area under the dashed portion of the curve will result in a severe underestimate of true performance.
6. As decreases decreases to zero while stays constant, Eqn. (17.35), as the latter is independent of .
7. As increases stays constant (it is independent of ) while approaches unity. As increases, the corresponding physical parameter increases, approaching unity, guaranteeing that every lesion is found. AFROC-AUC approaches one.
8. Over the range (0,0) to , the slope of the AFROC decreases monotonically. It is infinite at the origin and zero at the end-point.

### 31.7.1 Chance level performance on AFROC

There appears to be a misconception<sup>24,25</sup> that chance level performance on an AFROC corresponds to the positive diagonal of the plot, yielding AFROC-

AUC of 0.5. Fig. 17.4 (A) shows that chance level performance corresponds to the horizontal line connecting the origin to (1,0) and a vertical dashed line connecting (1,0) to (1,1) corresponding to AFROC-AUC of zero. If the lesion perceptual contrast is zero, then no lesions are found and all marks are NLs. The AFROC-AUC FOM, namely the probability that a lesion rating exceeds the ratings of NLs on non-diseased cases, see §14.2, is zero.

### 31.7.2 The reader who does not yield any marks

Suppose the radiologist does not mark any case, as in §13.4.2.2, resulting in an empty data file. One possibility is that the radiologist did not interpret the cases and simply “whizzed” through them. In this situation, the radiologist is not performing the diagnostic task. The AFROC operating point is stuck at the origin and connecting the straight-line extension yields AFROC-AUC = 0.5, would be incorrect as it implies finite performance (any value greater than zero for AFROC-AUC implies some degree of expertise). All models of observer performance assume that the observer is behaving rationally<sup>10</sup>, so this possibility is not analyzable. On the other hand, there is the real possibility that the radiologist did not detect any lesions and did not mark any non-diseased case. Assuming the radiologist is behaving rationally, one needs an explanation for AFROC-AUC = 0.5. It turns out that this observer is perfect at not generating NLs on non-diseased cases, so no patient is recalled incorrectly. The radiologist needs to get some credit for this ability, and this is the explanation of AFROC-AUC = 0.5 for a rational observer. Since this radiologist’s LLF = zero obviously the radiologist is far from perfect. The radiologist needs to be trained to find lesions. A suitable training set would consist of diseased cases only. The FROC data from such a dataset could be analyzed using the AFROC1 figure of merit, which could be used to measure the improvement of the radiologist in finding lesions. There is no point wasting non-diseased cases on this radiologist, as the radiologist has proven perfect performance on them by not generating NL marks on any non-diseased case.

The following code, mainNoMarks.R, illustrates how this observer can be simulated.

The corresponding AFROC plot, obtained by sourcing this code, is shown in Fig. 17.5.

Fig. 5 here

Fig. 17.5: The case of the rational observer who does not mark any image, who operates at (0,0) on the AFROC and for whom the AFROC plot consists only of the straight line extension connecting the origin to (1,1) and AFROC-AUC = 0.5. This observer has better performance (specifically unit case-level specificity) than the worst observer shown in Fig. 17.4 (A) who yielded AFROC-AUC = 0 (zero sensitivity and zero specificity). [This figure was generated by sourcing file mainNoMarks.R]

The explanation lies in the values of the chosen parameters. The parameter was set to 0.001 as setting it to zero would create a divide by zero error when is calculated. Instead one sets to 0.000001, so . With such small the probability is almost zero that any case will have a NL mark – according to the Poisson distribution, the probability of no mark is . Of course, the fact that is close to zero means the , so no lesion is found.

The corresponding code output is shown below.

This tells us that FROC-AUC = 0, ROC-AUC = 0.5 and AFROC-AUC = 0.5. FROC-AUC is meaningless as the operating point is stuck at the origin and one has no idea where it is supposed to end. ROC-AUC = 0.5 means that the observer is showing chance-level performance at the task of separating non-diseased and diseased cases. The ROC paradigm does not credit the observer for avoiding marking non-diseased cases. AFROC-AUC = 0.5 credits the observer for not marking any non-diseased cases but there is no credit for unmarked lesions. The difference between the ROC and FROC paradigms, both predicting AUC = 0.5, but these have different meanings, is because FROC is a location specific paradigm but ROC is not.

## 31.8 Discussion / Summary

This chapter has detailed ROC, FROC and AFROC curves predicted by the radiological search model (RSM). All RSM-predicted curves share the constrained end-point property that is qualitatively different from previous ROC models. In my experience, it is a property that most researchers in this field have difficulty accepting. There is too much history going back to the early 1940s, of the ROC curve extending from (0,0) to (1,1) that one has to let go of, and this can be difficult.

I am not aware of any direct evidence that radiologists can move the operating point continuously in the range (0,0) to (1,1) in search tasks, so the existence of such an ROC is tantamount to an assumption. Algorithmic observers that do not involve the element of search can extend continuously to (1,1). An example of an algorithmic observer not involving search is a diagnostic test that rates the results of a laboratory measurement, e.g., the A1C measure of blood glucose for presence of a disease. If A1C > 6.5% the patient is diagnosed as diabetic. By moving the threshold from infinity to -infinity, and assuming a large population of patients, one can trace out the entire ROC curve from the origin to (1,1). This is because every patient yields an A1C value. Now imagine that some finite fraction of the test results are “lost in the mail”; then the ROC curve, calculated over all patients, would have the constrained end-point property, albeit due to an unreasonable cause.

The situation in medical imaging involving search tasks is qualitatively different. Not every case yields a decision variable. There is a reasonable cause for this – to

render a decision variable sample the radiologist must find something suspicious to report, and if none is found, there is no decision variable to report. The ROC curve calculated over all patients would exhibit the constrained end-point property, even in the limit of an infinite number of patients. If calculated over only those patients that yielded at least one mark, the ROC curve would extend from (0,0) to (1,1) but then one would be ignoring the cases with no marks, which represent valuable information: unmarked non-diseased cases represent perfect decisions and unmarked diseased cases represent worst-case decisions.

RSM-predicted ROC, FROC and AFROC curves were derived (wAFROC is implemented in the Rjafroc). These were used to demonstrate that the FROC is a poor descriptor of performance. Since almost all work to date, including some by me TBA 47,48, has used FROC curves to measure performance, this is going to be difficulty for some to accept. The examples in Fig. 17.6 (A-F) and Fig. 17.7 (A-B) should convince one that the FROC curve is indeed a poor measure of performance. The only situation where one can safely use the FROC curve is if the two modalities produce curves extending over the same NLF range. This can happen with two variants of a CAD algorithm, but rarely with radiologist observers.

A unique feature is that the RSM provides measures of search and lesion-classification performance. It bears repeating that search performance is the ability to find lesions while avoiding finding non-lesions. Search performance can be determined from the position of the ROC end-point (which in turn is determined by RSM-based fitting of ROC data, Chapter 19). The perpendicular distance between the end-point and the chance diagonal is, apart from a factor of 1.414, a measure of search performance. All ROC models that predict continuous curves extending to (1,1), imply zero search performance.

Lesion-classification performance is measured by the AUC value corresponding to the parameter. Lesion-classification performance is the ability to discriminate between LLs and NLs, not between diseased and non-diseased cases: the latter is measured by RSM-AUC. There is a close analogy between the two ways of measuring lesion-classification performance and CAD used to find lesions in screening mammography vs. CAD used in the diagnostic context to determine if a lesion found at screening is actually malignant. The former is termed CADe, for CAD detection, which in my opinion, is slightly misleading as at screening lesions are found not detected (“detection” is “discover or identify the presence or existence of something”, correct localization is not necessarily implied; the more precise term is “localize”). In the diagnostic context one has CADx, for CAD diagnostic, i.e., given a specific region of the image, is the region malignant?

Search and lesion-classification performance can be used as “diagnostic aids” to optimize performance of a reader. For example, if search performance is low, then training using mainly non-diseased cases is called for, so the resident learns the different variants of non-diseased tissues that can appear to be true lesions. If lesion-classification performance is low then training with diseased cases only

is called for, so the resident learns the distinguishing features characterizing true lesions from non-diseased tissues that fake true lesions.

Finally, evidence for the RSM is summarized. Its correspondence to the empirical Kundel-Nodine model of visual search that is grounded in eye-tracking measurements. It reduces in the limit of large  $n$ , which guarantees that every case will yield a decision variable sample, to the binormal model; the predicted pdfs in this limit are not strictly normal, but deviations from normality would require very large sample size to demonstrate. Examples were given where even with 1200 cases the binormal model provides statistically good fits, as judged by the chi-square goodness of fit statistic, Table 17.2. Since the binormal model has proven quite successful in describing a large body of data, it satisfying that the RSM can mimic it in the limit of large  $n$ . The RSM explains most empirical results regarding binormal model fits: the common finding that  $b < 1$ ; that  $b$  decreases with increasing lesion pSNR (large and / or  $\sigma$ ); and the finding that the difference in means divided by the difference in standard deviations is fairly constant for a fixed experimental situation, Table 17.3. The RSM explains data degeneracy, especially for radiologists with high expertise.

The contaminated binormal model2-4 (CBM), Chapter 20, which models the diseased distribution as having two peaks, one at zero and the other at a constrained value, also explains the empirical observation that b-parameter  $< 1$  and data degeneracy. Because it allows the ROC curve to go continuously to (1,1), CBM does not completely account for search performance – it accounts for search when it comes to finding lesions, but not for avoiding finding non-lesions.

I do not want to leave the impression that RSM is the ultimate model. The current model does not predict satisfaction of search (SOS) effects<sup>27-29</sup>. Attempts to incorporate SOS effects in the RSM are in the early research stage. As stated earlier, the RSM is a first-order model: a lot of interesting science remains to be uncovered.

### 31.8.1 The Wagner review

The two RSM papers<sup>12,13</sup> were honored by being included in a list of 25 papers the “Highlights of 2006” in Physics in Medicine and Biology. As stated by the publisher: “I am delighted to present a special collection of articles that highlight the very best research published in Physics in Medicine and Biology in 2006. Articles were selected for their presentation of outstanding new research, receipt of the highest praise from our international referees, and the highest number of downloads from the journal website.”

One of the reviewers was the late Dr. Robert (“Bob”) Wagner – he had an open-minded approach to imaging science that is lacking these days, and a unique writing style. I reproduce one of his comments with minor edits, as it pertains to the most interesting and misunderstood prediction of the RSM, namely its constrained end-point property.

I'm thinking here about the straight-line piece of the ROC curve from the max to (1, 1). 1. This can be thought of as resulting from two overlapping uniform distributions (thus guessing) far to the left in decision space (rather than delta functions). Please think some more about this point—because it might make better contact with the classical literature. 2. BTW – it just occurs to me (based on the classical early ROC work of Swets & co.) – that there is a test that can resolve the issue that I struggled with in my earlier remarks. The experimenter can try to force the reader to provide further data that will fill in the space above the max point. If the results are a straight line, then the reader would just be guessing – as implied by the present model. If the results are concave downward, then further information has been extracted from the data. This could require a great amount of data to sort out—but it's an interesting point (at least to me).

Dr. Wagner made two interesting points. With his passing, I have been deprived of the penetrating and incisive evaluation of his ongoing work, which I deeply miss. Here is my response (ca. 2006):

The need for delta functions at negative infinity can be seen from the following argument. Let us postulate two constrained width pdfs with the same shapes but different areas, centered at a common value far to the left in decision space, but not at negative infinity. These pdfs would also yield a straight-line portion to the ROC curve. However, they would be inconsistent with the search model assumption that some images yield no decision variable samples and therefore cannot be rated in bin ROC:2 or higher. Therefore, if the distributions are as postulated above then choice of a cutoff in the neighborhood of the overlap would result in some of these images being rated 2 or higher, contradicting the RSM assumption. The delta function pdfs at negative infinity are seen to be a consequence of the search model.

One could argue that when the observer sees nothing to report then he starts guessing and indeed this would enable the observer to move along the dashed portion of the curve. This argument implies that the observer knows when the threshold is at negative infinity, at which point the observer turns on the guessing mechanism (the observer who always guesses would move along the chance diagonal). In my judgment, this is unreasonable. The existence of two thresholds, one for moving along the non-guessing portion and one for switching to the guessing mode would require abandoning the concept of a single decision rule. To preserve this concept one needs the delta functions at negative infinity.

Regarding Dr. Wagner's second point, it would require a great amount of data to sort out whether forcing the observer to guess would fill in the dashed portion of the curve, but I doubt it is worth the effort. Given the bad consequences of guessing (incorrect recalls) I believe that in the clinical situation, the radiologist will not knowingly guess. If the radiologist sees nothing to report, nothing will be reported. In addition, I believe that forcing the observer, to prove some research point, is not a good idea.

### 31.9 References

1. Chakraborty DP. Computer analysis of mammography phantom images (CAMPi): An application to the measurement of microcalcification image quality of directly acquired digital images. *Medical Physics.* 1997;24(8):1269-1277.
2. Chakraborty DP, Eckert MP. Quantitative versus subjective evaluation of mammography accreditation phantom images. *Medical Physics.* 1995;22(2):133-143.
3. Chakraborty DP, Yoon H-J, Mello-Thoms C. Application of threshold-bias independent analysis to eye-tracking and FROC data. *Academic Radiology.* 2012;In press.
4. Chakraborty DP. ROC Curves predicted by a model of visual search. *Phys Med Biol.* 2006;51:3463-3482.
5. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. *Phys Med Biol.* 2006;51:3449-3462.

## Chapter 32

# Search and classification performances

### 32.1 TBA How much finished

10%

### 32.2 Introduction

The preceding chapter described the radiological search model (RSM) for FROC data. This chapter describes predictions of the RSM and how they compare with evidence. The starting point is the inferred ROC curve. While mathematically rather complicated, the results are important because they are needed to derive the ROC-likelihood function, which is used to estimate RSM parameters from ROC data in TBA Chapter 19. The preceding sentence should lead the inquisitive reader to the question: *since the ROC paradigm ignores search, how is it possible to derive parameters of a model of search from the ROC curve?* The answer is that the *shape* of the ROC curve contains information about the RSM parameters. It is fundamentally different from predictions of all conventional ROC models: binormal (Dorfman and Alf, 1969), contaminated binormal model (Dorfman and Berbaum, 2000), bigamma (Dorfman et al., 1997) and proper ROC (Metz and Pan, 1999), namely it has a *constrained end-point property*, while all other models predict that the *end-point*, namely the uppermost non-trivial point on the ROC, reached at infinitely low reporting threshold, is (1,1), while the RSM predicts it does not reach (1,1). The nature of search is such that the limiting end-point is constrained to be below and to the left of (1,1). This key difference, allows one to estimate search parameters from ROC data.

Next, the RSM is used to predict FROC and AFROC curves. Two following sections show how search performance and lesion-classification performance can be quantified from the location of the ROC end-point. Search performance is the ability to find lesions while avoiding finding non-lesions, and lesion-classification performance is the ability, having found a suspicious region, to correctly classify it; if classified as a NL it would not be marked (in the mind of the observer every mark is a potential LL, albeit at different confidence levels). Note that lesion-classification is different from classification between diseased and non-diseased cases, which is measured by the ROC-AUC. Based on the ROC/FROC/AFROC curve predictions of the RSM, a comparison is presented between area measures that can be calculated from FROC data, and this leads to an important conclusion, namely the FROC curve is a poor descriptor of search performance and that the AFROC/wAFROC are preferred. This will come as a surprise (shock?) to most researchers somewhat familiar with this field, since the overwhelming majority of users of FROC methods, particularly in CAD, have relied on the FROC curve. Finally, evidence for the validity of the RSM is presented.

### 32.3 Quantifying search performance #rsm-search-search-performance}

Fig. 6 here Fig. 17.6: This figure shows a typical population ROC curve labeled (a) predicted by ROC models that do not account for search performance. Search performance is defined as the ability to find lesions while avoiding non-lesions. The end-point of such a curve is at (1,1), denoted by the filled circle. By adopting a sufficiently low reporting threshold the observer can continuously move the operating point from (0,0) to (1,1). The curve labeled (b) is a typical RSM-predicted ROC curve. The end-point is downward and leftward shifted relative to (1,1), as indicated by the filled square. The observer cannot move the operating point continuously all the way from (0,0) to (1,1) because a constrained fraction of images contain no marks. The fractions of unmarked non-diseased and diseased cases determine the location of the end-point, respectively. The observer can move the operating point continuously from the origin to the end-point and no further. The location of the end-point is a measure of search performance. Higher search performance is characterized by the end-point moving upwards and to the left, ideally to (0,1) which corresponds to perfect search performance. The perpendicular distance from the end-point to the chance diagonal (c) multiplied by  $\sqrt{2}$ , i.e., , is defined as a measure of search performance. Lesion-classification performance is defined as the implied AUC of two unit variance normal distributions separated by the parameter of the search model. It measures the ability, having found a suspicious region, to correctly classify it as a true lesion. The code for this plot is in file mainQuantifySearchPerformance.R.

In Fig. 17.6, the line labeled (a) is a conventional model ROC curve ending at

(1,1), the filled circle, while (b) shows a typical search model ROC curve ending at a point below and to the left of (1,1), the filled square. The location of the end-point of the RSM-predicted curve determines the search performance of the observer. The square root of two times the perpendicular distance (the subscript s is for search) from the end-point to the chance diagonal in Fig. 17.6, the line labeled (c), is defined as search-performance, denoted S. For example, if and then the end-point is (0,1) and  $S = 1$ . This observer has perfect search performance since no NLs are found and all lesions were found; the perpendicular distance from (0,1) to the chance diagonal is  $1/\sqrt{2}$ , which multiplied by  $\sqrt{2}$  yields unity. Search performance ranges from 0 to 1. Using geometry, Eqn. (17.1) and Eqn. (17.2), it follows that:

$$\dots \quad (17.37)$$

Therefore, search performance is given by:

$$\dots \quad (17.38)$$

The second form in Eqn. (17.38) shows S in terms of the physical (i.e., primed) parameters; it shows that search performance is the product of two terms: the probability of finding lesions times the probability of avoiding finding any non-lesions. This puts into a mathematical form the qualitative definition of search performance as the ability to find lesions while avoiding finding non-lesions. Since at least one parameter is needed to describe each of these probabilities, quantifying search requires at least two parameters.

Applying this definition to the case of the rational observer who does not generates any marks, one sees that the observer's search performance is zero . This emphasizes the point that not generating NLs is not enough; one must also be able to find lesions. It is also consistent with the fact that the origin lies on the positive diagonal of the ROC, implying zero perpendicular distance between them, i.e., .

## 32.4 Quantifying classification performance

To avoid misunderstanding, I emphasize that lesion-classification performance is being used in a different sense from that used in ROC methodology, where classification is between diseased and non-diseased cases, not between diseased and non-diseased regions, i.e., latent NLs and latent LLs, as in the current context.

Having found a suspicious region, how good is the observer at correctly classifying true lesions and non-lesions? Lesion-classification performance C is determined by the parameter, and is defined by the implied AUC of unit variance normal distributions separated by .

$$\dots \quad (17.39)$$

It ranges from 0.5 to 1. Only one parameter is needed for this, so one needs three parameters to quantify search and lesion-classification performance.

### **32.4.1 Lesion-classification performance and the 2AFC LKE task**

It should be obvious that lesion-classification performance is similar to what is commonly measured in model-observer research using the location-known-exactly (LKE) paradigm. In this paradigm, one uses 2AFC methods as in Fig. 4.3, but one could use the ratings method as long as the lesion is cued (i.e., pointed to). On diseased cases, the lesion is cued, but to control for false positives, one must also cue a similar region on non-diseased cases, as in Fig. 4.3. In that figure, the lesion, present in one of the two images, is always in the center of one of the two fields. Sometimes cross hairs are used to indicate where the observer should be looking. The probability of a correct choice in the 2AFC task is , i.e., AUC conditioned on the (possible) position of the lesion being cued. Since the lesion is cued, search performance of the observer is irrelevant, and one expects . The reason for the inequality is that on a non-diseased case, the location being cued, in all likelihood, does not correspond to a latent NL found by the observer's search mechanism. Latent NLs are more suspicious for disease than other locations in the case. measures the separation parameter between latent NLs and LLs. The separation parameter between latent LLs and a researcher chosen location is likely to be larger. This is because latent NLs are more suspicious for disease than a researcher chosen location. It is known that performance under this condition exceeds that in a free-search 2AFC or ROC study, denoted AUC, where the lesion is not cued and it could be anywhere. This should be obvious – pointing to the possible location of the lesion takes out the need for searching the rest of the image, which introduces the possibility of not finding the lesion and / or finding non-lesions. One expects the following ordering: . is expected to be the least, as there is uncertainty about possible lesion location. is expected to be next in order, as now uncertainty has been reduced, and the observer's task is to pick between two cued locations, one a latent NL and the other a latent LL. is expected to be highest, as now the observer's task is to pick between two cued locations, one a latent LL and the other a researcher chosen location, most likely not a latent NL. Data supporting the expected inequality is presented in §19.5.4.6.

### **32.4.2 Significance of measuring search and lesion-classification performance**

The ability to quantify search and lesion-classification performance from a single paradigm (ROC) study is highly significant, going well-beyond modeling the ROC curve. ROC-AUC measures how well an observer is able to separate two groups of patients, a group of diseased patients from a group of non-diseased

patients. While important, it does not inform us about how the observer goes about doing this and what is limiting performance (an exception the CBM model which yields information about how good the observer is at finding lesions but does not account for the ability of the observer to avoid NLs on non-diseased cases). In contrast, the search and lesion-classification measures described above can be used as a “diagnostic aid” in determining what is limiting performance. If search performance is poor, it indicates that the observer needs to be trained on many non-diseased cases, and learn the variants of non-diseased anatomy and learn not to confuse them for lesions. On the other hand, if lesion-classification performance is poor, then one needs to train the observer using images where the location of a possible lesion is cued, and the observer’s task is to determine if the cued location is a real lesion. The classic example here is breast CAD, where the designer level ROC curve goes almost all the way to (1,1) implying poor search performance, while lesion-classification performance could actually be quite good, because CAD has access to the pixel values and the ability to apply complex algorithms to properly classify lesions as benign or malignant.

Of course, before one can realize these benefits, one needs a way of estimating the end-point shown in Fig. 17.6 plot (b). The observer will generally not oblige by reporting every suspicious region. RSM based curve fitting is needed to estimate the end-point’s location, Chapter 19.

## 32.5 Discussion / Summary

This chapter has detailed ROC, FROC and AFROC curves predicted by the radiological search model (RSM). All RSM-predicted curves share the constrained end-point property that is qualitatively different from previous ROC models. In my experience, it is a property that most researchers in this field have difficulty accepting. There is too much history going back to the early 1940s, of the ROC curve extending from (0,0) to (1,1) that one has to let go of, and this can be difficult.

I am not aware of any direct evidence that radiologists can move the operating point continuously in the range (0,0) to (1,1) in search tasks, so the existence of such an ROC is tantamount to an assumption. Algorithmic observers that do not involve the element of search can extend continuously to (1,1). An example of an algorithmic observer not involving search is a diagnostic test that rates the results of a laboratory measurement, e.g., the A1C measure of blood glucose for presence of a disease. If A1C > 6.5% the patient is diagnosed as diabetic. By moving the threshold from infinity to -infinity, and assuming a large population of patients, one can trace out the entire ROC curve from the origin to (1,1). This is because every patient yields an A1C value. Now imagine that some finite fraction of the test results are “lost in the mail”; then the ROC curve, calculated over all patients, would have the constrained end-point property, albeit due to an unreasonable cause.

The situation in medical imaging involving search tasks is qualitatively different. Not every case yields a decision variable. There is a reasonable cause for this – to render a decision variable sample the radiologist must find something suspicious to report, and if none is found, there is no decision variable to report. The ROC curve calculated over all patients would exhibit the constrained end-point property, even in the limit of an infinite number of patients. If calculated over only those patients that yielded at least one mark, the ROC curve would extend from (0,0) to (1,1) but then one would be ignoring the cases with no marks, which represent valuable information: unmarked non-diseased cases represent perfect decisions and unmarked diseased cases represent worst-case decisions.

RSM-predicted ROC, FROC and AFROC curves were derived (wAFROC is implemented in the Rjafroc). These were used to demonstrate that the FROC is a poor descriptor of performance. Since almost all work to date, including some by me 47,48, has used FROC curves to measure performance, this is going to be difficult for some to accept. The examples in Fig. 17.6 (A- F) and Fig. 17.7 (A-B) should convince one that the FROC curve is indeed a poor measure of performance. The only situation where one can safely use the FROC curve is if the two modalities produce curves extending over the same NLF range. This can happen with two variants of a CAD algorithm, but rarely with radiologist observers.

A unique feature is that the RSM provides measures of search and lesion-classification performance. It bears repeating that search performance is the ability to find lesions while avoiding finding non-lesions. Search performance can be determined from the position of the ROC end-point (which in turn is determined by RSM-based fitting of ROC data, Chapter 19). The perpendicular distance between the end-point and the chance diagonal is, apart from a factor of 1.414, a measure of search performance. All ROC models that predict continuous curves extending to (1,1), imply zero search performance.

Lesion-classification performance is measured by the AUC value corresponding to the parameter. Lesion-classification performance is the ability to discriminate between LLs and NLs, not between diseased and non-diseased cases: the latter is measured by RSM-AUC. There is a close analogy between the two ways of measuring lesion-classification performance and CAD used to find lesions in screening mammography vs. CAD used in the diagnostic context to determine if a lesion found at screening is actually malignant. The former is termed CADe, for CAD detection, which in my opinion, is slightly misleading as at screening lesions are found not detected (“detection” is “discover or identify the presence or existence of something”, correct localization is not necessarily implied; the more precise term is “localize”). In the diagnostic context one has CADx, for CAD diagnostic, i.e., given a specific region of the image, is the region malignant?

Search and lesion-classification performance can be used as “diagnostic aids” to optimize performance of a reader. For example, if search performance is low, then training using mainly non-diseased cases is called for, so the resident learns

the different variants of non-diseased tissues that can appear to be true lesions. If lesion-classification performance is low then training with diseased cases only is called for, so the resident learns the distinguishing features characterizing true lesions from non-diseased tissues that fake true lesions.

Finally, evidence for the RSM is summarized. Its correspondence to the empirical Kundel-Nodine model of visual search that is grounded in eye-tracking measurements. It reduces in the limit of large  $\lambda$ , which guarantees that every case will yield a decision variable sample, to the binormal model; the predicted pdfs in this limit are not strictly normal, but deviations from normality would require very large sample size to demonstrate. Examples were given where even with 1200 cases the binormal model provides statistically good fits, as judged by the chi-square goodness of fit statistic, Table 17.2. Since the binormal model has proven quite successful in describing a large body of data, it satisfying that the RSM can mimic it in the limit of large  $\lambda$ . The RSM explains most empirical results regarding binormal model fits: the common finding that  $b < 1$ ; that  $b$  decreases with increasing lesion pSNR (large and / or  $\lambda$ ); and the finding that the difference in means divided by the difference in standard deviations is fairly constant for a fixed experimental situation, Table 17.3. The RSM explains data degeneracy, especially for radiologists with high expertise.

The contaminated binormal model2-4 (CBM), Chapter 20, which models the diseased distribution as having two peaks, one at zero and the other at a constrained value, also explains the empirical observation that b-parameter  $< 1$  and data degeneracy. Because it allows the ROC curve to go continuously to (1,1), CBM does not completely account for search performance – it accounts for search when it comes to finding lesions, but not for avoiding finding non-lesions.

I do not want to leave the impression that RSM is the ultimate model. The current model does not predict satisfaction of search (SOS) effects<sup>27-29</sup>. Attempts to incorporate SOS effects in the RSM are in the early research stage. As stated earlier, the RSM is a first-order model: a lot of interesting science remains to be uncovered.

### 32.5.1 The Wagner review

The two RSM papers<sup>12,13</sup> were honored by being included in a list of 25 papers the “Highlights of 2006” in Physics in Medicine and Biology. As stated by the publisher: “I am delighted to present a special collection of articles that highlight the very best research published in Physics in Medicine and Biology in 2006. Articles were selected for their presentation of outstanding new research, receipt of the highest praise from our international referees, and the highest number of downloads from the journal website.

One of the reviewers was the late Dr. Robert (“Bob”) Wagner – he had an open-minded approach to imaging science that is lacking these days, and a unique writing style. I reproduce one of his comments with minor edits, as it pertains

to the most interesting and misunderstood prediction of the RSM, namely its constrained end-point property.

I'm thinking here about the straight-line piece of the ROC curve from the max to  $(1, 1)$ . 1. This can be thought of as resulting from two overlapping uniform distributions (thus guessing) far to the left in decision space (rather than delta functions). Please think some more about this point—because it might make better contact with the classical literature. 2. BTW – it just occurs to me (based on the classical early ROC work of Swets & co.) – that there is a test that can resolve the issue that I struggled with in my earlier remarks. The experimenter can try to force the reader to provide further data that will fill in the space above the max point. If the results are a straight line, then the reader would just be guessing – as implied by the present model. If the results are concave downward, then further information has been extracted from the data. This could require a great amount of data to sort out—but it's an interesting point (at least to me).

Dr. Wagner made two interesting points. With his passing, I have been deprived of the penetrating and incisive evaluation of his ongoing work, which I deeply miss. Here is my response (ca. 2006):

The need for delta functions at negative infinity can be seen from the following argument. Let us postulate two constrained width pdfs with the same shapes but different areas, centered at a common value far to the left in decision space, but not at negative infinity. These pdfs would also yield a straight-line portion to the ROC curve. However, they would be inconsistent with the search model assumption that some images yield no decision variable samples and therefore cannot be rated in bin ROC:2 or higher. Therefore, if the distributions are as postulated above then choice of a cutoff in the neighborhood of the overlap would result in some of these images being rated 2 or higher, contradicting the RSM assumption. The delta function pdfs at negative infinity are seen to be a consequence of the search model.

One could argue that when the observer sees nothing to report then he starts guessing and indeed this would enable the observer to move along the dashed portion of the curve. This argument implies that the observer knows when the threshold is at negative infinity, at which point the observer turns on the guessing mechanism (the observer who always guesses would move along the chance diagonal). In my judgment, this is unreasonable. The existence of two thresholds, one for moving along the non-guessing portion and one for switching to the guessing mode would require abandoning the concept of a single decision rule. To preserve this concept one needs the delta functions at negative infinity.

Regarding Dr. Wagner's second point, it would require a great amount of data to sort out whether forcing the observer to guess would fill in the dashed portion of the curve, but I doubt it is worth the effort. Given the bad consequences of guessing (incorrect recalls) I believe that in the clinical situation, the radiologist will never guess. If the radiologist sees nothing to report, nothing will be re-

ported. In addition, I believe that forcing the observer, to prove some research point, is not a good idea.

## 32.6 References

1. Chakraborty DP. Computer analysis of mammography phantom images (CAMPI): An application to the measurement of microcalcification image quality of directly acquired digital images. *Medical Physics*. 1997;24(8):1269-1277.
2. Chakraborty DP, Eckert MP. Quantitative versus subjective evaluation of mammography accreditation phantom images. *Medical Physics*. 1995;22(2):133-143.
3. Chakraborty DP, Yoon H-J, Mello-Thoms C. Application of threshold-bias independent analysis to eye-tracking and FROC data. *Academic Radiology*. 2012;In press.
4. Chakraborty DP. ROC Curves predicted by a model of visual search. *Phys Med Biol*. 2006;51:3463–3482.
5. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. *Phys Med Biol*. 2006;51:3449–3462.



# Chapter 33

## The FROC should not be used to measure performance

### 33.1 TBA How much finished

10%

### 33.2 Introduction

The preceding chapter described the radiological search model (RSM) for FROC data. This chapter describes predictions of the RSM and how they compare with evidence. The starting point is the inferred ROC curve. While mathematically rather complicated, the results are important because they are needed to derive the ROC-likelihood function, which is used to estimate RSM parameters from ROC data in TBA Chapter 19. The preceding sentence should lead the inquisitive reader to the question: *since the ROC paradigm ignores search, how is it possible to derive parameters of a model of search from the ROC curve?* The answer is that the *shape* of the ROC curve contains information about the RSM parameters. It is fundamentally different from predictions of all conventional ROC models: binormal (Dorfman and Alf, 1969), contaminated binormal model (Dorfman and Berbaum, 2000), bigamma (Dorfman et al., 1997) and proper ROC (Metz and Pan, 1999), namely it has a *constrained end-point property*, while all other models predict that the *end-point*, namely the uppermost non-trivial point on the ROC, reached at infinitely low reporting threshold, is (1,1), while the RSM predicts it does not reach (1,1). The nature of search is such

that the limiting end-point is constrained to be below and to the left of (1,1). This key difference, allows one to estimate search parameters from ROC data. Next, the RSM is used to predict FROC and AFROC curves. Two following sections show how search performance and lesion-classification performance can be quantified from the location of the ROC end-point. Search performance is the ability to find lesions while avoiding finding non-lesions, and lesion-classification performance is the ability, having found a suspicious region, to correctly classify it; if classified as a NL it would not be marked (in the mind of the observer every mark is a potential LL, albeit at different confidence levels). Note that lesion-classification is different from classification between diseased and non-diseased cases, which is measured by the ROC-AUC. Based on the ROC/FROC/AFROC curve predictions of the RSM, a comparison is presented between area measures that can be calculated from FROC data, and this leads to an important conclusion, namely the FROC curve is a poor descriptor of search performance and that the AFROC/wAFROC are preferred. This will come as a surprise (shock?) to most researchers somewhat familiar with this field, since the overwhelming majority of users of FROC methods, particularly in CAD, have relied on the FROC curve. Finally, evidence for the validity of the RSM is presented.

### 33.3 The FROC curve is a poor descriptor of search performance

Why is the FROC curve is a bad descriptor of performance? The basic reason is that it is unconstrained in the x-direction<sup>26</sup>. Experts do not “move” as much along the positive x-direction as non-experts and partial area measures lose their meaning. Another reason is that it depends on the marks; unmarked non-diseased cases – representing perfect decisions - are not taken into account. The only meaningful comparison between two FROC curves occurs when they have a common NLF range, but this is rarely the case. As predicted by the RSM, a common range of NLF occurs when the two curves differ only in the parameter: if and are the same, then Eqn. (17.30) predicts the two curves will have identical . As shown below with numerical integration, this is the only situation where the area under the FROC tracks the area under the ROC, where the latter is regarded as the gold standard.

The code in file mainIsFrocGood.R, Online Appendix 17.E, calculates, by numerical integration, the areas under the full FROC, ROC and AFROC curves. Each full curve consists of the continuously accessible part plus any straight-line extension to (1,1), if applicable.

Fig. 7 here

The code is divided into 3 parts: \* Part I, lines 15 – 64, calculates , and for varying , with ; \* Part II, lines 66 – 115, calculates the same AUCs for varying

### 33.3. THE FROC CURVE IS A POOR DESCRIPTOR OF SEARCH PERFORMANCE555

, with ; and \* Part III, lines 117 – 159, calculates the same AUCs for varying , with .

This code takes a few minutes to complete running. The plots generated by this code are shown in Fig. 17.7 (A - F). The first column indicates which RSM parameter is being varied, ROC-AUC = is plotted along the x-axis, while is plotted along the y-axis in the left plot and is plotted along the y-axis in the right plot. The idea is that is the gold standard as it measures basic classification ability between diseased and non-diseased cases. So, for a valid figure of merit, the quantity plotted along the y-axis should monotonically increase with the gold standard, i.e., the slope should be positive. This is always true for but is not always true for ; it is only true when is varied, which, as was noted above, is the only situation when the range of integration along the NLF axis is constant.

Fig. 17.7 (A- F): Plots of plots along the x-axis, while is plotted along the y-axis in the left plot and is plotted along the y-axis in the right plot. Plots (A) and (B) correspond to varying , with and ; approximate slope AFROC vs. ROC = 2.00; plots (C) and (D) correspond to varying , with and ; approximate slope AFROC vs. ROC = 1.84; and plots (E) and (F) correspond to varying , with and ; approximate slope AFROC vs. ROC = 1.42. Regarding as the gold standard, the quantity plotted along the y-axis should be monotonic with the gold standard. This is always true for but is not always true for : it is only true for the varying . FROC-AUC is not constrained to unity; see plot (C); the AFROC-AUC is always in the range 0 to 1; see plots (B), (D) and (F). These plots were generated by mainIsFrocGood.R.

The plots of the FROC-AUC in (A) and (C) are non-linear and have negative slope. In contrast, the AFROC-AUCs have a quasi linear dependence on ROC-AUC. [The empirically determined slopes are printed by the code. For plot B the slope is 2.00, for plot D the slope is 1.84 and for plot F the slope is 1.42. These slopes indicate how much an ROC-AUC effect-size is amplified in the AFROC FOM. If only is different between two modalities, the amplification is almost exactly a factor of two. In the worst-case scenario, if only is different, the amplification is a factor of 1.42. In general, all three quantities could be different; one expects an intermediate amplification of the effect-size, in the range 1.4 to 2.]

One could argue that the above comparison is unfair to FROC as it considers the whole area under the FROC, while most users would use a point measure or a partial area measure, e.g., LLF @ selected NLF. The problem then is that some readers (especially the really good ones) cannot be analyzed as all of their operating points could to the left of the selected NLF value, and one would need to extrapolate outside the range of observed values in order to get the desired LLF @ selected NLF. For other readers, the data lying to the right of the selected NLF value does not contribute to the measure, resulting in loss of measurement accuracy

It is instructive to consider the extreme cases of a perfect observer and the worst

observer to see how the two methods of plotting would deal with defining the average observer. To make the comparison easier, consider that the lesions are small compared to the image area, so that the chance of a random LL is very small.

Fig. 17.8: (A) FROC curves for expert observer: vertical line extending from (0,0) to (0,1) and worst observer: horizontal line over the indicated NLF range. It is not possible to define an average FROC curve, as a common NLF range for the two observers does not exist. (B) Corresponding AFROC curves. AFROC-AUC for a perfect observer is unity (the area includes that under the dashed section extending from (0,1) to (1,1)). The corresponding area for the worst observer is zero, and the average AFROC curve is a straight line parallel to the x-axis at ordinate of 0.5, so the area under the average AFROC-AUC is 0.5 (unlike the ROC-AUC, AFROC-AUC = 0.5 does not denote worst possible performance). This plot was generated by the code in mainBestWorstObserver.R.

The perfect observer ( $LLF = 1 @ NLF = 0$ ) and the worst observer ( $LLF = 0 @ NLF < \text{some constrained value}$ ) both yield identical areas (zero) under the FROC curves. and it is not possible to define an average FROC curve, Fig. 17.7 (A). Because the two plots do not share a common range of abscissa values one cannot define an average FROC curve. In contrast, the AFROC is contained to the unit square and the area under the AFROC curve, Fig. 17.7 (B), is unity for the perfect observer (the area includes that under the dashed section extending from (0,1) to (1,1)). The corresponding area for the worst observer is zero, and the average AFROC curve is a straight line parallel to the x-axis at ordinate of 0.5, so the area under the average AFROC is 0.5 (as already noted, unlike the ROC area, AFROC area = 0.5 does not denote worst possible performance).

The FROC curve depends only on the marks. A valid FOM should reward correct decisions and penalize incorrect ones on all cases (in my judgment, the use of partial area measures, widespread in the literature, needs to be reconsidered). Unmarked non-diseased cases are perfect decisions, but these are not accounted for in the FROC curve (they indirectly affect the curve by the leftward movement of the uppermost point, all the way to  $NLF = 0$  for a perfect observer, but these decisions are not accounted for in FROC curve based partial area measures). The area under the horizontal dashed curve in the AFROC curve shown in Fig. 17.7 (B) is due to unmarked images. See §14.4.2 for further discussion of the meaning of the area under the dashed portion of the AFROC plot.

Finally, FP marks on diseased cases don't have the same negative connotation as those on non-diseased cases, since, following diagnostic workup, it is possible that the cancer will be found on the recalled cases, but, unlike the AFROC, both contribute to the FROC x-axis.

The RSM is a first-order model: a lot of interesting science remains to be uncovered. It does not account for the satisfaction of search (SOS) effects<sup>27-29</sup> observed in medical imaging. It is as if the radiologist senses that an image is possibly diseased, without being able to pinpoint the specific reason, and

therefore adopts a more cautious reporting style. They are more reluctant to mark NLs on diseased than on non-diseased cases. This means the probability the a LL rating exceeds the rating of a NL on diseased cases is not equal to the probability that a LL rating exceeds the rating of a NL on non-diseased cases:

. (17.40)

Therefore, inclusion of inter-comparisons between LLs and NLs on diseased cases would make the figure of merit depend on disease prevalence, thereby destroying a desirable property of a valid figure of merit. This is another reason for excluding such comparisons on diseased cases in the AFROC/wAFROC figures of merit.

### 33.3.1 Clinically relevant portion of an operating characteristic #rsm-goodbye-froc-clinically-relevant}

The reason for the quotes is that in my experience this term is used rather loosely in the literature. There is a serious misconception that the “clinically relevant” part of an operating characteristic is the steep portion emanating from the origin. The purpose of this section is to clarify this notion. One needs to go back the definition of the FROC, particularly the linear plot, Fig. 14.2, showing how the raw plot is generated as a virtual threshold is moved from the far right to the far left. While this plot applies to the AFROC, the essential idea is the same. One orders the LL marks (red dots) from left to right in increasing order according to their z-samples. Likewise, the NL marks (green dots) are also ordered from left to right in increasing order according to their z-samples. As the virtual threshold is moved to the left, starting from , mostly red dots and occasional green dots are crossed; each time a red dot is crossed the operating point moves up by  $1/(\text{total number of lesions})$  and each time a greed dot is crossed the operating point moves to the right by  $1/(\text{total number of cases})$ . This causes the operating point to rise, starting from the origin and move upward and to the right. The steep portion of the plot corresponds to crossings by LL and NL marks with high z-samples: it is the contribution of mostly easily visible lesions and the occasional NL. All observers are expected to localize the easy lesions, and there is nothing “clinically significant” about this. This argument applies to all operating characteristics. The clinical significance arises from the application. In a screening application, it is important to maintain high sensitivity at a reasonable specificity. In fact Jiang, Metz and Nishikawa<sup>30</sup> had it right when they proposed the area above a preselected high sensitivity threshold divided by . Such a measure would emphasize the upper right corner of the ROC curve, not the steep portion near the origin. In the screening context, most of the z-samples (99.5% to be more precise) are from non-diseased cases, and only 0.5% is from diseased cases. This implies the “clinically relevant” part of the plot is near the upper right corner of the ROC plot. With the FROC a

normalized area above a preselected cannot be defined. On the other hand, the AFROC is amenable to such a partial area measure as is, of course, the ROC.

To do it right, one needs to include the costs and benefits of correct and incorrect decisions on diseased and non-diseased cases, the prevalence of disease and the actual population distribution of the z-samples for non-diseased and diseased cases, and perform a weighted average over the entire ROC or AFROC curve. In the screening context, this would tend to weight the upper end of the curve. This is not an easy problem but it can be solved.

### 33.4 Discussion / Summary

This chapter has detailed ROC, FROC and AFROC curves predicted by the radiological search model (RSM). All RSM-predicted curves share the constrained end-point property that is qualitatively different from previous ROC models. In my experience, it is a property that most researchers in this field have difficulty accepting. There is too much history going back to the early 1940s, of the ROC curve extending from (0,0) to (1,1) that one has to let go of, and this can be difficult.

I am not aware of any direct evidence that radiologists can move the operating point continuously in the range (0,0) to (1,1) in search tasks, so the existence of such an ROC is tantamount to an assumption. Algorithmic observers that do not involve the element of search can extend continuously to (1,1). An example of an algorithmic observer not involving search is a diagnostic test that rates the results of a laboratory measurement, e.g., the A1C measure of blood glucose for presence of a disease. If A1C > 6.5% the patient is diagnosed as diabetic. By moving the threshold from infinity to -infinity, and assuming a large population of patients, one can trace out the entire ROC curve from the origin to (1,1). This is because every patient yields an A1C value. Now imagine that some finite fraction of the test results are “lost in the mail”; then the ROC curve, calculated over all patients, would have the constrained end-point property, albeit due to an unreasonable cause.

The situation in medical imaging involving search tasks is qualitatively different. Not every case yields a decision variable. There is a reasonable cause for this – to render a decision variable sample the radiologist must find something suspicious to report, and if none is found, there is no decision variable to report. The ROC curve calculated over all patients would exhibit the constrained end-point property, even in the limit of an infinite number of patients. If calculated over only those patients that yielded at least one mark, the ROC curve would extend from (0,0) to (1,1) but then one would be ignoring the cases with no marks, which represent valuable information: unmarked non-diseased cases represent perfect decisions and unmarked diseased cases represent worst-case decisions.

RSM-predicted ROC, FROC and AFROC curves were derived (wAFROC is

implemented in the Rjafroc). These were used to demonstrate that the FROC is a poor descriptor of performance. Since almost all work to date, including some by me 47,48, has used FROC curves to measure performance, this is going to be difficult for some to accept. The examples in Fig. 17.6 (A- F) and Fig. 17.7 (A-B) should convince one that the FROC curve is indeed a poor measure of performance. The only situation where one can safely use the FROC curve is if the two modalities produce curves extending over the same NLF range. This can happen with two variants of a CAD algorithm, but rarely with radiologist observers.

A unique feature is that the RSM provides measures of search and lesion-classification performance. It bears repeating that search performance is the ability to find lesions while avoiding finding non-lesions. Search performance can be determined from the position of the ROC end-point (which in turn is determined by RSM-based fitting of ROC data, Chapter 19). The perpendicular distance between the end-point and the chance diagonal is, apart from a factor of 1.414, a measure of search performance. All ROC models that predict continuous curves extending to (1,1), imply zero search performance.

Lesion-classification performance is measured by the AUC value corresponding to the parameter. Lesion-classification performance is the ability to discriminate between LLs and NLs, not between diseased and non-diseased cases: the latter is measured by RSM-AUC. There is a close analogy between the two ways of measuring lesion-classification performance and CAD used to find lesions in screening mammography vs. CAD used in the diagnostic context to determine if a lesion found at screening is actually malignant. The former is termed CADe, for CAD detection, which, in my opinion, is slightly misleading as at screening lesions are found not detected (“detection” is “discover or identify the presence or existence of something”, correct localization is not necessarily implied; the more precise term is “localize”). In the diagnostic context one has CADx, for CAD diagnostic, i.e., given a specific region of the image, is the region malignant?

Search and lesion-classification performance can be used as “diagnostic aids” to optimize performance of a reader. For example, if search performance is low, then training using mainly non-diseased cases is called for, so the resident learns the different variants of non-diseased tissues that can appear to be true lesions. If lesion-classification performance is low then training with diseased cases only is called for, so the resident learns the distinguishing features characterizing true lesions from non-diseased tissues that fake true lesions.

Finally, evidence for the RSM is summarized. Its correspondence to the empirical Kundel-Nodine model of visual search that is grounded in eye-tracking measurements. It reduces in the limit of large , which guarantees that every case will yield a decision variable sample, to the binormal model; the predicted pdfs in this limit are not strictly normal, but deviations from normality would require very large sample size to demonstrate. Examples were given where even with 1200 cases the binormal model provides statistically good fits, as judged

by the chi-square goodness of fit statistic, Table 17.2. Since the binormal model has proven quite successful in describing a large body of data, it satisfying that the RSM can mimic it in the limit of large . The RSM explains most empirical results regarding binormal model fits: the common finding that  $b < 1$ ; that  $b$  decreases with increasing lesion pSNR (large and / or ); and the finding that the difference in means divided by the difference in standard deviations is fairly constant for a fixed experimental situation, Table 17.3. The RSM explains data degeneracy, especially for radiologists with high expertise.

The contaminated binormal model2-4 (CBM), Chapter 20, which models the diseased distribution as having two peaks, one at zero and the other at a constrained value, also explains the empirical observation that b-parameter  $< 1$  and data degeneracy. Because it allows the ROC curve to go continuously to (1,1), CBM does not completely account for search performance – it accounts for search when it comes to finding lesions, but not for avoiding finding non-lesions.

I do not want to leave the impression that RSM is the ultimate model. The current model does not predict satisfaction of search (SOS) effects<sup>27-29</sup>. Attempts to incorporate SOS effects in the RSM are in the early research stage. As stated earlier, the RSM is a first-order model: a lot of interesting science remains to be uncovered.

### 33.4.1 The Wagner review

The two RSM papers<sup>12,13</sup> were honored by being included in a list of 25 papers the “Highlights of 2006” in Physics in Medicine and Biology. As stated by the publisher: ”I am delighted to present a special collection of articles that highlight the very best research published in Physics in Medicine and Biology in 2006. Articles were selected for their presentation of outstanding new research, receipt of the highest praise from our international referees, and the highest number of downloads from the journal website.

One of the reviewers was the late Dr. Robert (“Bob”) Wagner – he had an open-minded approach to imaging science that is lacking these days, and a unique writing style. I reproduce one of his comments with minor edits, as it pertains to the most interesting and misunderstood prediction of the RSM, namely its constrained end-point property.

I’m thinking here about the straight-line piece of the ROC curve from the max to (1, 1). 1. This can be thought of as resulting from two overlapping uniform distributions (thus guessing) far to the left in decision space (rather than delta functions). Please think some more about this point–because it might make better contact with the classical literature. 2. BTW – it just occurs to me (based on the classical early ROC work of Swets & co.) – that there is a test that can resolve the issue that I struggled with in my earlier remarks. The experimenter can try to force the reader to provide further data that will fill in the space above the max point. If the results are a straight line, then the reader

would just be guessing – as implied by the present model. If the results are concave downward, then further information has been extracted from the data. This could require a great amount of data to sort out—but it's an interesting point (at least to me).

Dr. Wagner made two interesting points. With his passing, I have been deprived of the penetrating and incisive evaluation of his ongoing work, which I deeply miss. Here is my response (ca. 2006):

The need for delta functions at negative infinity can be seen from the following argument. Let us postulate two constrained width pdfs with the same shapes but different areas, centered at a common value far to the left in decision space, but not at negative infinity. These pdfs would also yield a straight-line portion to the ROC curve. However, they would be inconsistent with the search model assumption that some images yield no decision variable samples and therefore cannot be rated in bin ROC:2 or higher. Therefore, if the distributions are as postulated above then choice of a cutoff in the neighborhood of the overlap would result in some of these images being rated 2 or higher, contradicting the RSM assumption. The delta function pdfs at negative infinity are seen to be a consequence of the search model.

One could argue that when the observer sees nothing to report then he starts guessing and indeed this would enable the observer to move along the dashed portion of the curve. This argument implies that the observer knows when the threshold is at negative infinity, at which point the observer turns on the guessing mechanism (the observer who always guesses would move along the chance diagonal). In my judgment, this is unreasonable. The existence of two thresholds, one for moving along the non-guessing portion and one for switching to the guessing mode would require abandoning the concept of a single decision rule. To preserve this concept one needs the delta functions at negative infinity.

Regarding Dr. Wagner's second point, it would require a great amount of data to sort out whether forcing the observer to guess would fill in the dashed portion of the curve, but I doubt it is worth the effort. Given the bad consequences of guessing (incorrect recalls) I believe that in the clinical situation, the radiologist will not knowingly guess. If the radiologist sees nothing to report, nothing will be reported. In addition, I believe that forcing the observer, to prove some research point, is not a good idea.

### 33.5 References

1. Chakraborty DP. Computer analysis of mammography phantom images (CAMPI): An application to the measurement of microcalcification image quality of directly acquired digital images. *Medical Physics*. 1997;24(8):1269-1277.

2. Chakraborty DP, Eckert MP. Quantitative versus subjective evaluation of mammography accreditation phantom images. *Medical Physics.* 1995;22(2):133-143.
3. Chakraborty DP, Yoon H-J, Mello-Thoms C. Application of threshold-bias independent analysis to eye-tracking and FROC data. *Academic Radiology.* 2012;In press.
4. Chakraborty DP. ROC Curves predicted by a model of visual search. *Phys Med Biol.* 2006;51:3463-3482.
5. Chakraborty DP. A search model and figure of merit for observer data acquired according to the free-response paradigm. *Phys Med Biol.* 2006;51:3449-3462.

# Chapter 34

## Analyzing FROC data

### 34.1 TBA How much finished

10%

### 34.2 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 R.Jafroc 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.

### 34.3 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-like<sup>2</sup> 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.]

## 34.4 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(plot1wAFROCPlot)` to `print(plot1wAFROCPoints)` 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(plot2wAFROCPlot)to print(plot2ROCPPlot)`, 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,  $p = 0.03054$ .

## 34.5 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.

## 34.6 Crossed-treatment analysis

This analysis was developed for a particular application<sup>6</sup> 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  $i$  ( $i = 1, 2, \dots, I$ ) and reader  $j$  ( $j = 1, 2, \dots, J$ ) 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  $\alpha = 0.05$ . 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.

### 34.7 Discussion / Summary

An IDL (Interactive Data Language, currently marketed by Exelis Visual Information Solutions, [www.exelisvis.com](http://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 sources<sup>7-9</sup>. 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 platform<sup>10</sup>. 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 analysis<sup>11</sup>, we close with a quote from respected sources<sup>12</sup> “(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.

## 34.8 References



# Chapter 35

## FROC sample size

### 35.1 TBA How much finished

10% TBA Merge the vignette into this ...

### 35.2 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  $\beta$ , typically chosen to be 20% - corresponding to 80% statistical power - and specified Type I error rate  $\alpha$ , 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 `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.

### 35.3 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 Fed-wAfroc.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-like<sup>2</sup> 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.

### 35.4 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 application<sup>6</sup> 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 ( $i = 1, 2, \dots, I$ ) and reader j ( $j = 1, 2, \dots, J$ ) 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  $\alpha = 0.05$ . 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.

### 35.5 FitRsmROC usage example

### 35.6 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.

### 35.7 References



# Chapter 36

## RSM fitting

### 36.1 TBA How much finished

10%

### 36.2 Introduction

The radiological search model (RSM) is based on what is known, via eye-tracking measurements, about how radiologists look at medical images (Kundel et al., 2007). The ability of this model to predict search and lesion-classification expertise was described in TBA Chapter 17. If one could estimate search and lesion-classification expertise from clinical datasets then one would know which of them is limiting performance. This would provide insight into the decision making efficiency of observers. For this potential to be realized, one has to be able to reliably estimate parameters of the RSM from data, and this turned out to be a difficult problem.

To put progress in this area in context a brief historical background is needed. I have worked on and off on the FROC estimation problem since 2002, and two persons (Dr. Hong-Jun Yoon and Xuetong Zhai) can attest to the effort. Initial attempts focused on fitting the FROC curve, in the (subsequently shown to be mistaken) belief that this was using *all* the data. In fact unmarked non-diseased cases, which are perfect decisions, are not taken into account in the FROC plot. In addition, there are degeneracy issues, which make parameter estimation difficult except in uninteresting situations. Early work involved maximization of the FROC likelihood function. This method was applied to seven designer-level CAD datasets. With CAD data one has a large number of marks and unmarked cases are relatively rare. However, only the CAD designer knows of their existence since in the clinic only a small fraction of the marks, those whose

$z$ -samples exceed a manufacturer-selected threshold, are actually shown to the radiologist. In other words the full FROC curve, extending to the end-point, is available to the CAD algorithm designer, which makes estimation of the end-point defining parameters  $\lambda'$ ,  $\nu'$  trivial. Estimating the remaining parameter of the RSM is then also relatively easy.

It was gradually recognized that the FROC curve based method worked only for designer level CAD data, and not for human observer data. Consequently, subsequent effort focused on ROC curve-based fitting, and this proved successful at fitting radiologist datasets, where detailed definition of the ROC curve is not available. A preliminary account of this work can be found in a conference proceeding (Chakraborty and Svahn, 2011).

*The reader should be surprised to read that the research eventually turned to ROC curve based fitting, which implies that one does not even need FROC data to estimate RSM parameters.* I have previously stated that the ROC paradigm ignores search, so how can one estimate search-model parameters from ROC data? The reason is that the *shape* of the ROC curve and the *position* of the upper-most observed operating point, depend on the RSM parameters, and this information can be used for a successful fitting method that is not susceptible to degeneracy<sup>1</sup>.

The chapter starts with fitting FROC curves. This is partly for historical reasons and to make contact with a method used by CAD designers. Then focus shifts to fitting ROC curves and comparing the RSM-based method to existing methods, namely the proper ROC (PROPROC) (Metz and Pan, 1999; Pan and Metz, 1997) and the contaminated binormal model (CBM) (Dorfman and Berbaum, 2000) methods, both of which are proper ROC fitting models. These are described in more detail in TBA Chapter 20. The comparison is based on a large number of interpretations, namely, 14 datasets comprising 43 modalities, 80 readers and 2012 cases, most of which are from my international collaborations. Besides providing further evidence for the validity of the RSM, the estimates of search and lesion-classification performance derived from the fitted parameters demonstrate that there is information in ROC data that is currently ignored by analyses that do not account for search performance. *Specifically, it shows that search performance is the bottleneck that is currently limiting radiologist performance.*

The ability to fit RSM to clinical datasets is critical to sample size estimation – this was the practical reason why the RSM fitting problem had to be solved. Sample size estimation requires relating the wAFROC-AUC FOM to the corresponding ROC-AUC FOM in order to obtain a physically meaningful effect-size. Lacking a mathematical relationship between them, comparing the effect-sizes in the two units would be like comparing “apples and oranges”. A mathematical relation is only possible if one has a parametric model that predicts both ROC

---

<sup>1</sup>Degenerate datasets are defined as those that do not provide any interior data points, i.e., all operating points lie on the edges of the ROC square, i.e., enclosed by the four lines defined by  $FPF = 0$  or  $1$  and  $TPF = 0$  or  $1$ .

and wAFROC curves, as does the RSM. Therefore, this chapter concludes with sample size estimation for FROC studies using the wAFROC FOM. However, as long as one can predict the appropriate operating characteristic using RSM parameters, the method can be extended to other paradigms, e.g., the location ROC (LROC) (Chakraborty and Yoon, 2008) paradigm.

### 36.3 FROC likelihood function

Recall that the likelihood function is the probability of observing the data as a function of the parameter values. FROC notation was summarized in TBA Table 13.1. Thresholds  $\vec{\zeta} \equiv (\zeta_0, \zeta_1, \dots, \zeta_{R_{FROC}+1},)$  were defined, where  $R_{FROC}$  is the number of FROC bins, and  $\zeta_0 = -\infty$  and  $\zeta_{R_{FROC}+1} = \infty$ . Since each z-sample is obtained by sampling an appropriately centered unit-variance normal distribution, the probability  $p_r$  that a latent NL will be marked and rated in FROC bin  $r$  and the probability  $q_r$  that a latent LL will be marked and rated in FROC bin  $r$  are given by:

$$\left. \begin{aligned} p_r &= \Phi(\zeta_{r+1}) - \Phi(\zeta_r) \\ q_r &= \Phi(\zeta_{r+1} - \mu) - \Phi(\zeta_r - \mu) \end{aligned} \right\} \quad (36.1)$$

Understanding these equations is easy: the CDF function evaluated at a threshold is the probability that a z-sample is less than the threshold. The first equation is the difference between the CDF functions of a unit-normal distribution evaluated at the two thresholds. This is the probability that the NL z-sample falls in bin FROC: $r$ . The second equation gives the probability that the LL z-sample falls in bin FROC: $r$ . The probabilities  $p_r$  and  $q_r$  individually sum to unity when all bins, including the zero bin, are included.

If NL and LL events are assumed independent, the contributions to the likelihood function can be separated, and one need not enumerate counts at the individual case-level; instead, in the description that follows, one enumerates NL and LL counts in the various bins over the whole dataset.

#### 36.3.1 Contribution of NLs

Define  $n$  (a random non-negative integer) as the total number of latent NLs in the dataset. The observed NL counts vector is  $\vec{n} \equiv (n_0, n_1, \dots, n_{R_{FROC}},)$ . Here  $n_r$  is the total number of NL counts in FROC ratings bin  $r$ ,  $n_0 = n - \sum_{r=1}^R n_r = n - N$ , is the *unknown number of unmarked latent NLs* and  $N$  is the total number of observed NLs in the dataset. The probability  $P(\vec{n} | n, \vec{\zeta})$  of observing the NL counts vector  $\vec{n}$  is (the factorials come from the multinomial distribution):

$$P(\vec{n} | n, \vec{\zeta}) = n! \prod_{r=0}^{R_{FROC}} \frac{p_r^{n_r}}{n_r!} \quad (36.2)$$

Since  $n$  is a random integer, the probability needs to be averaged over its Poisson distribution, i.e., one is calculating the expected value, yielding:

$$P(\vec{n} | \lambda', \vec{\zeta}) = \text{pmf}_{\text{Poi}}(n, K\lambda') P(\vec{n} | n, \vec{\zeta}) \quad (36.3)$$

In this expression  $K = K_1 + K_2$  is the total number of cases.  $\text{pmf}_{\text{Poi}}(n, K\lambda')$  of the Poisson distribution yields the probability of  $n$  counts from a Poisson distribution with mean  $K\lambda'$ . The multiplication by the total number of cases is required because one is counting the total number of latent NLs over the entire dataset. The lower limit on  $n$  is needed because  $n$  cannot be smaller than  $N$ , the total number of observed NL counts. The left hand side of Eqn. (36.3) is the probability of observing the NL counts vector  $\vec{n}$  as a function of RSM parameters. Not surprisingly, since NLs are sampled from a zero-mean normal distribution, the  $\mu$  parameter does not enter the above expression.

### 36.3.2 Contribution of LLs

Likewise, define  $l$  (a non-negative random integer) the total number of latent LLs in the dataset and the LL counts vector is  $\vec{l} \equiv (l_0, l_1, \dots, l_{R_{FROC}})$ . Here  $l_r$  is the number of LL counts in FROC ratings bin  $r$ ,  $l_0 = l - \sum_{r=1}^{R_{FROC}} l_r = l - L$  is the *known* number of unmarked latent LLs and  $L$  is the total number of observed LLs in the dataset. The probability  $P(\vec{l} | l, \mu, \vec{\zeta})$  of observing the LL counts vector  $\vec{l}$  is:

$$P(\vec{l} | l, \mu, \vec{\zeta}) = l! \prod_{r=0}^{R_{FROC}} \frac{q_r^{l_r}}{l_r!} \quad (36.4)$$

The above probability needs to be averaged over the binomial distribution of  $l$ :

$$P(\vec{l} | l, \mu, \nu', \vec{\zeta}) = \sum_{l=L}^{L_{tot}} \text{pmf}_{\text{Bin}}(l, L_T, \nu') P(\vec{l} | l, \mu, \vec{\zeta}) \quad (36.5)$$

In this expression  $L_{tot}$  is the total number of lesions in the dataset and the lower limit on  $l$  is needed because it cannot be smaller than  $L$ , the total number of observed LLs. Performing the two summations using Maple, multiplying the two probabilities and taking the logarithm yields the final expression for the log-likelihood function (Yoon et al., 2007):

$$LL_{FROC} \equiv LL_{FROC}(\vec{n}, \vec{l} | \mu, \lambda', \nu') = \sum_{r=1}^{R_{FROC}} \{n_r \log(p_r) + l_r \log(q_r)\} + N \log(\lambda') + L \log(\nu') - K \lambda' (1 - p_0) + (L_T - L) \log(1 - \nu') \quad (36.6)$$

### 36.3.3 Degeneracy problems

The product  $\lambda' (1 - p_0) = \lambda' \Phi(-\zeta_1)$  reveals degeneracy in the sense that two quantities appear as a product, so that they cannot be individually separated. The effect of increasing  $\lambda'$  can be counteracted by increasing  $\zeta_1$ ; increasing  $\lambda'$  yields more latent NLs but increasing  $\zeta_1$  results in fewer of them being marked. The two possibilities cannot be distinguished. A similar degeneracy occurs in the term involving the product  $-\nu' + \nu' q_0 = -\nu'(1 - q_0) = -\nu' \Phi(\mu - \zeta_1)$ , where increasing  $\nu'$  can be counter balanced by decreasing  $\mu - \zeta_1$ , i.e., by increasing  $\zeta_1$ . Again, the effect of increasing  $\nu'$  is to produce more latent LLs, but increasing  $\zeta_1$  results in fewer of them being marked.

*This is the fundamental problem with fitting RSM FROC curves to radiologist FROC data.*

## 36.4 IDCA Likelihood function

In the limit  $\zeta_1 \rightarrow -\infty$ ,  $p_0 \rightarrow 0$  and  $q_0 \rightarrow 0$ , and TBA Eqn. (18.6) reduces to:

$$LL_{FROC}^{IDCA} = \sum_{r=1}^{R_{FROC}} \{n_r \log(p_r) + l_r \log(q_r)\} + N \log(\lambda') + L \log(\nu') - K \lambda' + (L_T - L) \log(1 - \nu') \quad (36.7)$$

*Notice that in the limit  $\zeta_1 \rightarrow -\infty$  the degeneracy problems just described go away.*

The superscript IDCA comes from “*initial detection and candidate analysis*” (Edwards et al., 2002). All CAD algorithms consist of an *initial detection* stage, which identifies possible *lesion candidates*. In the second stage the algorithm analyzes each candidate lesion, *candidate analysis*, to get a probability of malignancy. If the probability of malignancy exceeds a threshold value selected by the CAD manufacturer, and this is accomplished based on a compromise between sensitivity and specificity, and see Chapter 39 for my solution to this problem, the location of each candidate lesion satisfying the criterion is shown to the radiologist, Fig. 36.1.

According to TBA Eqn. (17.30), in the limit  $\zeta_1 \rightarrow -\infty$  the end-point coordinates of the FROC curve represent estimates of  $\lambda', \nu'$  respectively:



Figure 36.1: A typical 4-view display of a patient mammogram with the CAD cues (the red arrows) turned on.

$$\left. \begin{array}{l} \lambda' = NLF_{max} \\ \nu' = LLF_{max} \end{array} \right\} \quad (36.8)$$

In other words, in this limit two of the three parameters of the RSM are trivially determined from the location of the observed end-point. Suppressing all parameter independent terms, the log-likelihood function, Eqn. (36.7), reduces to:

$$LL_{FROC}^{IDCA} = \sum_{r=1}^{R_{FROC}} \{n_r \log(p_r) + l_r \log(q_r)\} + \dots \quad (36.9)$$

Since the ignored terms in Eqn. (36.9) are independent of model parameters they do not affect the maximization. The equation contains only one parameter, namely  $\mu$ , which is implicit in the definition of  $q_r$ , Eqn. (36.1).

Eqn. (36.9) resembles the log-likelihood function for the binormal model, since, according to TBA Eqn. (6.37), the LL function for the binormal model with  $R_{FROC}$  bins, is <sup>2</sup>:

$$LL_{ROC} = \sum_{r=1}^{R_{FROC}} \{K_{1r} \log((\Phi(\zeta_{r+1}) - \Phi(\zeta_r))) + K_{2r} \log((\Phi(b\zeta_{r+1} - a) - \Phi(b\zeta_r - a)))\} \quad (36.10)$$

In this equation  $K_{1r}$  is the number of counts in bin  $r$  of an ROC study consisting of  $R_{FROC}$  bins. Define the unequal-variance binormal model versions of Eqn. (36.1) as follows:

$$\left. \begin{array}{l} p'_r = \Phi(\zeta_{r+1}) - \Phi(\zeta_r) \\ q'_r = \Phi(b\zeta_{r+1} - a) - \Phi(b\zeta_r - a) \end{array} \right\} \quad (36.11)$$

Here  $(a, b)$  are the parameters the unequal variance binormal model. Then Eqn. (36.10) becomes,

$$LL_{ROC} = \sum_{r=1}^{R_{FROC}} \{K_{1r} \log(p'_r) + K_{2r} \log(q'_r)\} \quad (36.12)$$

- With the identifications  $K_{1r} \rightarrow n_r$  and  $K_{2r} \rightarrow l_r$ , Eqn. (36.10) looks exactly like Eqn. (36.9). This implies that binormal ROC fitting method can be used to determine  $a$  and  $b$ . Notice that instead of fitting an equal

---

<sup>2</sup>The number of ROC bins exceeds the number of FROC bins by one.

variance binormal model to determine the remaining single remaining  $\mu$  parameter of the RSM, one is using an unequal-variance binormal model with two parameters,  $a$  and  $b$ . It turns out that the extra parameter helps. It gives some flexibility to the fitting curve to match the data.

- This method of fitting FROC data was well known to CAD researchers but was first formalized in (Edwards et al., 2002).
- Regard the NL marks as non-diseased “cases” ( $K_{1r} \rightarrow n_r$ ) and the LL marks as diseased “cases” ( $K_{2r} \rightarrow l_r$ ). Construct a pseudo-ROC counts table, analogous to TBA Table 4.1, where  $n_r$  is defined as the pseudo-FP counts in ratings bin  $r$ , and likewise,  $l_r$  is defined as the pseudo-TP counts in ratings bin  $r$ . The pseudo-ROC counts table has the same structure as the ROC counts table and can be fitted by the binormal model (or other alternatives).
- The pseudo-FP and pseudo-TP counts can be used to define pseudo-FPF and pseudo-TPF in the usual manner; the respective denominators are the total number of NL and LL counts, respectively. These probabilities define the pseudo-ROC operating points.
- The prefix “pseudo” is needed because one is regarding localized regions in a case as independent “cases”. Since the fitting algorithm assumes each rating is from an independent case, one is violating a basic assumption, but with CAD data it appears one can get away with it, because the method yields good fits, especially with the extra parameter.
- The fitted FROC curve is obtained by scaling (i.e., multiplying) the ROC curve along the y-axis by  $LLF_{max}$  and along the x-axis by  $NLF_{max}$ . The method is illustrated in Fig. 36.2.

Fig. 36.2: The IDCA method of fitting designer-level CAD FROC data. In the upper half of the figure, the y-axis of the pseudo-ROC is pseudo-TPF and the x-axis is pseudo-FPF. The method is illustrated for a dataset with four FROC bins. Regarding the NLs and LLs as non-diseased and diseased cases, respectively, one constructs a table similar to Table 4.1, but this time with only four ROC bins (i.e., three non-trivial operating points). This defines the four operating points, the filled circles, including the trivial one at the upper right corner, shown in the upper half of the plot. One fits the ratings counts data using, for example, the binormal model, yielding the continuous line (based on experience the unequal variance binormal model is needed; the equal variance model does not fit as well). In practice, the operating points will not fall exactly on the fitted line. Finally, one scales (or “stretches”, or multiplies) the y-axis by  $\nu'$ . Likewise, the x-axis is scaled by  $\lambda'$ . This yields the continuous line shown in the lower half of the figure. Upon adding the FROC operating points one finds that they are magically fitted by the line, which is a scaled replica of the ROC fit in the upper curve.

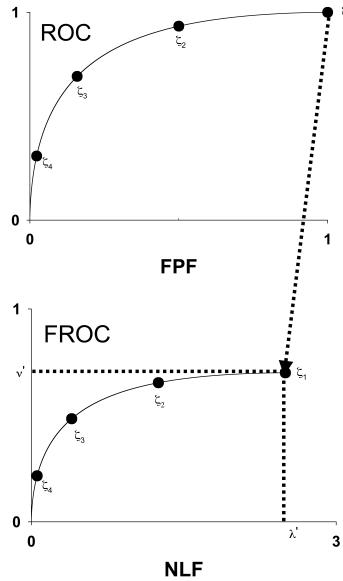


Figure 36.2: The IDCA method of fitting designer-level CAD FROC data.

Reference has already been made to the fact that it is necessary to assume  $\zeta_1 = -\infty$  in order to remove the degeneracy problem. This is also evident from the fact that the uppermost point in Fig. 36.2 is at (1,1). A point at the upper-right corner must correspond to  $\zeta_1 = -\infty$ , another confirmation of this assumption.

Assuming binormal fitting is employed, yielding parameters  $a$  and  $b$ , the equations defining the IDCA fitted FROC curve are, see TBA Eqn. (6.19) and Eqn. (6.20):

$$\left. \begin{aligned} NLF(\zeta) &= \lambda' \Phi(-\zeta) \\ LLF(\zeta) &= \nu' \Phi(a - b\zeta) \end{aligned} \right\} \quad (36.13)$$

The RSM predicted FROC curve is repeated below for convenience,

$$\left. \begin{aligned} NLF(\zeta) &= \lambda' \Phi(-\zeta) \\ LLF(\zeta) &= \nu' \Phi(\mu - \zeta) \end{aligned} \right\} \quad (36.14)$$

IDCA uses the *unequal variance* binormal model to fit the pseudo-ROC, which of course opens up the possibility of an inappropriate chance-line crossing and a predicted FROC curve that is non-monotonically increasing with NLF (this is always present with IDCA fits, but one would need to examine the curve near

the end-point very closely to see it). In practice the unequal variance model gives visually good fits for CAD datasets.

In fact, IDCA yields excellent fits to some designer-level FROC datasets. However, the issue is not with the quality of the fits, rather the appropriateness of the FROC curve as a measure of performance, especially for human observers. For CAD the method works, so if one wished one could use IDCA to fit designer level CAD FROC data. However, with closely spaced operating points, the empirical FROC would also work and it does not involve any fitting assumptions. The issue is not fitting designer level CAD data but comparing stand-alone performance of designer level CAD to radiologists, and this is not solved by IDCA, which works for designer level CAD, but not for human observers. The latter do not report every suspicious region, no matter how low its confidence level, so the IDCA assumption  $\zeta_1 \rightarrow -\infty$  is invalid. The problem of analyzing standalone performance of CAD against a group of radiologists interpreting the same cases is addressed in TBA Chapter 22.

### 36.5 ROC Likelihood function

The second attempt used the ROC likelihood function. In TBA Chapter 17 expressions were derived for the coordinates  $(x,y)$  of the ROC curve predicted by the RSM, see Eqn. (17.8) and Eqn. (17.16).

$$\text{FPF}(\zeta, \lambda') = 1 - \exp\left(-\frac{\lambda'}{2} \left[1 - \text{erf}\left(\frac{\zeta}{\sqrt{2}}\right)\right]\right) \quad (36.15)$$

$$y \equiv y(\zeta, \mu, \lambda', \nu', \bar{f}_L) = 1 - \sum_{L=1}^{L_{max}} f_L \times \left[ 1 - \left( 1 - \frac{\nu'}{2} + \frac{\nu'}{2} \text{erf}\left(\frac{\zeta - \mu}{\sqrt{2}}\right) \right)^L \exp\left(-\frac{\lambda'}{2} \left[1 - \text{erf}\left(\frac{\zeta}{\sqrt{2}}\right)\right]\right) \right] \quad (36.16)$$

Let  $(F_r, T_r)$  denote the number of false positives and true positives, respectively, in ROC rating bin  $r$  defined by thresholds  $[\zeta_r, \zeta_{r+1})$ , for  $r = 0, 1, \dots, R_{FROC}$ . The range of  $r$  shows explicitly that  $R_{FROC}$  FROC ratings correspond to  $R_{FROC} + 1$  ROC bins<sup>3</sup>. Note that  $(F_0, T_0)$  represent the *known* numbers of non-diseased and diseased cases, respectively, with no marks,  $(F_1, T_1)$  represent the numbers of non-diseased and diseased cases, respectively, with highest rating equal to one, etc. The probability  $P_{1r}$  of a count in non-diseased ROC bin  $r$  is<sup>4</sup>:

---

<sup>3</sup>The rating bookkeeping can be confusing. Basically,  $r = 0$  corresponds to unmarked cases,  $r = 1$  corresponds to cases where the highest rated FROC mark was rated 1, etc., and  $r = R_{FROC}$  corresponds to cases where the highest rated FROC mark was rated  $R_{FROC}$ .

<sup>4</sup>One needs to subtract the CDF evaluated at  $r+1$  from that at  $r$ ; the CDF is the complement of  $x$ , which results in the reversal. It should also make sense because the higher indexed  $x$  is to the right of the lower indexed one. Recall that the operating points are numbered starting from the top-right and working down.

$$P_{1r} = x(\zeta_r) - x(\zeta_{r+1}) \quad (36.17)$$

Likewise, the probability  $P_{2r}$  of a count in diseased ROC bin  $r$  is:

$$P_{2r} = y(\zeta_r) - y(\zeta_{r+1}) \quad (36.18)$$

The likelihood function is, ignoring combinatorial factors that do not depend on parameters:

$$(P_{1r})^{F_r} (P_{2r})^{T_r}$$

The log-likelihood function is:

$$LL_{ROC}(\mu, \lambda', \nu', \vec{f}_L) = \sum_{r=0}^{R_{FROC}} [F_r \log(P_{1r}) + T_r \log(P_{2r})] \quad (36.19)$$

The area  $AUC_{ROC}^{RSM}(\mu, \lambda', \nu', \vec{f}_L)$  under the parametric RSM-ROC curve was obtained by numerical integration:

$$AUC_{ROC}^{RSM}(\mu, \lambda', \nu', \vec{f}_L) = \int_{x=0}^1 y(\mu, \lambda', \nu', \vec{f}_L) dx \quad (36.20)$$

The total number of parameters to be estimated, including the  $R_{FROC}$  thresholds, is  $3 + R_{FROC}$ . Maximizing the likelihood function yields parameter estimates. The Broyden–Fletcher–Goldfarb–Shanno (BFGS) (Shanno and Kettler, 1970; Shanno, 1970; Goldfarb, 1970; Fletcher, 1970, 2013; Broyden, 1970) minimization algorithm, as implemented as function `mle2()` in the R-package `bbmle` (Bolker and R Development Core Team, 2020) was used to minimize the negative of the likelihood function. Since the BFGS algorithm varies each parameter in an unrestricted range  $(-\infty, \infty)$ , which would cause problems (e.g., RSM physical parameters cannot be negative and thresholds need to be properly ordered), appropriate variable transformations (both “forward” and “inverse”) were used so that parameters supplied to the log-likelihood function were always in the valid range, irrespective of values chosen by the BFGS algorithm.

The algorithm calculates the goodness of fit statistic using the method described in TBA §6.4.2. Because of the additional parameter, the degrees-of-freedom (df) of the chisquare goodness of fit statistic is  $R_{FROC}-3$ . One can appreciate that calculating goodness of fit for the RSM can fail in situations, where the corresponding statistic can be calculated for binormal model, e.g., three (non-trivial) ROC operating points, corresponding to  $df = 1$ . With FROC data one

needs at least four (non – trivial) ROC operating points, each defined by bins with at least five counts in both non-diseased and diseased categories.<sup>5</sup>

## 36.6 FitRsmROC implementation

The `RJafroc` function `FitRsmROC()` fits an RSM-predicted ROC curve to a binned single-modality single-reader ROC dataset. It is called by `ret <- FitRsmROC(binnedRocData, lesDistr, trt = 1, rdr = 1)`, where `binnedRocData` is a binned ROC dataset, `lesDistr` is the lesion distribution vector (normalized histogram) in the dataset and `trt` and `rdr` are the desired treatment and reader to extract from the dataset, each of which defaults to one.

The return value `ret` is a `list` with the following elements:

- `ret$mu` The mean of the diseased distribution relative to the non-diseased one
- `ret$\lambda` The Poisson parameter describing the distribution of latent NLs per case
- `ret$\nu` The binomial success probability describing the distribution of latent LLs per diseased case
- `ret$zetas` The RSM cutoffs, zetas or thresholds
- `ret$AUC` The RSM fitted ROC-AUC
- `ret$StdAUC` The standard deviation of AUC
- `ret$NLLIni` The initial value of negative LL
- `ret$NLLFin` The final value of negative LL
- `ret$ChisqrFitStats` The chisquare goodness of fit results
- `ret$covMat` The covariance matrix of the parameters
- `ret$fittedPlot` A `ggplot2` object containing the fitted operating characteristic along with the empirical operating points. Use `print` to display the object

---

<sup>5</sup>With three operating points, each defined by bins with at least five counts in both non-diseased and diseased categories, the number of usable ROC bins is four. Subtracting three one gets  $df = 1$ , and the statistic can be calculated. However, because of the extra RSM parameter, the corresponding  $df = 0$ .

## 36.7 FitRsmROC usage example

- The following example uses the *first* treatment of the “FED” dataset, `dataset04`, which is a 5 treatment 4 radiologist FROC dataset acquired by Dr. Federica Zanca et. al. (Zanca et al., 2009). The dataset has 5 treatments and 4 readers and 200 cases and was acquired on a 5-point integer scale, i.e., it is already binned. If not one needs to bin the dataset using `DfBinDataset()`. I need to emphasize this point: **if the dataset represents continuous ratings, as with a CAD algorithm, one must bin the dataset to (ideally) about 5-6 bins**. The number of parameters that must be estimated increases with the number of bins (because for each additional bin one needs to estimate an additional cutoff parameter).

```
rocData <- DfFroc2Roc(dataset04)
lesDistr <- UtilLesionDistr(dataset04)[,2]
ret <- FitRsmRoc(rocData, lesDistr = lesDistr)
```

The lesion distribution vector is 0.69, 0.2, 0.11. This means that fraction 0.69 of abnormal cases contain one lesion each, fraction 0.2 contain two lesions each and fraction 0.11 contain three lesions each. The fitting algorithm needs to know the distribution of lesions per case, as the fitted curve depends on this distribution. For example, all else being equal, if all abnormal cases contain one lesion, the ROC curve will be lower than if all abnormal cases contain three lesions. With increased number of lesions per case TPF increases, as there is greater chance that at least one the lesions will be marked.

The fitted parameter values are as follows (all cutoffs excepting  $\zeta_1$ , the chi-square statistic (NA for this dataset) and the covariance matrix are not shown):

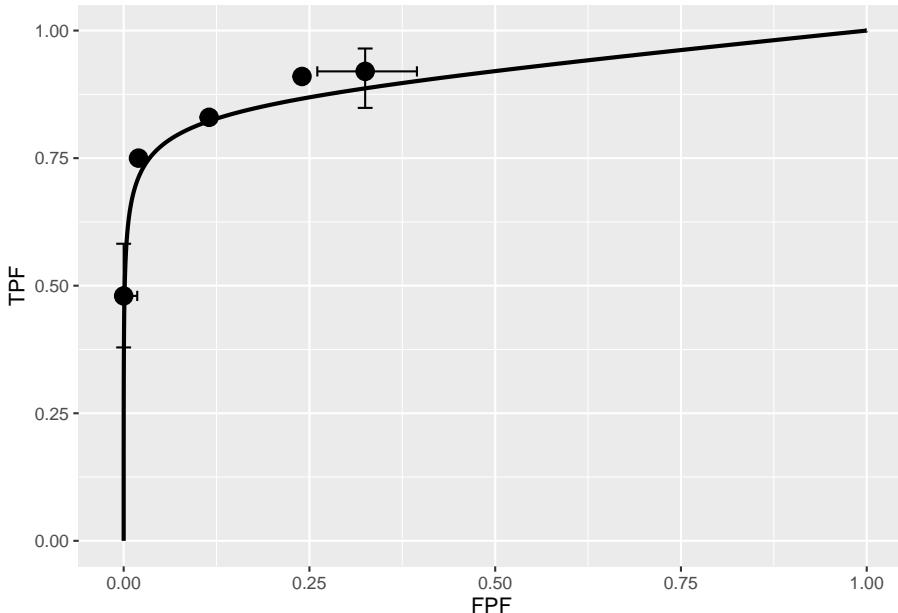
- $\mu = 3.6551363$
- $\lambda' = 9.8734529$
- $\nu' = 0.7963126$
- $\zeta_1 = 1.5006824$
- $AUC = 0.9065157$
- $\sigma(AUC) = 0.0231988$
- $NLLIni = 281.4024966$
- $NLLFin = 267.2673434$

The meaning of the parameters is as follows. The separation parameter  $\mu$  is 3.66. The relatively large separation will result in good classification performance. The large  $\lambda'$  parameter means that on the average the observer generates 9.87 latent NL marks per image. However, because of the relatively large value of  $\zeta_1$ , i.e., 1.5, only fraction 0.067 of these are actually marked, resulting in 0.66 actual marks per image. The fitting program decreased the negative of

the log-likelihood function from 281.4024966 to 267.2673434. A decrease in negative log-likelihood is equivalent to an increase in the likelihood, which is as expected, as the function maximizes the log-likelihood. Because the RSM contains 3 parameters, which is one more than other ROC models, the chisquare goodness of fit statistic usually cannot be calculated, except for large datasets - the criterion of 5 counts in each bin for true positives and false positives is usually hard to meet.

Shown next is the fitted plot. Error bars (exact 95% confidence intervals) are only shown for the lowest and highest operating points.

```
print(ret$fittedPlot)
```



The fitted ROC curve is proper: it's slope decreases monotonically as one moves up the curve thereby ruling out hooks such as are predicted by the binormal model. The area under the proper ROC is 0.907 which will be shown in a subsequent chapter to be identical to that yielded by other proper ROC fitting methods and higher than the binormal model fitted value.

## 36.8 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.

### 36.9 References



# Chapter 37

## Three proper ROC fits

### 37.1 TBA How much finished

75%

### 37.2 Introduction

A proper ROC curve is one whose slope decreases monotonically as the operating point moves up the curve, a consequence of which is that a proper ROC does not display an inappropriate chance line crossing followed by a sharp upward turn, i.e., a “hook”, usually near the (1,1) upper right corner.

There are three methods for fitting proper curves to ROC datasets:

- The radiological search model (RSM) described in Chapter 36,
- The PROPROC (proper ROC) and CBM (contaminated binormal model) described in TBA Chapter 20.

This chapter compares these methods for a number of datasets. Comparing the RSM to the binormal model would be inappropriate, as the latter does not predict proper ROCs.

- Both RSM and CBM are implemented in `RJafroc`.
- PROPROC is implemented in Windows software <sup>1</sup> available here, last accessed 1/4/21.

---

<sup>1</sup>OR DBM-MRMC 2.5, Sept. 04, 2014; this version, used in this chapter, is no longer distributed but is available from me upon request.

### 37.3 Applications

The RSM, PROPROC and CBM algorithms were applied to the 14 embedded datasets described in 37.11. The datasets have already been analyzed and the location of pre-analyzed results files are in 37.13.

```
datasetNames <- c("TONY", "VD", "FR",
                  "FED", "JT", "MAG",
                  "OPT", "PEN", "NICO",
                  "RUS", "DOB1", "DOB2",
                  "DOB3", "FZR")
```

In the following we focus on just two ROC datasets, which have been widely used in the literature to illustrate ROC methodological advances, namely the Van Dyke (VD) and the Franken (FR) datasets.

#### 37.3.1 Application to two datasets

- The code uses the function `Compare3ProperRocFits()`.
- The code file is `R/compare-3-fits/Compare3ProperRocFits.R`.
- `startIndx` is the first index to analyze and `endIndx` is the last.
- In the current example `startIndx = 2` and `endIndx = 3`; i.e., two datasets are analyzed corresponding to `datasetNames[2]` and `datasetNames[3]`, i.e., the VD and FR datasets.<sup>2</sup>
- `reAnalyze` is set to `FALSE` causing pre-analyzed results (to be found in directory `R/compare-3-fits/RSM6`) to be retrieved. If `reAnalyze` is `TRUE` the analysis is repeated, leading to possibly slightly different results (the maximum likelihood parameter-search algorithm has inherent randomness aimed at avoiding finding false local maxima).
- The fitted parameter results are contained in `ret$allResults` and the *composite plots* (i.e., 3 combined plots corresponding to the three proper ROC fitting algorithms) are contained in `ret$allPlots`.
- These are saved to lists `plotArr` and `resultsArr`.

```
startIndx <- 2
endIndx <- 3
ret <- Compare3ProperRocFits(datasetNames,
                               startIndx = startIndx,
                               endIndx = endIndx,
                               reAnalyze = FALSE)

resultsArr <- plotArr <- array(list(),
```

---

<sup>2</sup>To analyze all datasets one sets `startIndx <- 1` and `endIndx <- 14`.

```

dim = c(endIndx - startIndx + 1))

for (f in 1:(endIndx-startIndx+1)) {
  plotArr[[f]] <- ret$allPlots[[f]]
  resultsArr[[f]] <- ret$allResults[[f]]
}

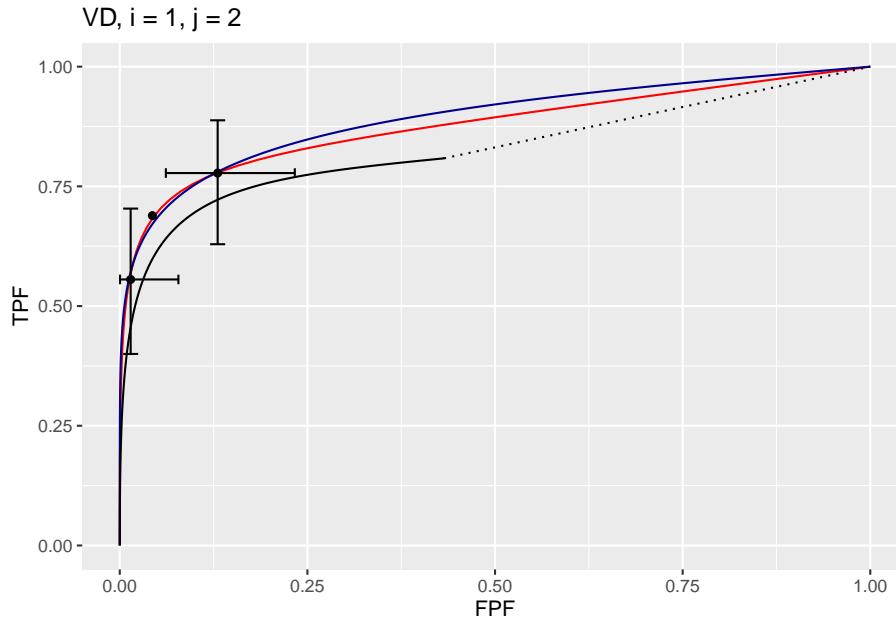
```

We show next how to display the composite plots.

## 37.4 Displaying composite plots

- The `plotArr` list contains plots for the two datasets. The Van Dyke plots are in `plotArr[[1]]` and the Franken in `plotArr[[2]]`.
- The Van Dyke plots contain  $I \times J = 2 \times 5 = 10$  composite plots, and similarly for the Franken dataset (both datasets consist of 2 treatments and 5 readers).
- The following shows how to display the composite plot for the Van Dyke dataset for treatment 1 and reader 2.

```
plotArr[[1]][[1,2]]
```

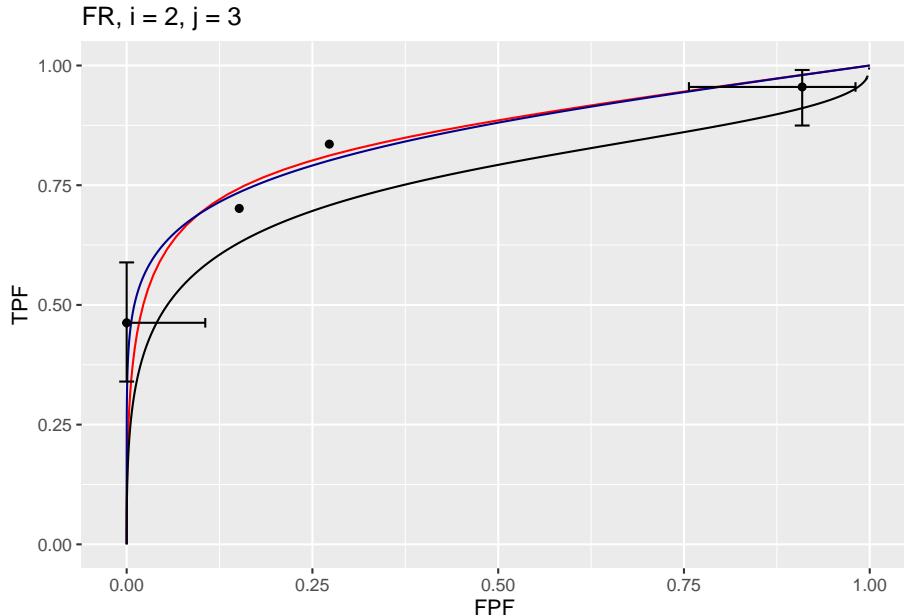


The plot is labeled **D2, i = 1, j = 2**, meaning the second dataset, the first treatment and the second reader. It contains 3 curves:

- The RSM fitted curve is in black. It is the only one with a dotted line connecting the uppermost continuously accessible operating point to (1,1).
- The PROPROC fitted curve is in red.
- The CBM fitted curve is in blue.
- Three operating points from the binned data are shown as well as exact 95% confidence intervals for the lowest and uppermost operating points.

The following example displays the composite plots for the Franken dataset, treatment 2 and reader 3:

```
plotArr[[2]][[2,3]]
```



Shown next is how to display the parameters corresponding to the fitted curves.

### 37.5 Displaying RSM parameters

The RSM has parameters:  $\mu$ ,  $\lambda'$ ,  $\nu'$  and  $\zeta_1$ . The parameters corresponding to the RSM plots are accessed as shown next.

- `resultsArr[[1]][[2]]$retRsm$mu` is the RSM  $\mu$  parameter for dataset 1 (i.e., Van Dyke dataset), treatment 1 and reader 2,
- `resultsArr[[1]][[2]]$retRsm$lambdaP` is the corresponding  $\lambda'$  parameter, and

- `resultsArr[[1]][[2]]$retRsm$nuP` is the corresponding  $\nu'$  parameter.
- `resultsArr[[1]][[2]]$retRsm$\zeta_1` is the corresponding  $\zeta_1$  parameter.
- Treatment 2 and reader 1 values would be accessed as `resultsArr[[1]][[6]]$retRsm$mu`, etc.
- More generally the values are accessed as `[[f]][[(i-1)*J+j]]`, where `f` is the dataset index, `i` is the treatment index, `j` is the reader index and `J` is the total number of readers.
- For the Van Dyke dataset `f = 1` and for the Franken dataset `f = 2`.

The first example displays RSM parameters for the Van Dyke dataset, treatment 1 and reader 2:

```
f <- 1; i <- 1; j <- 2; J <- 5
cat("RSM parameters, Van Dyke Dataset, treatment 1, reader 2:",
"\nmu = ",      resultsArr[[f]][[(i-1)*J+j]]$retRsm$mu,
"\nlambdaP = ",  resultsArr[[f]][[(i-1)*J+j]]$retRsm$lambdaP,
"\nnuP = ",      resultsArr[[f]][[(i-1)*J+j]]$retRsm$nuP,
"\nzeta_1 = ",    as.numeric(resultsArr[[f]][[(i-1)*J+j]]$retRsm$\zetas[1]),
"\nAUC = ",       resultsArr[[f]][[(i-1)*J+j]]$retRsm$AUC,
"\nsigma_AUC = ", as.numeric(resultsArr[[f]][[(i-1)*J+j]]$retRsm$StdAUC),
"\nNLLini = ",   resultsArr[[f]][[(i-1)*J+j]]$retRsm$NLLIni,
"\nNLLfin = ",   resultsArr[[f]][[(i-1)*J+j]]$retRsm$NLLFin)

## RSM parameters, Van Dyke Dataset, treatment 1, reader 2:
## mu =  2.201413
## lambdaP =  0.2569453
## nuP =  0.7524016
## zeta_1 = -0.1097901
## AUC =  0.8653694
## sigma_AUC =  0.04740562
## NLLini =  96.48516
## NLLfin =  85.86244
```

The next example displays RSM parameters for the Franken dataset, treatment 2 and reader 3:

```
f <- 2; i <- 2; j <- 3; J <- 5

## RSM parameters, Franken dataset, treatment 2, reader 3:
## mu =  3.287996
## lambdaP =  9.371198
## nuP =  0.7186006
## zeta_1 =  1.646943
```

```
## AUC = 0.8234519
## sigma_AUC = 0.04054005
## NLLini = 128.91
## NLLfin = 122.4996
```

The first four values are the fitted values for the RSM parameters  $\mu$ ,  $\lambda'$ ,  $\nu'$  and  $\zeta_1$ . The next value is the AUC under the fitted RSM curve followed by its standard error. The last two values are the initial and final values of negative log-likelihood <sup>3</sup>.

## 37.6 Displaying CBM parameters

CBM has parameters  $\mu_{CBM}$ ,  $\alpha$  and  $\zeta_1$ . The next example displays CBM parameters and AUC etc. for the Van Dyke dataset, treatment 1 and reader 2:

```
f <- 1;i <- 1; j <- 2;J <- 5
cat("CBM parameters, Van Dyke Dataset, treatment 1, reader 2:",
"\nmu = ",           resultsArr[[f]][[(i-1)*J+j]]$retCbm$mu,
"\nalpha = ",          resultsArr[[f]][[(i-1)*J+j]]$retCbm$alpha,
"\nzeta_1 = ",         as.numeric(resultsArr[[f]][[(i-1)*J+j]]$retCbm$zetas[1]),
"\nAUC = ",            resultsArr[[f]][[(i-1)*J+j]]$retCbm$AUC,
"\nsigma_AUC = ",      as.numeric(resultsArr[[f]][[(i-1)*J+j]]$retCbm$StdAUC),
"\nNLLini = ",          resultsArr[[f]][[(i-1)*J+j]]$retCbm$NLLIni,
"\nNLLfin = ",          resultsArr[[f]][[(i-1)*J+j]]$retCbm$NLLFin)

## CBM parameters, Van Dyke Dataset, treatment 1, reader 2:
## mu = 2.745791
## alpha = 0.7931264
## zeta_1 = 1.125028
## AUC = 0.8758668
## sigma_AUC = 0.03964492
## NLLini = 86.23289
## NLLfin = 85.88459
```

The next example displays CBM parameters for the Franken dataset, treatment 2 and reader 3:

```
f <- 2;i <- 2; j <- 3;J <- 5
```

---

<sup>3</sup>The initial value is calculated using initial estimates of parameters and the final value is that resulting from the log-likelihood maximization procedure. Since negative log-likelihood is being *minimized*, the final value is smaller than the initial value.

```
## CBM parameters, Franken dataset, treatment 2, reader 3:
## mu = 2.533668
## alpha = 0.6892561
## zeta_1 = 0.3097191
## AUC = 0.8194009
## sigma_AUC = 0.03968962
## NLLini = 122.6812
## NLLfin = 122.5604
```

The first three values are the fitted values for the CBM parameters  $\mu$ ,  $\alpha$  and  $\zeta_1$ . The next value is the AUC under the fitted CBM curve followed by its standard error. The last two values are the initial and final values of negative log-likelihood.

## 37.7 Displaying PROPROC parameters

PROPROC displayed parameters are  $c$  and  $d_a$ . The next example displays PROPROC parameters for the Van Dyke dataset, treatment 1 and reader 2:

```
f <- 1; i <- 1; j <- 2; J <- 5
cat("PROPROC parameters, Van Dyke Dataset, treatment 1, reader 2:",
"\nc = ",      resultsArr[[f]][[(i-1)*J+j]]$c1,
"\nd_a = ",    resultsArr[[f]][[(i-1)*J+j]]$da,
"\nAUC = ",    resultsArr[[f]][[(i-1)*J+j]]$aucProp)

## PROPROC parameters, Van Dyke Dataset, treatment 1, reader 2:
## c = -0.2809004
## d_a = 1.731472
## AUC = 0.8910714
```

The values are identical to those listed for treatment 1 and reader 2 in Fig. 37.7. Other statistics, such as standard error of AUC, are not provided by PROPROC software.

The next example displays PROPROC parameters for the Franken dataset, treatment 2 and reader 3:

```
f <- 2; i <- 2; j <- 3; J <- 5

## PROPROC parameters, Franken dataset, treatment 2, reader 3:
## c = -0.4420007
## d_a = 0.9836615
## AUC = 0.8252824
```

All 10 composite plots for the Van Dyke dataset are shown in the Appendix to this chapter, 37.14.

The next section provides an overview of the most salient findings from analyzing the datasets.

## 37.8 Overview of findings

With 14 datasets the total number of individual modality-reader combinations is 236: in other words, there are 236 datasets to each of which three algorithms were applied. It is easy to be overwhelmed by the numbers and this section summarizes the most important conclusion: *for each dataset, treatment and reader, the three fitting methods are consistent with a single method-independent AUC.*

If the AUCs of the three methods are identical the following relations hold with slopes equal to unity:

$$\left. \begin{array}{l} AUC_{PRO} = m_{PR} AUC_{PRO} \\ AUC_{CBM} = m_{CR} AUC_{PRO} \\ m_{PR} = 1 \\ m_{CR} = 1 \end{array} \right\} \quad (37.1)$$

The abbreviations are as follows:

- PRO = PROPROC
- PR = PROPROC vs. RSM
- CR = CBM vs. RSM.

For each dataset the plot of PROPROC AUC vs. RSM AUC should be linear with zero intercept and slope  $m_{PR}$ . The reason for the *zero intercept* is that if one of the AUCs indicates zero performance the other AUC must also be zero. Likewise, chance level performance (AUC = 0.5) must be common to all method of estimating AUC. Finally, perfect performance must be common to all methods. All of these conditions require a zero-intercept linear fit.

### 37.8.1 Slopes

Denote PROPROC AUC for dataset  $f$ , treatment  $i$  and reader  $j$  by  $\theta_{fij}^{PRO}$ . Likewise, the corresponding RSM and CBM values are denoted by  $\theta_{fij}^{RSM}$  and  $\theta_{fij}^{CBM}$ , respectively. For a given dataset the slope of the PROPROC values vs. the RSM values is denoted  $m_{PR,f}$ . The (grand) average over all datasets

is denoted  $m_{\bullet}^{PR}$ . Likewise, the average of the CBM AUC values vs. the RSM value is denoted  $m_{\bullet}^{CR}$ .

An analysis was conducted to determine the average slopes and a bootstrap analysis was conducted to determine the corresponding confidence intervals.

The code for calculating the average slopes is in `R/compare-3-fits/slopesConvVsRsm.R` and that for calculating the bootstrap confidence intervals is in `R/compare-3-fits/slopesAucsConvVsRsmCI.R`.

```
ret <- slopesConvVsRsm(datasetNames)
retCI <- slopesAucsConvVsRsmCI(datasetNames)
```

The call to function `slopesConvVsRsm()` returns `ret`, which contains, for each of 14 datasets, two plots and two slopes. For example:

- `ret$p1[[2]]` is the plot of  $\theta_{2ij}^{PRO}$  vs.  $\theta_{2ij}^{RSM}$  for the Van Dyke dataset.
- `ret$p2[[2]]` is the plot of  $\theta_{2ij}^{CBM}$  vs.  $\theta_{2ij}^{RSM}$  for the Van Dyke dataset.
- `ret$m_pro_rsm` has two columns, each of length 14, the slopes  $m_{PR,f}$  for the datasets (indexed by `f`) and the corresponding  $R^2$  values. The first column is `ret$m_pro_rsm[[1]]` and the second is `ret$m_pro_rsm[[2]]`.
- `ret$m_cbm_rsm` has two columns, each of length 14, the slopes  $m_{CR,f}$  for the datasets and the corresponding  $R^2$  values.

Likewise,

- `ret$p1[[3]]` is the plot of  $\theta_{3ij}^{PRO}$  vs.  $\theta_{3ij}^{RSM}$  for the Franken dataset.
- `ret$p2[[3]]` is the plot of  $\theta_{3ij}^{CBM}$  vs.  $\theta_{3ij}^{RSM}$  for the Franken dataset.

As examples, `ret$p1[[2]]` is the plot of  $\theta_{2ij}^{PRO}$  vs.  $\theta_{2ij}^{RSM}$  for the Van Dyke dataset and `ret$p1[[3]]` is the plot of  $\theta_{2ij}^{CBM}$  vs.  $\theta_{2ij}^{RSM}$  for the Van Dyke dataset, shown next. Each plot has the constrained linear fit superposed on the data points; each data point represents a distinct modality-reader combination.

The next plot shows corresponding plots for the Franken dataset.

The average slopes and  $R^2$  values ( $R^2$  is the fraction of variance explained by the constrained straight line fit) are listed in Table 37.1.

The slopes and  $R^2$  values for the Van Dyke dataset are shown next:

```
##          m-PR      R2-PR      m-CR      R2-CR
##  VD 1.006127 0.999773 1.000699 0.9999832
```

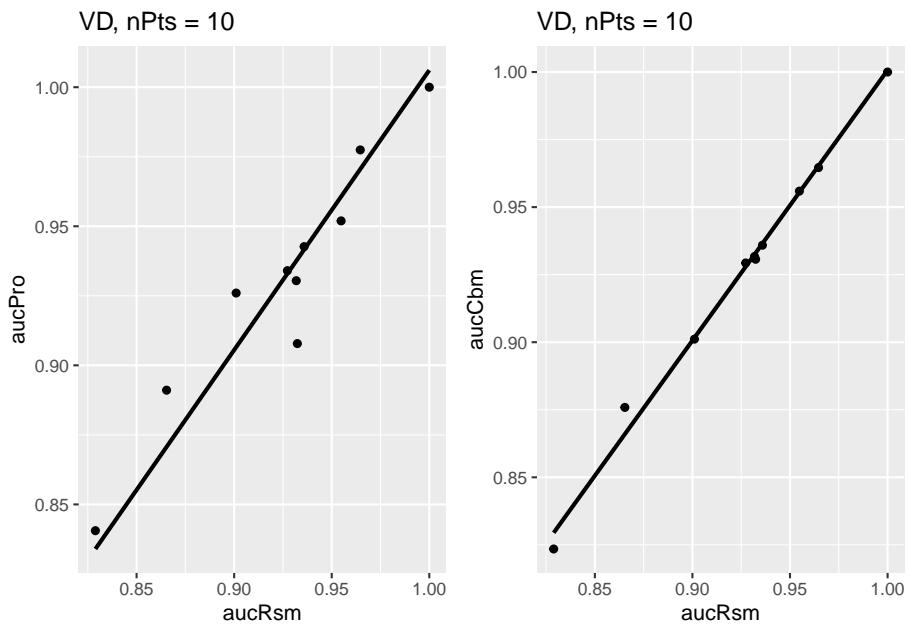


Figure 37.1: Van Dyke dataset: Left plot is PROPROC-AUC vs. RSM-AUC with the superposed constrained linear fit. The number of data points is  $n_{\text{Pts}} = 10$ . Right plot is CBM-AUC vs. RSM-AUC.

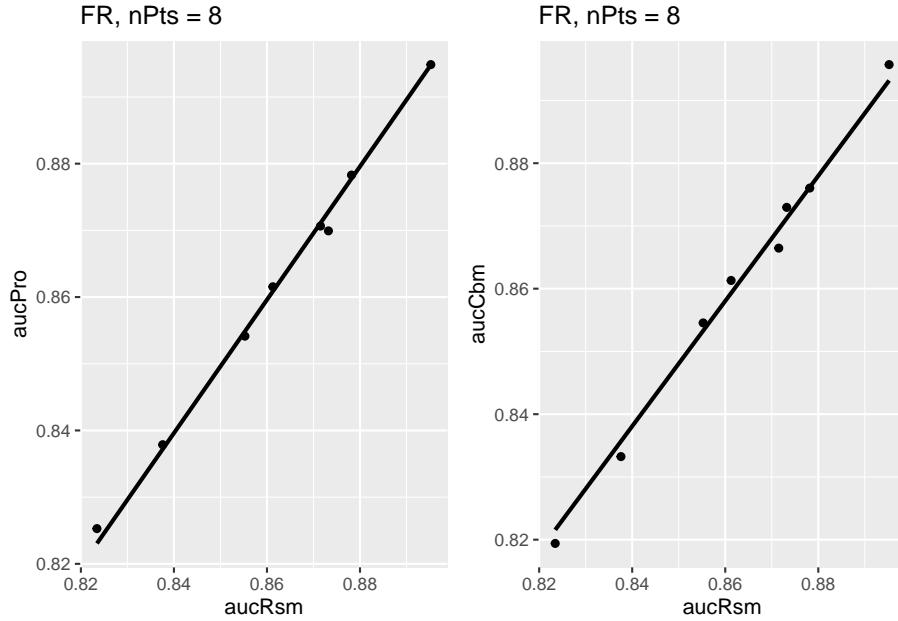


Figure 37.2: Similar to previous plot, for Franken dataset.

### 37.8.2 Confidence intervals

The call to `slopesAucsConvVsRsmCI` returns `retCI`, containing the results of the bootstrap analysis (note the bullet symbols • denoting averages over 14 datasets):

- `retCI$cislopeProRsm` confidence interval for  $m_{\bullet}^{PR}$
- `retCI$cislopeCbmRsm` confidence interval for  $m_{\bullet}^{CR}$
- `retCI$histSlopeProRsm` histogram plot for 200 bootstrap values of  $m_{\bullet}^{PR}$
- `retCI$histSlopeCbmRsm` histogram plot for 200 bootstrap values of  $m_{\bullet}^{CR}$
- `retCI$ciAvgAucRsm` confidence interval for 200 bootstrap values of  $\theta_{\bullet}^{RSM}$
- `retCI$ciAvgAucPro` confidence interval for 200 bootstrap values of  $\theta_{\bullet}^{PRO}$
- `retCI$ciAvgAucCbm` confidence interval for 200 bootstrap values of  $\theta_{\bullet}^{CBM}$

As examples,

```
##          m-PR      m-CR
## 2.5% 1.005092 0.9919886
## 97.5% 1.012285 0.9966149
```

The CI for  $m_{\bullet}^{PR}$  is slightly above unity, while that for  $m_{\bullet}^{CR}$  is slightly below. Shown next is the histogram plot for  $m_{\bullet}^{PR}$  (left plot) and  $m_{\bullet}^{CR}$  (right plot). Quantiles of these histograms were used to compute the confidence intervals cited above.

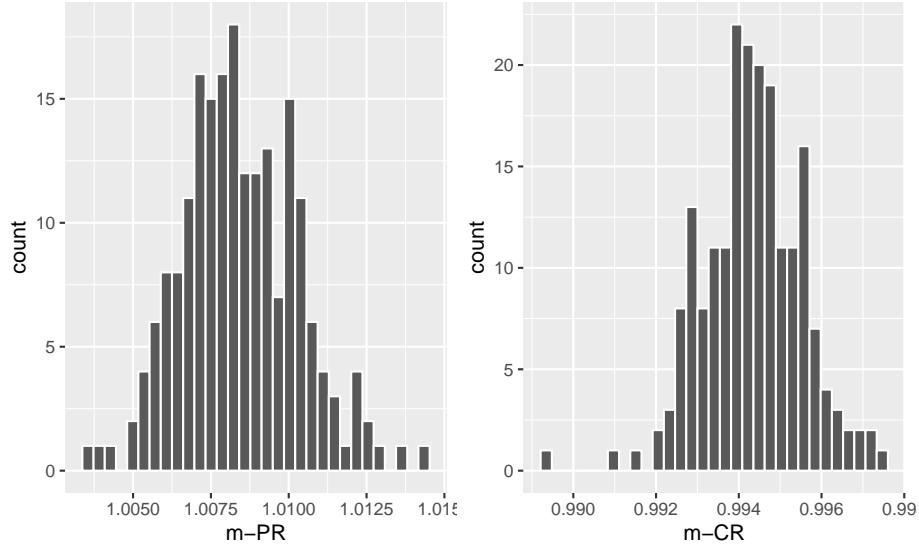


Figure 37.3: Histograms of slope PROPROC AUC vs. RSM AUC (left) and slope CBM AUC vs. RSM AUC (right).

### 37.8.3 Summary of slopes and confidence intervals

Table 37.1: Summary of slopes and correlations for the two constrained fits: PROPROC AUC vs. RSM AUC and CBM AUC vs. RSM AUC. The average of each slope equals unity to within 0.6 percent.

	$m_{PR}$	$R^2_{PR}$	$m_{CR}$	$R^2_{CR}$
TONY	1.0002	0.9997	0.9933	0.9997
VD	1.0061	0.9998	1.0007	1
FR	0.9995	1	0.9977	1
FED	1.0146	0.9998	0.9999	0.9999
JT	0.9964	0.9995	0.9972	1
MAG	1.036	0.9983	0.9953	1
OPT	1.0184	0.9997	1.0059	0.9997
PEN	1.0081	0.9996	0.9976	1
NICO	0.9843	0.9998	0.997	1
RUS	0.9989	0.9999	0.9921	0.9999
DOB1	1.0262	0.9963	0.9886	0.9962
DOB2	1.0056	0.9987	0.971	0.9978
DOB3	1.0211	0.998	0.9847	0.9986
FZR	1.0027	0.9999	0.9996	1
AVG	1.0084	0.9992	0.9943	0.9994
CI	(1.005, 1.012)	NA	(0.992, 0.997)	NA

Table 37.1: The first column, labeled  $m_{PR}$ , shows results of fitting straight lines, constrained to go through the origin, to fitted PROPROC AUC vs. RSM AUC results, for each of the 14 datasets, as labeled. The second column, labeled  $R^2_{PR}$ , lists the square of the correlation coefficient for each fit. The third and fourth columns list the corresponding values for the CBM AUC vs. RSM AUC fits. The second last row lists the averages (AVG) and the last row lists the 95 percent confidence intervals (CI) for the average slopes.

## 37.9 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 TBA 36. This was followed by Swensson's approach<sup>37</sup>, subsequently shown to be equivalent to my earlier work, as far as predicting the FROC curve was concerned TBA 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<sup>12</sup>.

This chapter describes an approach to fitting ROC curves, instead of FROC curves, using the RSM. Fits were described for 14 datasets, comprising 236 distinct treatment-reader combinations. All fits and parameter values are viewable in the online “Supplemental Material” directory corresponding to this chapter. Validity of fit was assessed by the chisquare goodness of fit p-value; unfortunately using adjacent bin combining this could not be calculated in most instances; ongoing research at other ways of validating the fits is underway. PROPROC and CBM were fitted to the same datasets, yielding further validation and insights. One of the insights was the finding that the AUCS were almost identical, with PROPROC yielding the highest value, followed by CBM and closely by the RSM. The PROPROC-AUC / CBM-AUC, vs. RSM-AUC straight-line fits, constrained to go through the origin, had slopes 1.0255 (1.021, 1.030) and 1.0097 (1.006, 1.013), respectively. The  $R^2$  values were generally in excess of 0.999, indicative of excellent fits.

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<sup>38</sup>. 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. A similar argument applies to LLs and NLs on diseased cases. The advantage of fitting to the ROC is that the coupling of NLs and LLs on diseased cases breaks the degeneracy problem described in §18.2.3.

The reader may wonder why I chose not to fit the wAFROC TBA. After all, it is the recommended figure of merit for FROC studies. While the methods described in this chapter are readily adapted to the wAFROC, they are more susceptible to degeneracy issues. The reason is that the y-axis is defined by LL-events, in other words by the parameters, while the x-axis is defined by the highest rated NL on non-diseased cases, in other words by the parameter. The

consequent decoupling of parameters leads to degeneracy of the type described in §18.2.3. This is avoided in ROC fitting because the y-axis is defined by LLs and NLs, in other words all parameters of the RSM are involved. The situation with the wAFROC is not quite as severe as with fitting FROC curves but it does have a problem with degeneracy. There are some ideas on how to improve the fits, perhaps by simultaneously fitting ROC and wAFROC-operating points, which amounts to putting constraints on the parameters, but for now this remains an open research subject. Empirical wAFROC, which is the current method implemented in RJafroc, is expected to have the same issues with variability of thresholds between treatments as the empirical ROC-AUC, as discussed in §5.9. So the fitting problem has to be solved. There is no need to fit the FROC, as it should never be used as a basis of a figure of merit for human observer studies; this is distinct from the severe degeneracy issues encountered with fitting it for human observers.

The application to a large number (236) of real datasets revealed that PROPROC has serious issues. These were apparently not revealed by the millions of simulations used to validate it<sup>39</sup>. To quote the cited reference, “The new algorithm never failed to converge and produced good fits for all of the several million datasets on which it was tested”. This is a good illustration of why simulations studies are not a good alternative to the method described in §18.5.1.3. In my experience this is a common misconception in this field, and is discussed further in the following chapter. Fig. 18.5, panels (J), (K) and (L) show that PROPROC, and to a lesser extent CBM, can, under some circumstances, severely overestimate performance. Recommendations regarding usage of PROPROC and CBM are deferred to Chapter 20.

The current ROC-based effort led to some interesting findings. The near equality of the AUCs predicted by the three proper ROC fitting methods, summarized in Table 18.4, has been noted, which is explained by the fact that proper ROC fitting methods represent different approaches to realizing an ideal observer, and the ideal observer must be unique, §18.6.

This chapter explores what is termed inter-correlations, between RSM and CBM parameters. Since they have similar physical meanings, the RSM and CBM separation parameters were found to be significantly correlated, = 0.86 (0.76, 0.89), as were the RSM and CBM parameters corresponding to the fraction of lesions that was actually visible, = 0.77 (0.68, 0.82). This type of correspondence between two different models can be interpreted as evidence of mutually reinforcing validity of each of the models.

The CBM method comes closest to the RSM in terms of yielding meaningful measures, but the fact that it allows the ROC curve to go continuously to (1,1) implies that it is not completely accounting for search performance, §17.8. There are two components to search performance: finding lesions and avoiding non-lesions. The CBM model accounts for finding lesions, but it does not account for avoiding suspicious regions that are non-diseased, an important characteristic of expert radiologists.

An important finding is the inverse correlation between search performance and lesion-classification performance, which suggest there could be tradeoffs in attempts to optimize them. As a simplistic illustration, a low-resolution gestalt-view of the image1, such as seen by the peripheral viewing mechanism, is expected to make it easier to rapidly spot deviations from the expected normal template described in Chapter 15. However, the observer may not be able to switch effectively between this and the high-resolution viewing mode necessary to correctly classify found suspicious region.

The main scientific conclusion of this chapter is that search-performance is the primary bottleneck in limiting observer performance. It is unfortunate that search is ignored in the ROC paradigm, usage of which is decreasing, albeit at an agonizingly slow rate. Evidence presented in this chapter should convince researchers to reconsider the focus of their investigations, most of which is currently directed at improving classification performance, which has been shown not to be the bottleneck. Another conclusion is that the three method of fitting ROC data yield almost identical AUCs. Relative to the RSM the PROPROC estimates are about 2.6% larger while CBM estimates are about 1% larger. This was a serendipitous finding that makes sense, in retrospect, but to the best of my knowledge is not known in the research community. PROPROC and to a lesser extent CBM are prone to severely overestimating performance in situations where the operating points are limited to a steep ascending section at the low end of false positive fraction scale. This parallels an earlier comment regarding the FROC, namely measurements derived from the steep part of the curve are unreliable, §17.10.1.

## 37.10 Appendices

### 37.11 Datasets

The datasets are embedded in the `RJafroc` package. They can be viewed in the help file of the package, a partial screen-shot of which is shown next <sup>4</sup>.

The datasets are identified in the code by `datasetdd` (where `dd` is an integer in the range 01 to 14) as follows:

- `dataset01` “TONY” FROC dataset (Chakraborty and Svahn, 2011)

```
## List of 3
## $ NL    : num [1:2, 1:5, 1:185, 1:3] 3 -Inf 3 -Inf 4 ...
```

---

<sup>4</sup>The raw datasets (Excel files) are in folder `R/compare-3-fits/Datasets` and file `R/compare-3-fits/loadDataFile.R` shows the correspondence between `datasetNames` and a dataset: for example, the Van Dyke dataset corresponds to file `VanDykeData.xlsx` in the `R/compare-3-fits/Datasets` folder.

<u>dataset01</u>	TONY FROC dataset
<u>dataset02</u>	Van Dyke ROC dataset
<u>dataset03</u>	Franken ROC dataset
<u>dataset04</u>	Federica Zanca FROC dataset
<u>dataset05</u>	John Thompson FROC dataset
<u>dataset06</u>	Magnus FROC dataset
<u>dataset07</u>	Lucy Warren FROC dataset
<u>dataset08</u>	Monica Penedo ROC dataset
<u>dataset09</u>	Nico Karssemeijer ROC dataset (CAD vs. radiologists)
<u>dataset10</u>	Marc Ruschin ROC dataset
<u>dataset11</u>	Dobbins 1 FROC dataset
<u>dataset12</u>	Dobbins 2 ROC dataset
<u>dataset13</u>	Dobbins 3 FROC dataset
<u>dataset14</u>	Federica Zanca real (as opposed to inferred) ROC dataset

Figure 37.4: Partial screen shot of ‘RJafroc’ help file showing the datasets included with the current distribution (v2.0.1).

```
## $ LL    : num [1:2, 1:5, 1:89, 1:2] 4 4 3 -Inf 3.5 ...
## $ LL_IL: logi NA

• dataset02 “VAN-DYKE” Van Dyke ROC dataset (Van Dyke et al., 1993)

## List of 3
## $ NL    : num [1:2, 1:5, 1:114, 1] 1 3 2 3 2 2 1 2 3 2 ...
## $ LL    : num [1:2, 1:5, 1:45, 1] 5 5 5 5 5 5 5 5 5 ...
## $ LL_IL: logi NA

• dataset03 “FRANKEN” Franken ROC dataset (Franken et al., 1992)

## List of 3
## $ NL    : num [1:2, 1:4, 1:100, 1] 3 3 4 3 3 3 4 1 1 3 ...
## $ LL    : num [1:2, 1:4, 1:67, 1] 5 5 4 4 5 4 4 5 2 2 ...
## $ LL_IL: logi NA

• dataset04 “FEDERICA” Federica Zanca FROC dataset (Zanca et al., 2009)

## List of 3
## $ NL    : num [1:5, 1:4, 1:200, 1:7] -Inf -Inf 1 -Inf -Inf ...
## $ LL    : num [1:5, 1:4, 1:100, 1:3] 4 5 4 5 4 3 5 4 4 3 ...
## $ LL_IL: logi NA

• dataset05 “THOMPSON” John Thompson FROC dataset (Thompson et al., 2014)
```

```

## List of 3
## $ NL    : num [1:2, 1:9, 1:92, 1:7] 4 5 -Inf -Inf 8 ...
## $ LL    : num [1:2, 1:9, 1:47, 1:3] 5 9 -Inf 10 8 ...
## $ LL_IL: logi NA

• dataset06 “MAGNUS” Magnus Bath FROC dataset (Vikgren et al., 2008)

## List of 3
## $ NL    : num [1:2, 1:4, 1:89, 1:17] 1 -Inf -Inf -Inf 1 ...
## $ LL    : num [1:2, 1:4, 1:42, 1:15] -Inf -Inf -Inf -Inf ...
## $ LL_IL: logi NA

• dataset07 “LUCY-WARREN” Lucy Warren FROC dataset (Warren et al., 2014)

## List of 3
## $ NL    : num [1:5, 1:7, 1:162, 1:4] 1 2 1 2 -Inf ...
## $ LL    : num [1:5, 1:7, 1:81, 1:3] 2 -Inf 2 -Inf 1 ...
## $ LL_IL: logi NA

• dataset08 “PENEDO” Monica Penedo FROC dataset (Penedo et al., 2005)

## List of 3
## $ NL    : num [1:5, 1:5, 1:112, 1] 3 2 3 2 3 0 0 4 0 2 ...
## $ LL    : num [1:5, 1:5, 1:64, 1] 3 2 4 3 3 3 3 4 4 3 ...
## $ LL_IL: logi NA

• dataset09 “NICO-CAD-ROC” Nico Karssemeijer ROC dataset (Hupse et al., 2013)

## List of 3
## $ NL    : num [1, 1:10, 1:200, 1] 28 0 14 0 16 0 31 0 0 0 ...
## $ LL    : num [1, 1:10, 1:80, 1] 29 12 13 10 41 67 61 51 67 0 ...
## $ LL_IL: logi NA

• dataset10 “RUSCHIN” Mark Ruschin ROC dataset (Ruschin et al., 2007)

## List of 3
## $ NL    : num [1:3, 1:8, 1:90, 1] 1 0 0 0 0 0 1 0 0 0 ...
## $ LL    : num [1:3, 1:8, 1:40, 1] 2 1 1 2 0 0 0 0 0 3 ...
## $ LL_IL: logi NA

```

- `dataset11` “DOBBINS-1” Dobbins I FROC dataset (Dobbins III et al., 2016)
- ```
## 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
```
- `dataset12` “DOBBINS-2” Dobbins II ROC dataset (Dobbins III et al., 2016)
- ```
## List of 3
## $ NL    : num [1:4, 1:5, 1:152, 1] -Inf -Inf -Inf -Inf -Inf ...
## $ LL    : num [1:4, 1:5, 1:88, 1] 3 4 4 -Inf -Inf ...
## $ LL_IL: logi NA
```
- `dataset13` “DOBBINS-3” Dobbins III FROC dataset (Dobbins III et al., 2016)
- ```
## List of 3
## $ NL    : num [1:4, 1:5, 1:158, 1:4] -Inf 3 -Inf 4 5 ...
## $ LL    : num [1:4, 1:5, 1:106, 1:15] -Inf -Inf -Inf -Inf -Inf ...
## $ LL_IL: logi NA
```
- `dataset14` “FEDERICA-REAL-ROC” Federica Zanca *real* ROC dataset (Zanca et al., 2012)
- ```
## List of 3
## $ NL    : num [1:2, 1:4, 1:200, 1] 2 2 2 2 1 3 2 2 3 1 ...
## $ LL    : num [1:2, 1:4, 1:100, 1] 6 5 6 4 5 5 5 5 5 4 ...
## $ LL_IL: logi NA
```

## 37.12 Location of PROPROC files

For each dataset PROPROC parameters were obtained by running the Windows software with PROPROC selected as the curve-fitting method. The results are saved to files that end with `propocnormareapooled.csv`<sup>5</sup> contained in “R/compare-3-fits/MRMCRuns/C/”, where C denotes the name of the dataset (for example, for the Van Dyke dataset, C = “VD”). Examples are shown in the next two screen-shots.

---

<sup>5</sup>In accordance with R-package policies white-spaces in the original PROPROC output file names have been removed.

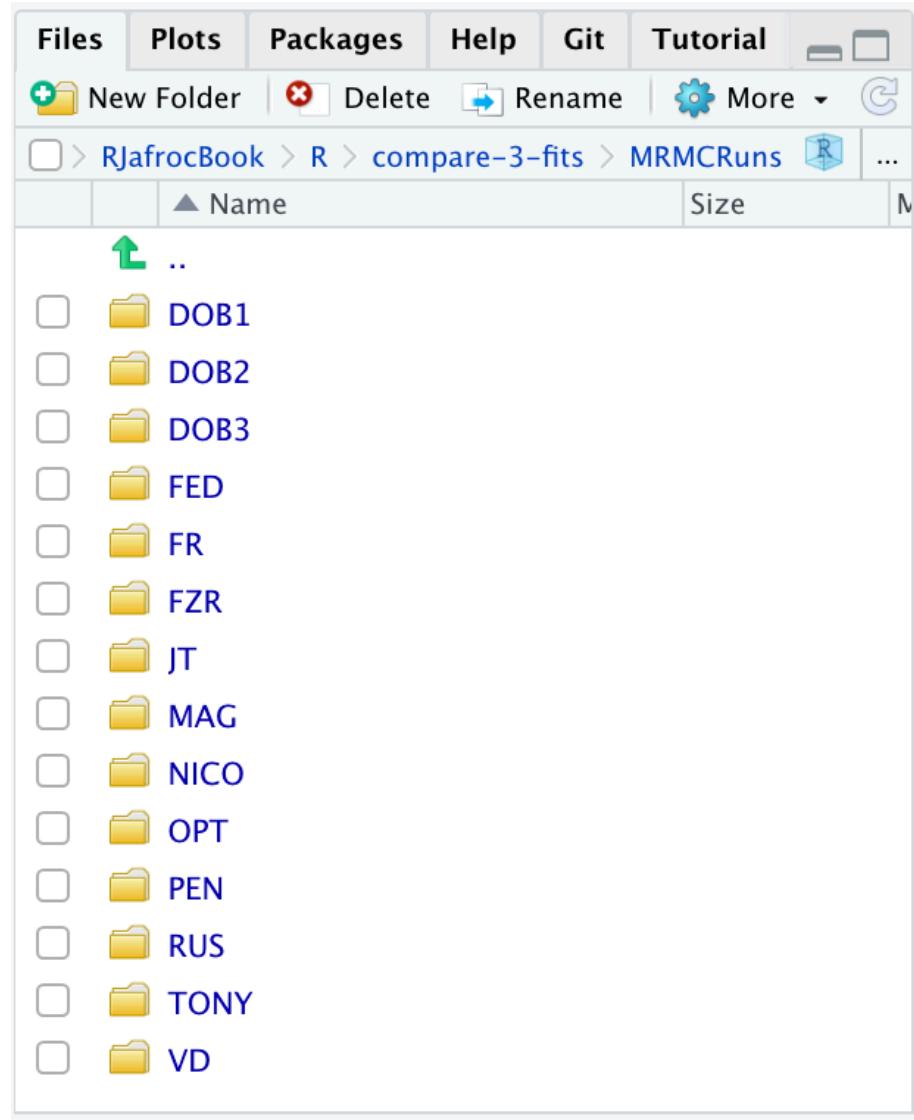


Figure 37.5: Screen shot (1 of 2) of ‘R/compare-3-fits/MRMCRuns‘ showing the folders containing the results of PROPROC analysis on 14 datasets.

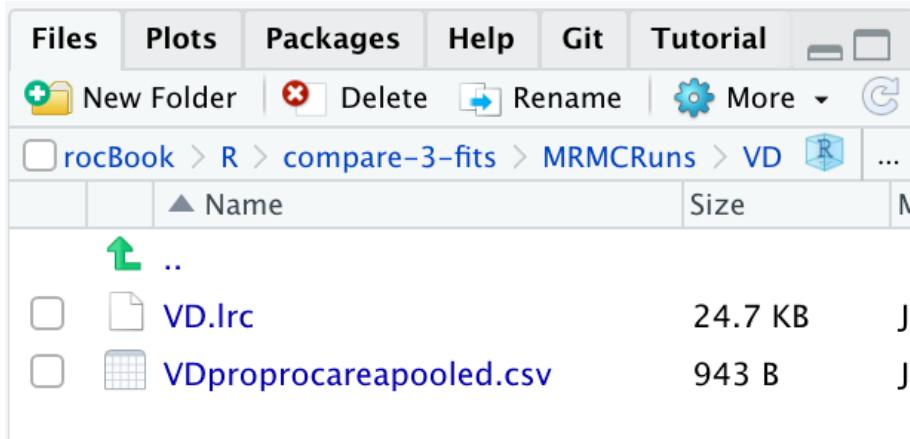


Figure 37.6: Screen shot (2 of 2) of ‘R/compare-3-fits/MRMCRuns/VD’ showing files containing the results of PROPROC analysis for the Van Dyke dataset.

The contents of R/compare-3-fits/MRMCRuns/VD/VDproprocnormareapooled.csv are shown next, see Fig. 37.7.<sup>6</sup> The PROPROC parameters  $c$  and  $d_a$  are in the last two columns. The column names are T = treatment; R = reader; return-code = undocumented value, area = PROPROC AUC; numCAT = number of ROC bins; adjPMean = undocumented value; c =  $c$  and d\_a =  $d_a$ , are the PROPROC parameters defined in (Metz and Pan, 1999).

R/afrocBook - mas						
<i>myRfci.lib</i> 19b-rm-3-fits.Rmd VDproprocareapooled.csv 82-froc-data-format.Rmd CompareH						
1	T,R,returnCode,area,numCAT,adjPMean,c,d_a,					
2	1, 1, 0, 0.9340403616, 5, 0.9340403616, -0.2980072344, 2.1255412315					
3	1, 2, 0, 0.8910714123, 4, 0.8910714123, -0.2890004255, 1.7314724686					
4	1, 3, 0, 0.8910714123, 4, 0.8910714123, -0.2890004255, 1.7314724686					
5	1, 4, 0, 0.9774594813, 4, 0.9774594813, -0.9315807469, 0.3005236787					
6	1, 5, 0, 0.8489557684, 5, 0.8489557684, -0.5074209320, 0.8958632862					
7	2, 1, 0, 0.9519359395, 5, 0.9519359395, -0.3212354325, 2.3481494976					
8	2, 2, 0, 0.9519359395, 5, 0.9519359395, -0.3212354325, 2.3481494976					
9	2, 3, 0, 0.9304317582, 5, 0.9304317582, -0.329982976, 2.0785431399					
10	2, 4, 3, 1.0000000000, 3, 1.0000000000, 1.0000000000, 0.2000000000					
11	2, 5, 0, 0.9426874140, 4, 0.9426874140, -0.5530831909, 2.0196598664					
12						

Figure 37.7: PROPROC output for the Van Dyke ROC data set.

### 37.13 Location of pre-analyzed results

The following screen shot shows the pre-analyzed files created by the function `Compare3ProperRocFits()` described below. Each file is named `allResultsC`, where C is the abbreviated name of the dataset (uppercase C denotes one or more uppercase characters; for example, C = VD denotes the Van Dyke dataset.).

<sup>6</sup>The VD.lrc file in this directory is the Van Dyke data formatted for input to OR DBM-MRMC 2.5.

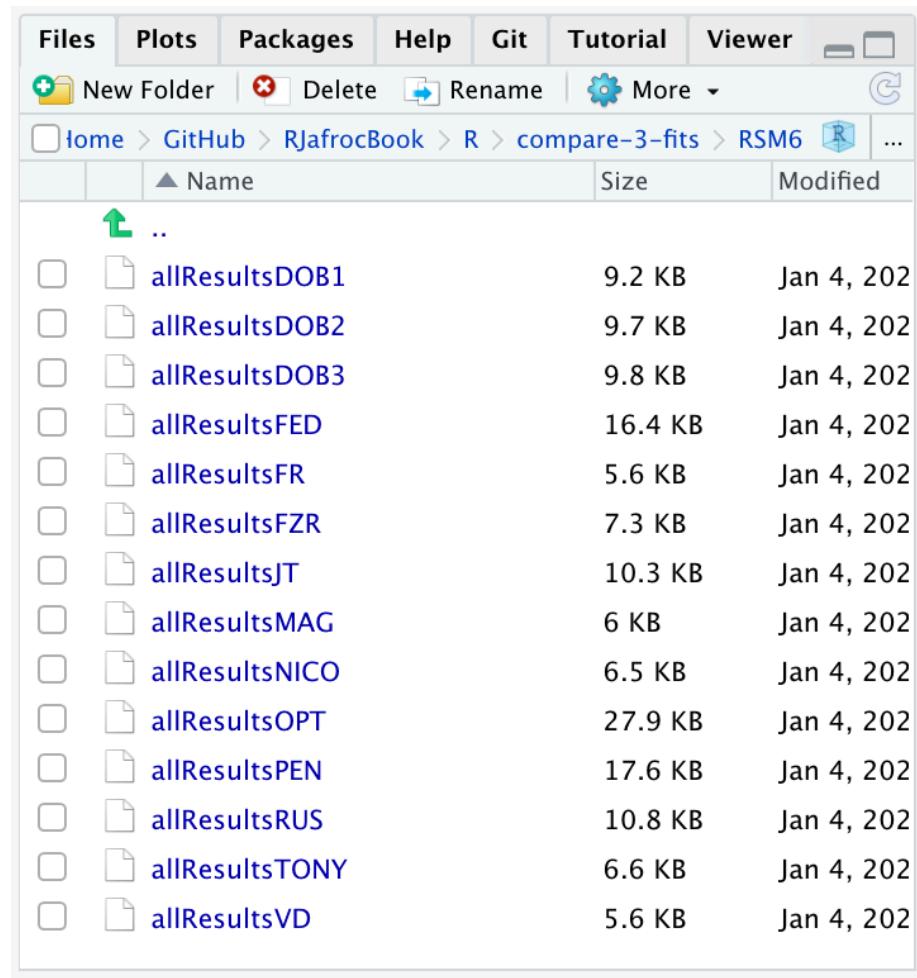


Figure 37.8: Screen shot of ‘R/compare-3-fits/RSM6‘ showing the results files created by ‘Compare3ProperRocFits()‘ .

### 37.14 Plots for Van Dyke dataset

The following plots are arranged in pairs, with the left plot corresponding to treatment 1 and the right to treatment 2.

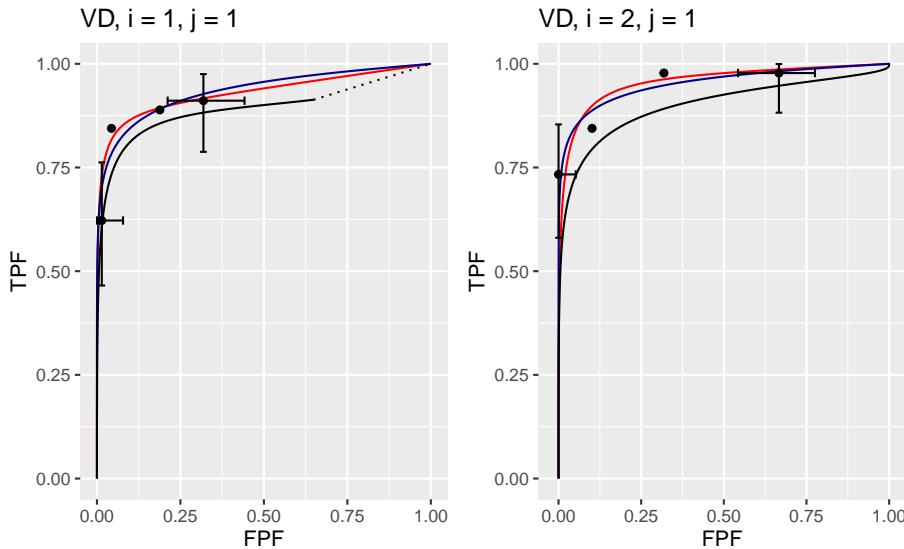


Figure 37.9: Composite plots in both treatments for Van Dyke dataset, reader 1.

The RSM parameter values for the treatment 2 plot are:  $\mu = 5.767237$ ,  $\lambda' = 2.7212621$ ,  $\nu' = 0.8021718$ ,  $\zeta_1 = -1.5717303$ . The corresponding CBM values are  $\mu = 5.4464738$ ,  $\alpha = 0.8023609$ ,  $\zeta_1 = -1.4253826$ . The RSM and CBM  $\mu$  parameters are very close and likewise the RSM  $\nu'$  and CBM  $\alpha$  parameters are very close - this is because they have similar physical meanings, which is investigated later in this chapter TBA. [The CBM does not have a parameter analogous to the RSM  $\lambda'$  parameter.]

The RSM parameters for the treatment 1 plot are:  $\mu = 3.1527627$ ,  $\lambda' = 9.9986154$ ,  $\nu' = 0.9899933$ ,  $\zeta_1 = 1.1733988$ . The corresponding CBM values are  $\mu = 2.1927712$ ,  $\alpha = 0.98$ ,  $\zeta_1 = -0.5168848$ .

### 37.15 References

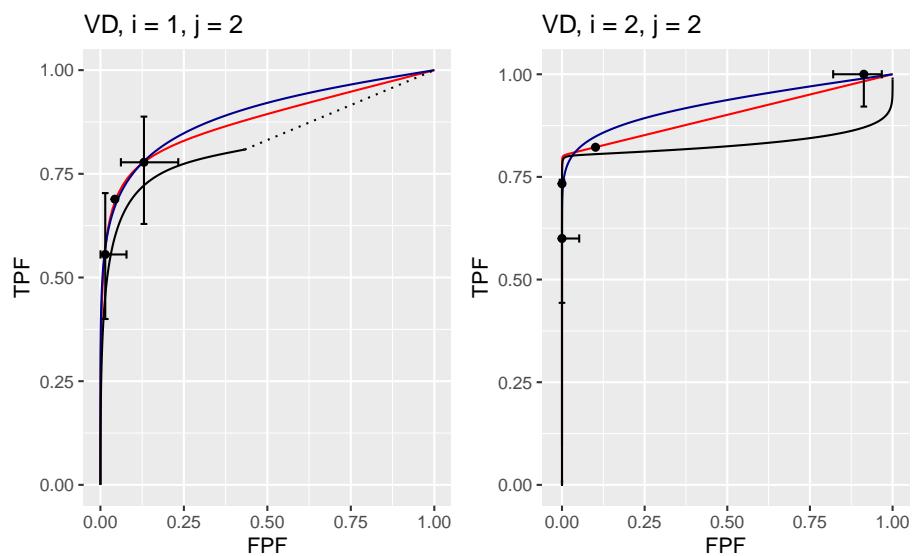


Figure 37.10: Composite plots in both treatments for Van Dyke dataset, reader 2. For treatment 2 the RSM and PROPROC fits are indistinguishable.

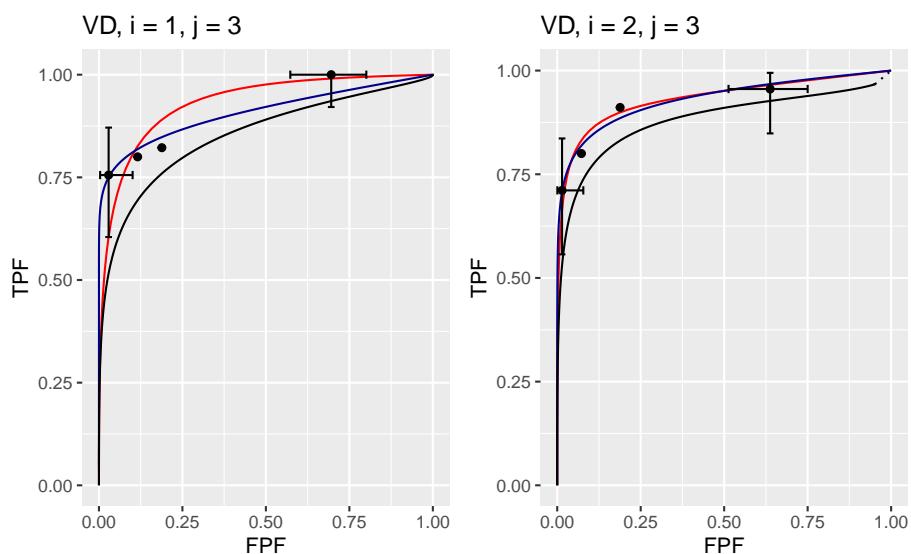


Figure 37.11: Composite plots in both treatments for Van Dyke dataset, reader 3.

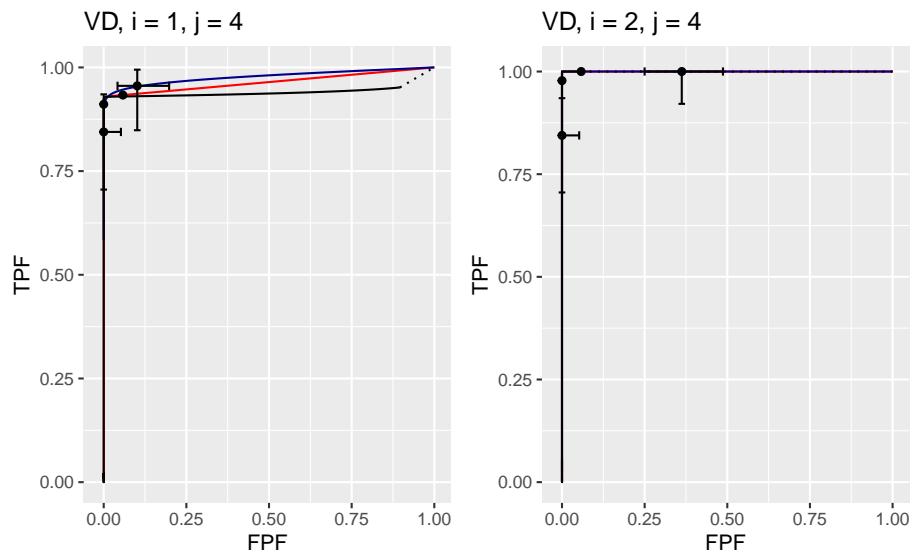


Figure 37.12: Composite plots in both treatments for Van Dyke dataset, reader 4. For treatment 2 the 3 plots are indistinguishable and each one has  $AUC = 1$ . The degeneracy is due to all operating points being on the axes of the unit square.

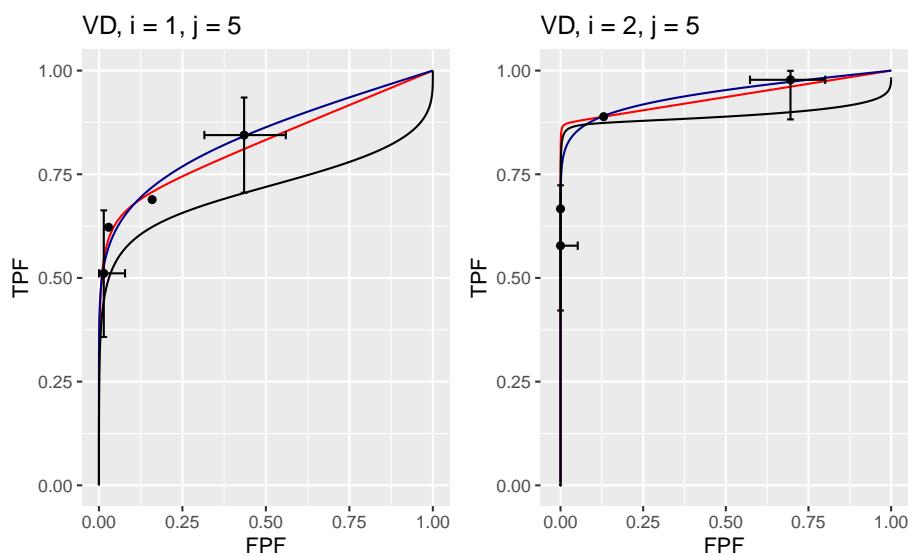


Figure 37.13: Composite plots in both treatments for Van Dyke dataset, reader 5.



**CAD**



# **Chapter 38**

## **Standalone CAD vs. Radiologists**

### **38.1 TBA How much finished**

10%

### **38.2 Abstract**

Computer aided detection (CAD) research for screening mammography has so far focused on measuring performance of radiologists with and without CAD. Typically a group of radiologists interpret a set of images with and without CAD assist. Standalone performance of CAD algorithms is rarely measured. The stated reason for this is that in the clinic CAD is never used alone, rather it is always used with radiologists. For this reason interest has been focused on the incremental improvement afforded by CAD.

Another reason for the lack of focus on standalone CAD performance is the lack of clear methodology for measuring standalone CAD performance. This chapter extends the methodology used in a recent study of standalone performance. The method is termed random-reader fixed case (1T-RRFC), since it only accounts for reader variability but does not account for case-variability. The extension includes the effect of case-sampling variability. Since in the proposed method CAD is treated as an additional reader within a single treatment, the method is termed one-treatment random-reader random-case (1T-RRRC) analysis. The new method is based on existing methodology allowing comparison of the average performance of readers in a single treatment to a specified value. The key modification is to regard the difference in performance between radiologists and

CAD as a figure of merit, to which the existing work is then directly applicable. The 1T-RRRC method was compared to 1T-RRFC. It was also compared to an unorthodox usage of conventional ROC (receiver operating characteristic) analysis software, termed 2T-RRRC analysis, which involves replicating the CAD ratings as many times as there are radiologists, to in effect simulate a second treatment, i.e., CAD is regarded as the second treatment. The proposed 1T-RRRC analysis has 3 random parameters as compared to 6 parameters in 2T-RRRC and one parameter in 1T-RRFC. As expected, since one is including an additional source of variability, both RRRC analyses (1T and 2T) yielded larger p-values and wider confidence intervals as compared to 1T-RRFC. For the F-statistic, degrees of freedom and p-value, both 1T-RRRC and 2T-RRRC analyses yielded exactly the same results. However, 2T-RRRC model parameter estimates were unrealistic; for example, it yields zero between-reader variance, whereas 1T-RRRC yielded the expected non-zero value. All three methods are implemented in an open-source R package `RJafroc`.

### 38.3 Keywords

Technology assessment, computer-aided detection (CAD), screening mammography, standalone performance, single-treatment multi-reader ROC analysis.

### 38.4 Introduction

In the US the majority of screening mammograms are analyzed by computer aided detection (CAD) algorithms (Rao et al., 2010). Almost all major imaging device manufacturers provide CAD as part of their imaging workstation display software. In the United States CAD is approved for use as a second reader (Petrick and Pastel, 2018), i.e., the radiologist first interprets the images (typically 4 views, 2 views of each breast) without CAD and then CAD information (i.e., cued suspicious regions, possibly shown with associated probabilities of malignancies) is shown and the radiologist has the opportunity to revise the initial interpretation. In response to the second reader usage, the evolution of CAD algorithms has been guided mainly by comparing observer performance of radiologists with and without CAD.

Clinical CAD systems sometimes only report the locations of suspicious regions, i.e., it may not provide ratings. However, a (continuous variable) malignancy index for every CAD-found suspicious region is available to the algorithm designer (Edwards et al., 2002). Standalone performance, i.e., performance of designer-level CAD by itself, regarded as an algorithmic reader, vs. radiologists, is rarely measured. In breast cancer screening I am aware of only one study (Hupse et al., 2013) where standalone performance was measured. [Standalone

performance has been measured in CAD for computed tomography colonography, chest radiography and three dimensional ultrasound (Hein et al., 2010; Summers et al., 2008; Taylor et al., 2006; De Boo et al., 2011; Tan et al., 2012)].

One possible reason for not measuring standalone performance of CAD is the lack of an accepted assessment methodology for such measurements. The purpose of this work is to remove that impediment. It describes a method for comparing standalone performance of designer-level CAD to radiologists interpreting the same cases and compares the method to those described in two recent publications (Hupse et al., 2013; Kooi et al., 2016).

## 38.5 Methods

Summarized are two recent studies of CAD vs. radiologists in mammography. This is followed by comments on the methodologies used in the two studies. The second study used multi-treatment multi-reader receiver operating characteristic (ROC) software in an unorthodox or unconventional way. A statistical model and analysis method is described that avoids unorthodox, and perhaps unjustified, use of ROC software and has fewer model parameters.

### 38.5.1 Studies assessing performance of CAD vs. radiologists

The first study (Hupse et al., 2013) measured performance in finding and localizing lesions in mammograms, i.e., visual search was involved, while the second study (Kooi et al., 2016) measured lesion classification performance between non-diseased and diseased regions of interest (ROIs) previously found on mammograms by an independent algorithmic reader, i.e., visual search was not involved.

#### 38.5.1.1 Study - 1

The first study (Hupse et al., 2013) compared standalone performance of a CAD device to that of 9 radiologists interpreting the same cases (120 non-diseased and 80 with a single malignant mass per case). It used the LROC (localization ROC) paradigm (Starr et al., 1975; Metz et al., 1976; Swensson, 1996b), in which the observer gives an overall rating for presence of disease (an integer 0 to 100 scale was used) and indicates the location of the most suspicious region. On a non-diseased case the rating is classified as a false positive (FP) but on a diseased case it is classified as a *correct localization* (CL) if the location is sufficiently close to the lesion, and otherwise it is classified as an *incorrect localization*. For a given reporting threshold, the number of correct localizations divided by the number of diseased cases estimates the probability of correct localization (PCL)

at that threshold. On non-diseased cases the number of false positives (FPs) divided by the number of non-diseased cases estimates the probability of a false positive, or false positive fraction (FPF), at that threshold. The plot of PCL (ordinate) vs. FPF defines the LROC curve. Study - 1 used as figures of merit (FOMs) the interpolated PCL at two values of FPF, specifically FPF = 0.05 and FPF = 0.2, denoted PCL<sub>0.05</sub> and PCL<sub>0.2</sub>, respectively. The t-test between the radiologist PCL<sub>FPF</sub> values and that of CAD was used to compute the two-sided p-value for rejecting the NH of equal performance. Study - 1 reported p-value = 0.17 for PCL<sub>0.05</sub> and p-value  $\leq 0.001$ , with CAD being inferior, for PCL<sub>0.2</sub>.

### 38.5.1.2 Study - 2

The second study (Kooi et al., 2016) used 199 diseased and 199 non-diseased ROIs extracted by an independent CAD algorithm. These were interpreted using the ROC paradigm (i.e., rating only, no localization required) by a different CAD algorithmic observer from that used to determine the ROIs, and by four expert radiologists. The figure of merit was the empirical area (AUC) under the respective ROC curves (one per radiologist and one for CAD). The p-value for the difference in AUCs between the average radiologist and CAD was determined using an unorthodox application of the Dorfman-Berbaum-Metz (Dorfman et al., 1992b) multiple-treatment multiple-reader multiple-case (DBM-MRMC) software with recent modifications (Hillis et al., 2008b). The unorthodox application was that in the input data file *radiologists and CAD were entered as two treatments*. In conventional (or orthodox) DBM-MRMC each reader provides two ratings per case and the data file would consist of paired ratings of a set of cases interpreted by 4 readers. To accommodate the paired data structure assumed by the software, the authors of Study - 2 *replicated the CAD ratings four times in the input data file*, as explained in the caption to Table 38.1. By this artifice they converted a single-treatment 5-reader (4 radiologists plus CAD) data file to a two-treatment 4-reader data file, in which the four readers in treatment 1 were the radiologists, and the four “readers” in treatment 2 were CAD replicated ratings. Note that for each case the four readers in the second treatment had identical ratings. In Table 1 the replicated CAD observers are labeled C1, C2, C3 and C4.

Study - 2 reported a not significant difference between CAD and the radiologists ( $p = 0.253$ ).

### 38.5.1.3 Comments

For the purpose of this work, which focuses on the respective analysis methods, the difference in observer performance paradigms between the two studies, namely a search paradigm in Study - 1 vs. an ROI classification paradigm in Study - 2, is inconsequential. The paired t-test used in Study - 1 treats the case-sample as fixed. In other words, the analysis is not accounting for case-sampling

Table 38.1: The differences between the data structures in conventional DBM-MRMC analysis and the unorthodox application of the software used in Study - 2. There are four radiologists, labeled R1, R2, R3 and R4 interpreting 398 cases labeled 1, 2, ..., 398, in two treatments, labeled 1 and 2. Sample ratings are shown only for the first and last radiologist and the first and last case. In the first four columns, labeled "Standard DBM-MRMC", each radiologist interprets each case twice. In the next four columns, labeled "Unorthodox DBM-MRMC", the radiologists interpret each case once. CAD ratings are replicated four times to effectively create the second "treatment". The quotations emphasize that there is, in fact, only one treatment. The replicated CAD observers are labeled C1, C2, C3 and C4.

Standard DBM-MRMC				Unorthodox DBM-MRMC			
Reader	Treatment	Case	Rating	Reader	Treatment	Case	Rating
R1	1	1	75	R1	1	1	75
...	...	...	...	...	...	...	...
R1	1	398	0	R1	1	398	0
...	...	...	...	...	...	...	...
R4	1	1	50	R4	1	1	50
...	...	...	...	...	...	...	...
R4	1	398	25	R4	1	398	25
R1	2	1	45	C1	2	1	55
...	...	...	...	...	...	...	...
R1	2	398	25	C1	2	398	5
...	...	...	...	...	...	...	...
R4	2	1	95	C4	2	1	55
...	...	...	...	...	...	...	...
R4	2	398	20	C4	2	398	5

variability but it is accounting for reader variability. While not explicitly stated, the reason for the unorthodox analysis in Study – 2 was the desire to include case-sampling variability.<sup>1</sup>

In what follows, the analysis in Study – 1 is referred to as random-reader fixed-case (1T-RRFC) while that in Study – 2 is referred to as dual-treatment random-reader random-case (2T-RRRC).

### 38.5.2 The 1T-RRFC analysis model

The sampling model for the FOM is:

$$\left. \begin{aligned} \theta_j &= \mu + R_j \\ (j &= 1, 2, \dots, J) \end{aligned} \right\} \quad (38.1)$$

Here  $\mu$  is a constant,  $\theta_j$  is the FOM for reader  $j$ , and  $R_j$  is the random contribution for reader  $j$  distributed as:

$$R_j \sim N(0, \sigma_R^2) \quad (38.2)$$

Because of the assumed normal distribution of  $R_j$ , in order to compare the readers to a fixed value, that of CAD denoted  $\theta_0$ , one uses the (unpaired) t-test, as done in Study – 1. As evident from the model, no allowance is made for case-sampling variability, which is the reason for calling it the 1T-RRFC method.

Performance of CAD on a fixed dataset does exhibit within-reader variability. The same algorithm applied repeatedly to a fixed dataset does not always produce the same mark-rating data. However, this source of CAD FOM variability is much smaller than inter-reader FOM variability of radiologists interpreting the same dataset. In fact the within-reader variability of radiologists is smaller than their inter-reader variability, and within-reader variability of CAD is even smaller still. For this reason one is justified in regarding  $\theta_0$  as a fixed quantity for a given dataset. Varying the dataset will result in different values for  $\theta_0$ , i.e., its case sampling variability needs to be accounted for, as done in the following analyses.

### 38.5.3 The 2T-RRRC analysis model

This could be termed the conventional or the orthodox method. There are two treatments and the study design is fully crossed: each reader interprets each case in each treatment, i.e., the data structure is as in the left half of Table 1.

---

<sup>1</sup>Prof. Karssemeijer (private communication, 10/27/2017) had consulted with a few ROC experts to determine if the procedure used in Study – 2 was valid, and while the experts thought it was probably valid they were not sure.

The following approach, termed 2T-RRRC, uses the Obuchowski and Rockette (OR) figure of merit sampling model (Obuchowski and Rockette, 1995b) instead of the pseudo-value-based model used in the original DBM paper (Dorfman et al., 1992b). For the empirical FOM, Hillis has shown the two to be equivalent (Hillis et al., 2005b).

The OR model is:

$$\theta_{ij\{c\}} = \mu + \tau_i + (\tau R)_{ij} + \epsilon_{ij\{c\}} \quad (38.3)$$

Assuming two treatments,  $i$  ( $i = 1, 2$ ) is the treatment index,  $j$  ( $j = 1, \dots, J$ ) is the reader index, and  $k$  ( $k = 1, \dots, K$ ) is the case index, and  $\theta_{ij\{c\}}$  is a figure of merit for reader  $j$  in treatment  $i$  and case-sample  $\{c\}$ . A case-sample is a set or ensemble of cases, diseased and non-diseased, and different integer values of  $c$  correspond to different case-samples.

The first two terms on the right hand side of Eqn. (38.3) are fixed effects (average performance and treatment effect, respectively). The next two terms are random effect variables that, by assumption, are sampled as follows:

$$\begin{aligned} R_j &\sim N(0, \sigma_R^2) \\ (\tau R)_{ij} &\sim N(0, \sigma_{\tau R}^2) \end{aligned} \quad (38.4)$$

The terms  $R_j$  represents the random treatment-independent contribution of reader  $j$ , modeled as a sample from a zero-mean normal distribution with variance  $\sigma_R^2$ ,  $(\tau R)_{ij}$  represents the random treatment-dependent contribution of reader  $j$  in treatment  $i$ , modeled as a sample from a zero-mean normal distribution with variance  $\sigma_{\tau R}^2$ . The sampling of the last (error) term is described by:

$$\epsilon_{ij\{c\}} \sim N_{I \times J}(\vec{0}, \Sigma) \quad (38.5)$$

Here  $N_{I \times J}$  is the  $I \times J$  variate normal distribution and  $\vec{0}$ , a  $I \times J$  length zero-vector, represents the mean of the distribution. The  $\{I \times J\} \times \{I \times J\}$  dimensional covariance matrix  $\Sigma$  is defined by 4 parameters, Var, Cov<sub>1</sub>, Cov<sub>2</sub>, Cov<sub>3</sub>, defined as follows:

$$\text{Cov}(\epsilon_{ij\{c\}}, \epsilon_{i'j'\{c\}}) = \begin{cases} \text{Var}(i = i', j = j') \\ \text{Cov1}(i \neq i', j = j') \\ \text{Cov2}(i = i', j \neq j') \\ \text{Cov3}(i \neq i', j \neq j') \end{cases} \quad (38.6)$$

Software {U of Iowa and RJafroc} yields estimates of all terms appearing on the right hand side of Eqn. (38.6). Excluding fixed effects, the model represented by Eqn. (38.3) contains six parameters:

$$\sigma_R^2, \sigma_{\tau R}^2, \text{Var}, \text{Cov}_1, \text{Cov}_2, \text{Cov}_3 \quad (38.7)$$

The meanings the last four terms are described in (Hillis, 2007a; Obuchowski and Rockette, 1995b; Hillis et al., 2005b; Chakraborty, 2017). Briefly, Var is the variance of a reader's FOMs, in a given treatment, over interpretations of different case-samples, averaged over readers and treatments; Cov<sub>1</sub>/Var is the correlation of a reader's FOMs, over interpretations of different case-samples in different treatments, averaged over all different-treatment same-reader pairings; Cov<sub>2</sub>/Var is the correlation of different reader's FOMs, over interpretations of different case-samples in the same treatment, averaged over all same-treatment different-reader pairings and finally, Cov<sub>3</sub>/Var is the correlation of different reader's FOMs, over interpretations of different case-samples in different treatments, averaged over all different-treatment different-reader pairings. One expects the following inequalities to hold:

$$\text{Var} \geq \text{Cov}_1 \geq \text{Cov}_2 \geq \text{Cov}_3 \quad (38.8)$$

In practice, since one is usually limited to one case-sample, i.e.,  $c = 1$ , resampling techniques (Efron and Tibshirani, 1994) – e.g., the jackknife – are used to estimate these terms.

### 38.5.4 The 1T-RRRC analysis model

This is the contribution of this work. The key difference from the approach in Study - 2 is to regard standalone CAD as a different reader, not as a different treatment. Therefore, needed is a single treatment method for analyzing readers and CAD, where the latter is regarded as an additional reader. Accordingly the proposed method is termed single-treatment RRRC (1T-RRRC) analysis.

The starting point is the (Obuchowski and Rockette, 1995b) model for a single treatment, which for the radiologists (i.e., *excluding* CAD) interpreting in a single-treatment reduces to the following model:

$$\theta_{j\{c\}} = \mu + R_j + \epsilon_{j\{c\}} \quad (38.9)$$

$\theta_{j\{c\}}$  is the figure of merit for radiologist  $j$  ( $j = 1, 2, \dots, J$ ) interpreting case-sample  $\{c\}$ ;  $R_j$  is the random effect of radiologist  $j$  and  $\epsilon_{j\{c\}}$  is the error term. For single-treatment multiple-reader interpretations the error term is distributed as:

$$\epsilon_{j\{c\}} \sim N_J(\vec{0}, \Sigma) \quad (38.10)$$

The  $J \times J$  covariance matrix  $\Sigma$  is defined by two parameters, Var and Cov<sub>2</sub>, as follows:

$$\Sigma_{jj'} = \text{Cov}(\epsilon_{j\{c\}}, \epsilon_{j'\{c\}}) = \begin{cases} \text{Var} & j = j' \\ \text{Cov}_2 & j \neq j' \end{cases} \quad (38.11)$$

The terms  $\text{Var}$  and  $\text{Cov}_2$  are estimated using resampling methods. Using the jackknife, and denoting the FOM with case  $k$  removed by  $\psi_{j(k)}$  (the index in parenthesis denotes deleted case  $k$ , and since one is dealing with a single case-sample, the case-sample index  $c$  is now superfluous). The covariance matrix is estimated using (the dot symbol represents an average over the replaced index):

$$\Sigma_{jj'}|_{\text{jack}} = \frac{K-1}{K} \sum_{k=1}^K (\psi_{j(k)} - \bar{\psi}_{j(\bullet)}) (\psi_{j'(k)} - \bar{\psi}_{j'(\bullet)}) \quad (38.12)$$

The final estimates of  $\text{Var}$  and  $\text{Cov}_2$  are averaged (indicated in the following equation by the angular brackets) over all pairings of radiologists satisfying the relevant equalities/inequalities shown just below the closing angular bracket:

$$\begin{aligned} \text{Var} &= \langle \Sigma_{jj'}|_{\text{jack}} \rangle_{j=j'} \\ \text{Cov}_2 &= \langle \Sigma_{jj'}|_{\text{jack}} \rangle_{j \neq j'} \end{aligned} \quad (38.13)$$

Hillis' formulae (Hillis et al., 2005b; Hillis, 2007a) permit one to test the NH:  $\mu = \mu_0$ , where  $\mu_0$  is a pre-specified constant. One could set  $\mu_0$  equal to the performance of CAD, but that would not be accounting for the fact that the performance of CAD is itself a random variable, whose case-sampling variability needs to be accounted for.

Instead, the following model was used for the figure of merit of the radiologists and CAD ( $j = 0$  is used to denote the CAD algorithmic reader):

$$\theta_{j\{c\}} = \theta_{0\{c\}} + \Delta\theta + R_j + \epsilon_{j\{c\}} \quad j = 1, 2, \dots, J \quad (38.14)$$

$\theta_{0\{c\}}$  is the CAD figure of merit for case-sample  $\{c\}$  and  $\Delta\theta$  is the average figure of merit increment of the radiologists over CAD. To reduce this model to one to which existing formulae are directly applicable, one subtracts the CAD figure of merit from each radiologist's figure of merit (for the same case-sample), and defines this as the difference figure of merit  $\psi_{j\{c\}}$ , i.e.,

$$\psi_{j\{c\}} = \theta_{j\{c\}} - \theta_{0\{c\}} \quad (38.15)$$

Then Eqn. (38.14) reduces to:

$$\psi_{j\{c\}} = \Delta\theta + R_j + \epsilon_{j\{c\}} \quad j = 1, 2, \dots, J \quad (38.16)$$

Eqn. (38.16) is identical in form to Eqn. (38.9) with the difference that the figure of merit on the left hand side of Eqn. (38.16) is a *difference FOM*, that between the radiologist's and CAD. Eqn. (38.16) describes a model for  $J$  radiologists interpreting a common case set, each of whose performances is measured relative to that of CAD. Under the NH the expected difference is zero: NH: $\Delta\theta = 0$ . The method (Hillis et al., 2005b; Hillis, 2007a) for single-treatment multiple-reader analysis is now directly applicable to the model described by Eqn. (38.16).

Apart from fixed effects, the model in Eqn. (38.16) contains three parameters:

$$\sigma_R^2, \text{Var}, \text{Cov}_2 \quad (38.17)$$

Setting  $\text{Var} = 0, \text{Cov}_2 = 0$  yields the 1T-RRFC model, which contains only one random parameter, namely  $\sigma_R^2$ . [One expects identical estimates of  $\sigma_R^2$  using 1T-RRFC, 2T-RRRC or 1T-RRRC analyses.]

## 38.6 Software implementation

The three analyses, namely random-reader fixed-case (1T-RRFC), dual-treatment random-reader random-case (2T-RRRC) and single-treatment random-reader random-case (1T-RRRC), are implemented in **RJafroc**, an R-package (Chakraborty et al., 2020a).

The following code shows usage of the software to generate the results corresponding to the three analyses. Note that **datasetCadLroc** is the LROC dataset and **dataset09** is the corresponding ROC dataset.

```
RRFC_1T_PCL_0_05 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.05, method = "1T-RRFC")
RRRC_2T_PCL_0_05 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.05, method = "2T-RRRC")
RRRC_1T_PCL_0_05 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.05, method = "1T-RRRC")

RRFC_1T_PCL_0_2 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.2, method = "1T-RRFC")
RRRC_2T_PCL_0_2 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.2, method = "2T-RRRC")
RRRC_1T_PCL_0_2 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.2, method = "1T-RRRC")

RRFC_1T_PCL_1 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 1, method = "1T-RRFC")
```

```

RRRC_2T_PCL_1 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 1, method = "2T-RRRC")
RRRC_1T_PCL_1 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 1, method = "1T-RRRC")

RRFC_1T_AUC <- StSignificanceTestingCadVsRad (dataset09,
FOM = "Wilcoxon", method = "1T-RRFC")
RRRC_2T_AUC <- StSignificanceTestingCadVsRad (dataset09,
FOM = "Wilcoxon", method = "2T-RRRC")
RRRC_1T_AUC <- StSignificanceTestingCadVsRad (dataset09,
FOM = "Wilcoxon", method = "1T-RRRC")

```

The results are organized as follows:

- RRFC\_1T\_PCL\_0\_05 contains the results of 1T-RRFC analysis for figure of merit =  $PCL_{0.05}$ .
- RRRC\_2T\_PCL\_0\_05 contains the results of 2T-RRFC analysis for figure of merit =  $PCL_{0.05}$ .
- RRRC\_1T\_PCL\_0\_05 contains the results of 1T-RRFC analysis for figure of merit =  $PCL_{0.05}$ .
- RRFC\_1T\_PCL\_0\_2 contains the results of 1T-RRFC analysis for figure of merit =  $PCL_{0.2}$ .
- RRRC\_2T\_PCL\_0\_2 contains the results of 2T-RRRC analysis for figure of merit =  $PCL_{0.2}$ .
- RRRC\_1T\_PCL\_0\_2 contains the results of 1T-RRRC analysis for figure of merit =  $PCL_{0.2}$ .
- RRFC\_1T\_AUC contains the results of 1T-RRFC analysis for the Wilcoxon figure of merit.
- RRRC\_2T\_AUC contains the results of 2T-RRRC analysis for the Wilcoxon figure of merit.
- RRRC\_1T\_AUC contains the results of 1T-RRRC analysis for the Wilcoxon figure of merit.

The structures of these objects are illustrated with examples in the Appendix.

## 38.7 Results

The three methods, in historical order 1T-RRFC, 2T-RRRC and 1T-RRRC, were applied to an LROC dataset similar to that used in Study – 1 (I thank Prof. Karssemeijer for making this dataset available).

Shown next, Table 38.2, are the significance testing results corresponding to the three analyses.

Table 38.2: Significance testing results of the analyses for an LROC dataset. Three sets of results, namely RRRC, 2T-RRRC and 1T-RRRC, are shown for each figure of merit (FOM). Because it is accounting for an additional source of variability, each of the rows labeled RRRC yields a larger p-value and wider confidence intervals than the corresponding row labeled 1T-RRFC. [ $\theta_0$  = FOM CAD;  $\theta_\bullet$  = average FOM of radiologists;  $\psi_\bullet$  = average FOM of radiologists minus CAD; CI= 95 percent confidence interval of quantity indicated by the subscript, F = F-statistic; ddf = denominator degrees of freedom; p = p-value for rejecting the null hypothesis:  $\psi_\bullet = 0$ .]

FOM	Analysis	$\theta_0$	$CI_{\theta_0}$	$\theta_\bullet$	$CI_{\theta_\bullet}$	$\psi_\bullet$	$CI_{\psi_\bullet}$	F	ddf	p
PCL_0_05	1T-RRFC	0	(4.18e-01, 5.68e-01)	4.93e-01	(3.76e-01, 6.11e-01) (2.93e-01, 6.94e-01)	4.33e-02 (-1.57e-01, 2.44e-01)	(-3.16e-02, 1.18e-01) (-1.57e-01, 2.44e-01)	1.77e+00	8e+00	2.2e-01
	2T-RRRC	4.5e-01	(2.58e-01, 6.42e-01)							
	1T-RRRC	NA								
PCL_0_2	1T-RRFC	0	(6.69e-01, 7.51e-01)	7.1e-01	(6.33e-01, 7.87e-01) (5.96e-01, 8.24e-01)	1.19e-01 (4.45e-03, 2.33e-01)	(7.78e-02, 1.59e-01) (4.45e-03, 2.33e-01)	4.5e+01	8e+00	1.51e-04
	2T-RRRC	5.92e-01	(4.78e-01, 7.05e-01)							
	1T-RRRC	NA								
PCL_1	1T-RRFC	0	(7.4e-01, 8.27e-01)	7.83e-01	(7.12e-01, 8.54e-01) (6.8e-01, 8.87e-01)	1.08e-01 (4.5e-03, 2.12e-01)	(6.48e-02, 1.52e-01) (4.5e-03, 2.12e-01)	3.3e+01	8e+00	4.33e-04
	2T-RRRC	6.75e-01	(5.71e-01, 7.79e-01)							
	1T-RRRC	NA								
Wilcoxon	1T-RRFC	0	(8.26e-01, 8.71e-01)	8.49e-01	(8.07e-01, 8.9e-01) (7.86e-01, 9.11e-01)	3.17e-02 (-3.1e-02, 9.45e-02)	(8.96e-03, 5.45e-02) (-3.1e-02, 9.45e-02)	1.03e+01	8e+00	1.24e-02
	2T-RRRC	8.17e-01	(7.52e-01, 8.82e-01)							
	1T-RRRC	NA								

Results are shown for the following FOMs: PCL<sub>0.05</sub>, PCL<sub>0.2</sub>, PCL<sub>1</sub>, and the empirical area (AUC) under the ROC curve estimated by the Wilcoxon statistic. The first two FOMs are identical to those used in Study – 1. Columns 3 and 4 list the CAD FOM  $\theta_0$  and its 95% confidence interval  $CI_{\theta_0}$ , columns 5 and 6 list the average radiologist FOM  $\theta_\bullet$  (the dot symbol represents an average over the radiologist index) and its 95% confidence interval  $CI_{\theta_\bullet}$ , columns 7 and 8 list the average difference FOM  $\psi_\bullet$ , i.e., radiologist minus CAD, and its 95% confidence interval  $CI_{\psi_\bullet}$ , and the last three columns list the F-statistic, the denominator degrees of freedom (ddf) and the p-value for rejecting the null hypothesis. The numerator degree of freedom of the F-statistic, not listed, is unity.

In Table 38.2 identical values in adjacent cells in vertical columns have been replaced by the common values. The last three columns show that 2T-RRRC and 1T-RRRC analyses yield *identical F-statistics, ddf and p-values*. So the intuition of the authors of Study – 2, that the unorthodox method of using DBM – MRMIC software to account for both reader and case-sampling variability,

turns out to be correct. If interest is solely in these statistics one is justified in using the unorthodox method.

Commented on next are other aspects of the results evident in Table 38.2.

1. Where a direct comparison is possible, namely 1T-RRFC analysis using and as FOMs, the p-values in Table 38.2 are similar to those reported in Study – 1.
2. All FOMs (i.e.,  $\theta_0$ ,  $\theta_\bullet$  and  $\psi_\bullet$ ) in Table 38.2 are independent of the method of analysis. However, the corresponding confidence intervals (i.e.,  $CI_{\theta_0}$ ,  $CI_{\theta_\bullet}$  and  $CI_{\psi_\bullet}$ ) depend on the analyses.
3. Since 1T-RRFC analysis ignores case sampling variability, the CAD figure of merit is a constant, with zero-width confidence interval. For compactness the CI is listed as 0, rather than two identical values in parentheses. The confidence interval listed for 2T-RRRC analyses is centered on the corresponding CAD value, as are all confidence intervals in Table 38.2.
4. The LROC FOMs increase as the value of FPF (the subscript) increases. This should be obvious, as PCL increases as FPF increases, a general feature of any partial curve based figure of merit.
5. The area (AUC) under the ROC is larger than the largest PCL value, i.e.,  $AUC \geq PCL_1$ . This too should be obvious from the general features of the LROC (Swensson, 1996b).
6. The p-value for either RRRC analyses (2T or 1T) is larger than the corresponding 1T-RRFC value. Accounting for case-sampling variability increases the p-value, leading to less possibility of finding a significant difference.
7. Partial curve-based FOMs, such as  $PCL_{FPF}$ , lead, depending on the choice of  $FPF$ , to different conclusions. The p-values generally decrease as FPF increases. Measuring performance on the steep part of the LROC curve (i.e., small FPF) needs to account for greater reader variability and risks lower statistical power.
8. Ignoring localization information (i.e., using the AUC FOM) led to a non-significant difference between CAD and the radiologists ( $p = 0.3210$ ), while the corresponding FOM yielded a significant difference ( $p = 0.0409$ ). Accounting for localization leads to a less “noisy” measurement. This has been demonstrated for the LROC paradigm (Swensson, 1996b) and I have demonstrated this for the FROC paradigm (Chakraborty, 2008).
9. For 1T-RRRC analysis, is listed as NA, for not applicable, since is not a model parameter, see Eqn. (38.16).

Shown next, Table 38.3, are the model-parameters corresponding to the three analyses.

Table 38.3: Parameter estimates for the analyses; NA = not applicable.

FOM	Analysis	$\sigma_R^2$	$\sigma_{\tau R}^2$	Cov1	Cov2	Cov3	Var
PCL_0_05	1T-RRFC	9.5e-03	NA	NA	NA	NA	NA
	2T-RRRC	1.84e-18	-5.71e-03	1.31e-03	6.01e-03	1.31e-03	1.65e-02
	1T-RRRC	9.5e-03	NA	NA	9.4e-03	NA	3.03e-02
PCL_0_2	1T-RRFC	2.81e-03	NA	NA	NA	NA	NA
	2T-RRRC	-7.59e-19	2.65e-04	7.61e-04	2.29e-03	7.61e-04	3.43e-03
	1T-RRRC	2.81e-03	NA	NA	3.07e-03	NA	5.34e-03
PCL_1	1T-RRFC	3.2e-03	NA	NA	NA	NA	NA
	2T-RRRC	1.63e-18	1e-03	6.43e-04	1.86e-03	6.43e-04	2.46e-03
	1T-RRRC	3.2e-03	NA	NA	2.44e-03	NA	3.64e-03
Wilcoxon	1T-RRFC	8.78e-04	NA	NA	NA	NA	NA
	2T-RRRC	2.98e-19	2.01e-04	2.62e-04	7.24e-04	2.62e-04	9.62e-04
	1T-RRRC	8.78e-04	NA	NA	9.24e-04	NA	1.4e-03

The following characteristics are evident from Table 38.3.

1. For 2T-RRRC analyses  $\sigma_R^2 = 0$ . Actually, the analysis yielded very small values, of the order of  $10^{-18}$  to  $10^{-19}$ , which, being smaller than double precision accuracy, were replaced by zeroes in Table 38.2.  $\sigma_R^2 = 0$  is clearly an incorrect result as the radiologists do not have identical performance. In contrast, 1T-RRRC analyses yielded more realistic values, identical to those obtained by 1T-RRFC analyses, and consistent with expectation – see comment following Eqn. (15).
2. Because 2T analysis found zero reader variability, it follows from the definitions of the covariances (Obuchowski and Rockette, 1995b), that  $Cov_1 = Cov_3 = 0$ , as evident in the table.
3. When they can be compared (i.e.,  $\sigma_R^2$ , Cov<sub>2</sub> and Var), all variance and covariance estimates were smaller for the 2T method than for the 1T method.
4. For the 2T method the expected inequalities, Eqn. (38.8), are not obeyed (specifically,  $Cov_1 \geq Cov_2 \geq Cov_3$  is not obeyed).

For an analysis method to be considered statistically valid it needs to be tested with simulations to determine if it has the proper null hypothesis behavior. The design of a ratings simulator to statistically match a given dataset is addressed in Chapter 23 of reference (Chakraborty, 2017). Using this simulator, the 1T-RRRC method had the expected null hypothesis behavior (Table 23.5, ibid).

## 38.8 Discussion

TBA TODOLAST The argument often made for not measuring standalone performance is that since CAD will be used only as a second reader, it is only necessary to measure performance of radiologists without and with CAD. It has been stated (Nishikawa and Pesce, 2011):

High stand-alone performance is neither a necessary nor a sufficient condition for CAD to be truly useful clinically.

Assessing CAD utility this way, i.e., by measuring performance with and without CAD, may have inadvertently set a low bar for CAD to be considered useful. As examples, CAD is not penalized for missing cancers as long as the radiologist finds them and CAD is not penalized for excessive false positives (FPs) as long as the radiologist ignores them. Moreover, since both such measurements include the variability of radiologists, there is additional noise introduced that presumably makes it harder to determine if the CAD system is optimal.

Described is an extension of the analysis used in Study – 1 that accounts for case sampling variability. It extends (Hillis et al., 2005b) single-treatment analysis to a situation where one of the “readers” is a special reader, and the desire is to compare performance of this reader to the average of the remaining readers. The method, along with two other methods, was used to analyze an LROC data set using different figures of merit.

1T-RRRC analyses yielded identical overall results (specifically the F-statistic, degrees of freedom and p-value) to those yielded by the unorthodox application of DBM-MRMC software, termed 2T-RRRC analyses, where the CAD reader is regarded as a second treatment. However, the values of the model parameters of the dual-treatment analysis lacked clear physical meanings. In particular, the result  $\sigma_R^2 = 0$  is clearly an artifact. One can only speculate as to what happens when software is used in a manner that it was not designed for: perhaps finding that all readers in the second treatment have identical FOMs led the software to yield  $\sigma_R^2 = 0$ . The single-treatment model has half as many parameters as the dual-treatment model and the parameters have clear physical meanings and the values are realistic.

The paradigm used to collect the observer performance data - e.g., receiver operating characteristic (ROC) (Metz, 1986), free-response ROC (FROC) (Chakraborty et al., 1986), location ROC (LROC) (Starr et al., 1975) or region of interest (ROI) (Obuchowski et al., 2000) - is irrelevant – all that is needed is a scalar performance measure for the actual paradigm used. In addition to PCL and AUC, RJafroc currently implements the partial area under the LROC, from FPF = 0 to a specified value as well other FROC-paradigm based FOMs.

While there is consensus that CAD works for microcalcifications, for masses its performance is controversial<sup>27,28</sup>. Two large clinical studies TBA 29,30

(222,135 and 684,956 women, respectively) showed that CAD actually had a detrimental effect on patient outcome. A more recent large clinical study has confirmed the negative view of CAD31 and there has been a call for ending Medicare reimbursement for CAD interpretations32.

In my opinion standalone performance is the most direct measure of CAD performance. Lack of clear-cut methodology to assess standalone CAD performance may have limited past CAD research. The current work hopefully removes that impediment. Going forward, assessment of standalone performance of CAD vs. expert radiologists is strongly encouraged.

## 38.9 Appendix

The structures of the R objects generated by the software are illustrated with three examples.

### 38.9.1 Example 1

The first example shows the structure of ‘RRFC\_1T\_PCL\_0\_2

```
print(fom_individual_rad)
#>      rdr1  rdr2  rdr3  rdr4      rdr5      rdr6  rdr7  rdr8  rdr9
#> 1 0.69453125 0.65 0.80625 0.725 0.65982143 0.76845238 0.7375 0.675 0.675
print(stats)
#>      fomCAD  avgRadFom avgDiffFom      varR      Tstat df      pval
#> 1 0.59166667 0.71017278 0.11850612 0.002808612 6.7083568 8 0.0001513964
print(ConfidenceIntervals)
#>      CIAvgRadFom CIAvgDiffFom
#> Lower  0.66943619 0.077769525
#> Upper  0.75090938 0.159242710
```

The results are displayed as three data frames.

The first data frame :

- `fom_individual_rad` shows the figures of merit for the nine radiologists in the study.

The next data frame summarizes the statistics.

- `fomCAD` is the figure of merit for CAD.
- `avgRadFom` is the average figure of merit of the nine radiologists in the study.

- `avgDiffFom` is the average difference figure of merit, RAD - CAD.
- `varR` is the variance of the figures of merit for the nine radiologists in the study.
- `Tstat` is the t-statistic for testing the NH that the average difference FOM `avgDiffFom` is zero, whose square is the F-statistic.
- `df` is the degrees of freedom of the t-statistic.
- `pval` is the p-value for rejecting the NH. In the example shown below the value is highly significant.

The last data frame summarizes the 95 percent confidence intervals.

- `CIAvgRadFom` is the 95 percent confidence interval, listed as pairs `Lower`, `Upper`, for `avgRadFom`.
- `CIAvgDiffFom` is the 95 percent confidence interval for `avgDiffFom`.
- If the pair `CIAvgDiffFom` excludes zero, the difference is statistically significant.
- In the example the interval excludes zero showing that the FOM difference is significant.

### 38.9.2 Example 2

The next example shows the structure of `RRRC_2T_PCL_0_2`.

```
print(fom_individual_rad)
#>      rdr1  rdr2  rdr3  rdr4      rdr5      rdr6  rdr7  rdr8  rdr9
#> 1 0.69453125 0.65 0.80625 0.725 0.65982143 0.76845238 0.7375 0.675 0.675
print(stats1)
#>      fomCAD  avgRadFom  avgDiffFom
#> 1 0.59166667 0.71017278 0.11850612
print(stats2)
#>      varR      varTR      cov1      cov2      cov3
#> 1 -7.5894152e-19 0.00026488983 0.00076136841 0.0022942211 0.00076136841
#>      Var      FStat      df      pval
#> 1 0.0034336373 4.1576797 937.24371 0.041726262
```

In addition to the quantities defined previously, the output contains the covariance matrix for the Obuchowski-Rockette model, summarized in Eqn. (38.3) – Eqn. (38.6).

- `varTR` is  $\sigma_{\tau R}^2$ .
- `cov1` is  $\text{Cov}_1$ .
- `cov2` is  $\text{Cov}_2$ .
- `cov3` is  $\text{Cov}_3$ .

- **Var** is Var.
- **FStat** is the F-statistic for testing the NH.
- **ndf** is the numerator degrees of freedom, equal to unity.
- **df** is denominator degrees of freedom of the F-statistic for testing the NH.
- **Tstat** is the t-statistic for testing the NH that the average difference FOM **avgDiffFom** is zero.
- **pval** is the p-value for rejecting the NH. In the example shown below the value is significant.

Notice that including the variability of cases results in a higher p-value for 2T-RRRC as compared to 1T-RRFC.

Shown next are the confidence interval statistics **x\$ciAvgRdrEachTrt** for the two treatments (“trt1” = CAD, “trt2” = RAD):

```
print(x$ciAvgRdrEachTrt)
#>           Estimate      StdErr       DF    CILower    CIUpper      Cov2
#> trt1 0.59166667 0.058028349      Inf 0.47793319 0.70540014 0.0033672893
#> trt2 0.71017278 0.039156365 193.10832 0.63294372 0.78740185 0.0012211529
```

- **Estimate** contains the difference FOM estimate.
- **StdErr** contains the standard estimate of the difference FOM estimate.
- **DF** contains the degrees of freedom of the t-statistic.
- **t** contains the value of the t-statistic.
- **PrGTt** contains the probability of exceeding the magnitude of the t-statistic.
- **CILower** is the lower confidence interval for the difference FOM.
- **CIUpper** is the upper confidence interval for the difference FOM.

Shown next are the confidence interval statistics **x\$ciDiffFom** between the two treatments (“trt1-trt2” = CAD - RAD):

```
print(x$ciDiffFom)
#>           Estimate      StdErr       DF          t      PrGTt      CILower
#> trt2-trt1 0.11850612 0.058118615 937.24371 2.0390389 0.041726262 0.004448434
#>           CIUpper
#> trt2-trt1 0.2325638
```

The difference figure of merit statistics are contained in a dataframe **x\$ciDiffFom** with elements:

- **Estimate** contains the difference FOM estimate.
- **StdErr** contains the standard estimate of the difference FOM estimate.
- **DF** contains the degrees of freedom of the t-statistic.

- `t` contains the value of the t-statistic.
- `PrGtt` contains the probability of exceeding the magnitude of the t-statistic.
- `CILower` is the lower confidence interval for the difference FOM.
- `CIUpper` is the upper confidence interval for the difference FOM.

The figures of merit statistic for the two treatments, 1 is CAD and 2 is RAD.

- `trt1`: statistics for CAD.
- `trt2`: statistics for RAD.
- `Cov2`: Cov<sub>2</sub> calculated over individual treatments.

### 38.9.3 Example 3

The last example shows the structure of `RRRC_1T_PCL_0_2`.

```
RRRC_1T_PCL_0_2
#> $fomCAD
#> [1] 0.59166667
#>
#> $fomRAD
#> [1] 0.69453125 0.65000000 0.80625000 0.72500000 0.65982143 0.76845238 0.73750000
#> [8] 0.67500000 0.67500000
#>
#> $avgRadFom
#> [1] 0.71017278
#>
#> $CIAvgRad
#> [1] 0.59611510 0.82423047
#>
#> $avgDiffFom
#> [1] 0.11850612
#>
#> $CIAvgDiffFom
#> [1] 0.004448434 0.232563801
#>
#> $varR
#> [1] 0.002808612
#>
#> $varError
#> [1] 0.0053445377
#>
#> $cov2
#> [1] 0.0030657054
```

```
#>
#> $Tstat
#>      rdr2
#> 2.0390389
#>
#> $df
#>      rdr2
#> 937.24371
#>
#> $pval
#>      rdr2
#> 0.041726262
```

The differences from RRFC\_1T\_PCL\_0\_2 are listed next:

- **varR** is  $\sigma_R^2$  of the single treatment model for comparing CAD to RAD, Eqn. (38.17).
- **cov2** is Cov<sub>2</sub> of the single treatment model for comparing CAD to RAD.
- **varError** is Var of the single treatment model for comparing CAD to RAD.

Notice that the RRRC\_1T\_PCL\_0\_2 p value, i.e., 0.04172626, is identical to that of RRRC\_2T\_PCL\_0\_2, i.e., 0.04172626.

### 38.10 References

# Chapter 39

## Optimal operating point on FROC

### 39.1 TBA How much finished

80%

Discussion and Intro need more work; coding is done

### 39.2 Introduction

This chapter deals with finding the optimal reporting threshold of an algorithmic observer, such as CAD. We assume that designer level FROC data is available for the algorithm, i.e., the data consists of mark-rating pairs, with continuous-scale ratings, and a decision needs to be made as to the optimal reporting threshold, i.e., the minimum rating of a mark before it is shown to the radiologist. This is a familiar problem faced by a CAD algorithm designer.

The problem has been solved in the context of ROC analysis (Metz, 1978), namely, the optimal operating point on the ROC corresponds to a slope determined by disease prevalence and the cost of decisions in the four basic binary paradigm categories: true and false positives and true and false negatives. In practice the costs are difficult to quantify. However, for equal numbers of diseased and non-diseased cases and equal costs it can be shown that the slope of the ROC curve at the optimal point is unity. For a proper ROC curve this corresponds to the point that maximizes the Youden-index (Youden, 1950), defined as the sum of sensitivity and specificity minus one. Typically it is maximized at the point that is closest to the (0,1) corner of the ROC.

CAD produces FROC data and lacking a procedure for setting it analytically, CAD manufacturers, in consultation with radiologists, set site-specific reporting thresholds. For example, if radiologists at a site are comfortable with more false marks as the price of potentially greater lesion-level sensitivity, the reporting threshold for them is adjusted downward.

This chapter describes an analytic method for finding the optimal reporting threshold. The method is based on maximizing AUC (area under curve) under the wAFROC curve. The method is compared to the Youden-index based method.

### 39.3 Methods

The ROC, FROC and wAFROC curves are completely defined by the RSM (radiological search model) parameters:  $\lambda$ ,  $\nu$ ,  $\mu$  and  $\zeta_1$ , which have the following meanings:

- The  $\mu$  parameter is the perceptual signal to noise ratio of lesions measured under location-known-exactly conditions. Higher values of  $\mu$  lead to increased overall performance of the algorithm.
- The intrinsic  $\lambda$  parameter determines the number of non-lesion localizations, NLs, per case (location level “false positives”). Lower values lead to fewer NL marks and increased algorithm performance. It is related to the physical  $\lambda'$  parameter by  $\lambda' = \lambda/\mu$ . The physical parameter  $\lambda'$  equals the mean of the assumed Poisson distribution of NLs per case.
- The intrinsic  $\nu$  parameter determines the probability of a lesion localizations, LLs, (location level “true positives”). Higher values lead to more LL marks. It is related to the physical  $\nu'$  parameter by  $\nu' = 1 - \exp(-\mu\nu)$ . The physical parameter  $\nu'$  equals the success probability of the assumed binomial distribution of LLs per case.
- The  $\zeta_1$  parameter determines if a suspicious region found by the algorithm is actually marked. The higher this value, the fewer the reported marks. The objective is to optimize  $\zeta_1$ .

In the following sections each of the first three parameters is varied in turn and the corresponding optimal  $\zeta_1$  determined by maximizing one of two figures of merit (FOMs), namely, the wAFROC-AUC and the Youden-index.

#### 39.3.1 Functions to be maximized

The functions to be maximized, wAFROC and Youden, are defined next:

- wAFROC-AUC is computed by `UtilAnalyticalAucsRSM`. Lines 2 - 19 returns `-wAFROC`, the *negative* of wAFROC-AUC. The negative sign is needed because the `optimize()` function, used later, finds the *minimum* of wAFROC-AUC. The first argument is  $\zeta_1$ , the variable to be varied to find the maximum. The remaining arguments passed to the function, needed to calculate the FOMs, are  $\mu$ ,  $\lambda$ ,  $\nu$ , `lesDistr` and `relWeights`. The last two specify the number of lesions per case and their weights. The following code below uses `lesDistr = c(0.5,0.5)`, i.e., half of the diseased cases contain one lesion and the rest contain two lesions, and `relWeights = c(0.5,0.5)`, which specifies equal weights to all lesions.
- The Youden-index is defined as the sum of sensitivity and specificity minus 1. Sensitivity is computed by `RSM_yROC` and specificity by `(1 - RSM_xROC)`. Lines 22 - 42 returns `-Youden`, the *negative* of the Youden-index.

```

1   wAFROC <- function (
2     zeta1,
3     mu,
4     lambda,
5     nu,
6     lesDistr,
7     relWeights) {
8     x <- UtilAnalyticalAucsRSM(
9       mu,
10      lambda,
11      nu, zeta1,
12      lesDistr,
13      relWeights)$aucwAFROC
14
15   # return negative of aucwAFROC
16   # (as optimize finds minimum of function)
17   return(-x)
18
19 }
20
21
22 Youden <- function (
23   zeta1,
24   mu,
25   lambda,
26   nu,
27   lesDistr,
28   relWeights) {
29   # add sensitivity and specificity
30   # and subtract 1, i.e., Youden's index

```

```

31   x <- RSM_yROC(
32     zeta1,
33     mu,
34     lambda,
35     nu,
36     lesDistr) +
37     (1 - RSM_xROC(zeta1, lambda/mu)) - 1
38   # return negative of Youden-index
39   # (as optimize finds minimum of function)
40   return(-x)
41
42 }
```

### 39.3.2 Vary lambda

For  $\mu = 2$  and  $\nu = 1$ , wAFROC-AUC and Youden-index based optimizations were performed for  $\lambda = 1, 5, 10, 15$ . The following quantities were calculated:

- `zetaOptArr`, a [2,4] array, the optimal thresholds  $\zeta_1$ ;
- `fomMaxArr`, a [2,4] array, the maximized values of wAFROC-AUC, using either wAFROC based or Youden-index based optimization; note that in the latter we report wAFROC-AUC even though the optimized quantity is the Youden-index.
- `rocAucArr`, a [2,4] array, the AUCs under the ROC curves corresponding to optimizations based on wAFROC-AUC or Youden-index;
- `nlfOptArr`, a [2,4] array, the abscissa of the optimal reporting point on the FROC curve corresponding to optimizations based on wAFROC-AUC or Youden-index;
- `llfOptArr`, a [2,4] array, the ordinate of the optimal reporting point on the FROC curve corresponding to optimizations based on wAFROC-AUC or Youden-index.

In each of these arrays the first index, `y` in the following code, denotes whether wAFROC-AUC is being maximized (`y = 1`, see lines 14 - 20) - or if Youden-index is being optimized (`y = 2`, see lines 39 - 45). The second index `i` in the following code, corresponds to  $\lambda$ .

```

1 mu <- 2
2 nu <- 1
3 lambdaArr <- c(1,5,10,15)
4 fomMaxArr <- array(dim = c(2,length(lambdaArr)))
5 zetaOptArr <- array(dim = c(2,length(lambdaArr)))
```

```

6  rocAucArr <- array(dim = c(2,length(lambdaArr)))
7  nlfOptArr <- array(dim = c(2,length(lambdaArr)))
8  llfOptArr <- array(dim = c(2,length(lambdaArr)))
9  lesDistr <- c(0.5, 0.5)
10 relWeights <- c(0.5, 0.5)
11 for (y in 1:2) {
12   for (i in 1:length(lambdaArr)) {
13     if (y == 1) {
14       x <- optimize(wAFROC,
15                     interval = c(-5,5),
16                     mu,
17                     lambdaArr[i],
18                     nu,
19                     lesDistr,
20                     relWeights)
21     zetaOptArr[y,i] <- x$minimum
22     fomMaxArr[y,i] <- -x$objective # safe to use objective here
23     rocAucArr[y,i] <- UtilAnalyticalAucsRSM(
24       mu,
25       lambdaArr[i],
26       nu,
27       zeta1 = x$minimum,
28       lesDistr,
29       relWeights)$aucROC
30     nlfOptArr[y,i] <- RSM_xFROC(
31       z = x$minimum,
32       mu,
33       lambda = lambdaArr[i])
34     llfOptArr[y,i] <- RSM_yFROC(
35       z = x$minimum,
36       mu,
37       nu)
38   } else if (y == 2) {
39     x <- optimize(Youden,
40                   interval = c(-5,5),
41                   mu,
42                   lambdaArr[i],
43                   nu,
44                   lesDistr,
45                   relWeights)
46     zetaOptArr[y,i] <- x$minimum
47     fomMaxArr[y,i] <- UtilAnalyticalAucsRSM(
48       mu,
49       lambdaArr[i],
50       nu,

```

```

51     zeta1 = x$minimum,
52     lesDistr,
53     relWeights)$aucwAFROC
54     rocAucArr[y,i] <- UtilAnalyticalAucsRSM(
55         mu,
56         lambdaArr[i],
57         nu,
58         zeta1 = x$minimum,
59         lesDistr,
60         relWeights)$aucROC
61     nlfOptArr[y,i] <- RSM_xFROC(
62         z = x$minimum,
63         mu,
64         lambda = lambdaArr[i])
65     llfOptArr[y,i] <- RSM_yFROC(
66         z = x$minimum, mu, nu)
67     } else stop("incorrect y")
68   }
69 }
```

Table 39.1 summarizes the results. The column labeled “FOM” shows the quantity being maximized, “lambda” corresponds to the 4 values of  $\lambda$ , “zeta1” is the optimal value of  $\zeta_1$  that maximizes FOM, “wAFROC” is the wAFROC-AUC, “ROC” is the AUC under the ROC curve, i.e., ROC-AUC, and “OptOpPt” is the optimal operating point on the FROC curve.

For the wAFROC-AUC based optimizations (first four rows of table), as  $\lambda$  increases:

- The optimal threshold  $\zeta_1$  increases;
- wAFROC-AUC decreases;
- ROC-AUC decreases;
- The optimal operating point moves to lower LLF values, i.e., lower values of location-level “sensitivity”.
- The advantage of wAFROC-AUC over Youden-index based optimizations, as measured by the differences between the corresponding wAFROC-AUCs, decreases with increasing  $\lambda$ : `fomMaxArr[1,] - fomMaxArr[2,]` = 0.024, 0.018, 0.007, 0.001, where the successive values correspond to  $\lambda = 1, 5, 10, 15$ .

The  $\lambda'$  Poisson parameter controls the average number of perceived NLs per case. For example, for  $\mu = 2$  and  $\lambda = 1$ , the average number is  $\lambda' = \lambda/\mu = 0.5$ , i.e., an average of one perceived NL every two non-diseased case. With increasing numbers of NLs per case it is necessary to increase the reporting threshold and LLF consequently decreases. Also, overall CAD performance, regardless of how it is measured (i.e., wAFROC-AUC or ROC-AUC), decreases.

Similar trends are observed for the Youden-index based optimizations (last four rows of table). However, Youden-index based optimizations compared as a group to wAFROC-AUC based optimizations show that Youden yields higher reporting thresholds, lower wAFROC-AUC, lower ROC-AUC and lower LLF values.

Table 39.1: Summary of optimization results for  $\mu = 2$ ,  $\nu = 1$  and different values of  $\lambda$ . The wAFROC column always displays wAFROC-AUC, even though the optimized quantity may the Youden-index, as in the last four rows.

FOM	lambda	zeta1	wAFROC	ROC	OptOpPt
wAFROC	1	-0.235	0.880	0.937	(0.296, 0.854)
	5	0.810	0.768	0.875	(0.522, 0.763)
	10	1.373	0.699	0.825	(0.424, 0.635)
	15	1.697	0.660	0.788	(0.336, 0.535)
Youden	1	0.802	0.856	0.915	(0.106, 0.765)
	5	1.438	0.750	0.842	(0.188, 0.616)
	10	1.690	0.693	0.801	(0.227, 0.538)
	15	1.832	0.658	0.776	(0.251, 0.490)

One could display 8 FROC plots, each corresponding to a row of the preceding table, but there is a more efficient method. The FROC curve is defined in terms of the RSM parameters as follows:

$$\left. \begin{aligned} NLF(\zeta, \lambda') &= \lambda' \Phi(-\zeta) \\ LLF(\zeta, \mu, \nu', \vec{f}_L) &= \nu' \Phi(\mu - \zeta) \end{aligned} \right\} \quad (39.1)$$

Here  $\vec{f}_L$  is the lesion-distribution vector,  $c(0.5, 0.5)$  in the current example.

The *end-point* of the FROC defined by  $(\lambda', \nu')$  is not to be confused with the *optimal* value of  $\zeta_1$ ; the former corresponds to  $\zeta_1 = -\infty$  while the latter is a finite value of  $\zeta_1$  as found by the optimization procedure.

Since the  $\Phi$  function ranges from one to unity, the *four FROC curves for different values of  $\lambda$  are scaled versions of a single curve whose x-axis ranges from 0 to 1*. The single curve corresponds to  $\lambda' = 1$  and the true curves are obtained by scaling this curve along the x-axis by the appropriate  $\lambda'$  factor. With this understanding one can display 4 FROC curves with a single FROC curve where the x-axis is  $NLF(\zeta, \lambda' = 1)$ . The true FROC curve is defined by:

$$\left. \begin{aligned} NLF(\zeta, \lambda') &= \lambda' NLF(\zeta, \lambda' = 1) \\ LLF(\zeta, \mu, \nu', \bar{f}_L) &= \nu' \Phi(\mu - \zeta) \end{aligned} \right\} \quad (39.2)$$

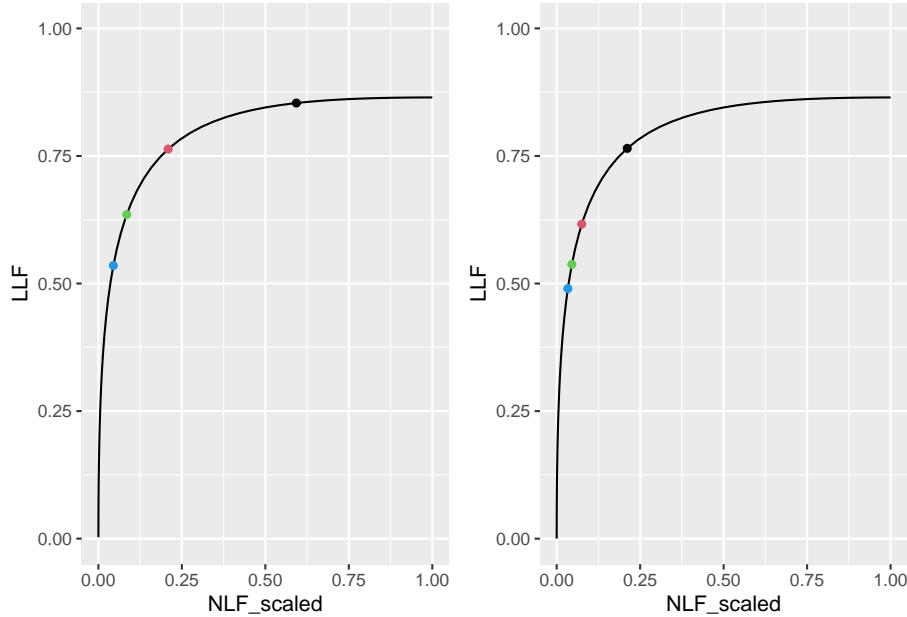


Figure 39.1: Left panel: maximized wAFROC AUC was used to find optimal  $\zeta_1$ . Right panel: maximized Youden-index was used to find optimal  $\zeta_1$ . Dot colors: black means  $\lambda = 1$ , red means  $\lambda = 5$ , green means  $\lambda = 10$  and blue means  $\lambda = 15$ .

The left panel in 39.1 shows the optimal operating points when wAFROC-AUC is maximized. The 4 operating points are color coded as follows:

- The black dot corresponds to  $\lambda = 1$ , i.e.,  $\lambda' = 1/2 = 0.5$ . In other words, the true FROC is obtained by *shrinking* the plot shown, including the superposed black dot, along the x-axis by a factor of 2.
- The red dot corresponds to  $\lambda' = 2.5$ . In other words, the true FROC is obtained by *magnifying* that shown, including the red dot, along the x-axis by a factor of 2.5.
- The green dot corresponds to  $\lambda' = 5$ .
- The blue dot corresponds to  $\lambda' = 7.5$ .

These plots illustrate the previous comments, namely, as  $\lambda$  increases, *the optimal operating point moves down the scaled curve*.

The right panel shows the optimal operating point when the Youden-index is maximized. It shows the same general features as the previous example but the group of four operating points in the right panel are below-left those in the left panel, representing higher values of optimal  $\zeta_1$ , i.e., a more stringent criteria. As seen in the preceding table the overly strict criteria, using Youden-index based optimization, leads to lower true performance: i.e., lower wAFROC-AUC and lower ROC-AUC, and lower LLF.

The FROC curve does not represent true performance. To visualize true performance one compares wAFROC curves.

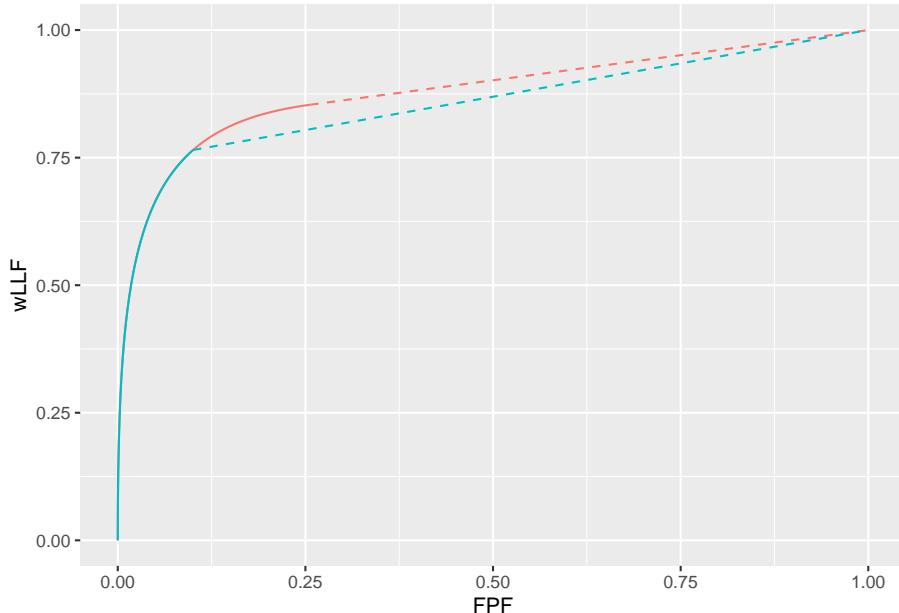


Figure 39.2: wAFROC curves for wAFROC-AUC and Youden-index based optimizations: both curves correspond to  $\mu = 2$ ,  $\nu = 1$  and  $\lambda = 1$ . The optimal reporting threshold  $\zeta_1$  is determined by the selected FOM. The red curve corresponds to FOM = wAFROC-AUC and the blue curve corresponds to FOM = Youden-index. The stricter reporting threshold found by the Youden-index based method sacrifices a considerable amount of area under the wAFROC. The two wAFROC-AUCs are 0.880 and 0.856, respectively.

Each curve ends at the optimal threshold listed in Table 39.1, namely  $\zeta_1 = -0.235$  for the red curve and  $\zeta_1 = 0.802$  for the blue curve. The lower performance represented by the blue curve, based on Youden-index maximization, is due to

the adoption of an overly strict threshold.

### 39.3.3 Vary nu

For  $\mu = 2$  and  $\lambda = 5$ , wAFROC-AUC and Youden-index based optimizations were performed for  $\nu = 0.1, 0.5, 1, 2$ . Table 39.2 summarizes the results.

Table 39.2: Summary of optimization results for  $\mu = 2$ ,  $\lambda = 5$  and different values of  $\nu$ . The wAFROC column always displays wAFROC-AUC, even though the optimized quantity may be the Youden-index, as in the last four rows.

FOM	nu	zeta1	wAFROC	ROC	OptOpPt
wAFROC	0.1	2.275	0.522	0.551	(0.029, 0.071)
	0.5	1.376	0.660	0.771	(0.211, 0.464)
	1	0.810	0.768	0.875	(0.522, 0.763)
	2	-0.311	0.841	0.915	(1.555, 0.971)
Youden	0.1	1.336	0.473	0.588	(0.227, 0.135)
	0.5	1.398	0.660	0.770	(0.203, 0.459)
	1	1.438	0.750	0.842	(0.188, 0.616)
	2	1.461	0.793	0.874	(0.180, 0.692)

Focusing on the wAFROC-AUC based optimizations (first four rows of table), as  $\nu$  increases:

- The optimal threshold  $\zeta_1$  decreases, resulting in more marks being reported; wAFROC-AUC increases; ROC-AUC increases and the optimal operating point on the FROC moves to higher LLF values, i.e., higher values of lesion-level “sensitivity”.

All of these are opposite to the effect of increasing  $\lambda$ . The  $\nu'$  binomial success probability parameter is the probability of a perceived LL event. For example, for  $\mu = 2$  and  $\nu = 0.1$ ,  $\nu' = 1 - \exp(-\mu\nu) = 0.1812692$ , i.e., an average of 18 percent of lesions present are found by the algorithm at the *initial detection* stage, using terminology in (Edwards et al., 2002).

With one exception similar trends are observed for the Youden-index based optimizations (last four rows of table). As a group Youden-index based optimizations (last four rows of table) compared to wAFROC-AUC based optimizations

show that the former yields higher reporting thresholds, lower wAFROC-AUC, lower ROC-AUC and lower LLF values.

The exception is that as  $\nu$  increases the optimal threshold increases, but more slowly. The increasing separation of the two underlying probability density functions that generate the ROC causes the optimal threshold to increase (similar to the explanation in Section 39.3.4).

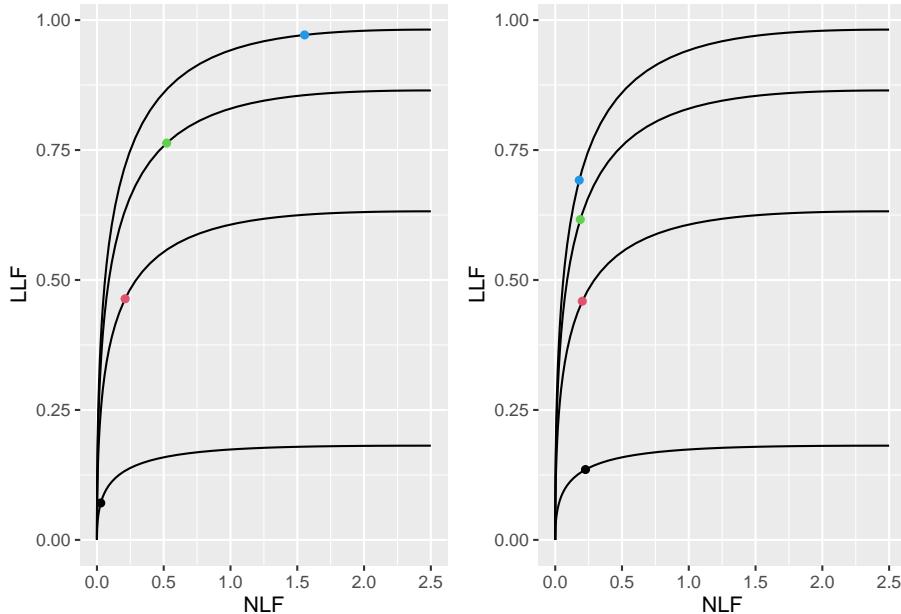


Figure 39.3: Left panel: maximized wAFROC-AUC was used to find optimal  $\zeta_1$ . Right panel: maximized Youden-index was used to find optimal  $\zeta_1$ . Dot colors: black means  $\nu = 0.1$ , red means  $\nu = 0.5$ , green means  $\nu = 1$  and blue means  $\nu = 2$ .

Fig. 39.3 shows the FROC curves with optimal operating points superimposed. The left panel corresponds to wAFROC-AUC based optimizations while the right panel corresponds to Youden-index based optimizations. These illustrate the previous comments, namely, as  $\nu$  increases, *the optimal operating point moves up the FROC curve*.

To visualize true performance one compares wAFROC curves.

Each curve ends at the optimal threshold listed in Table 39.2, namely  $\zeta_1 = -0.311$  for the red curve and  $\zeta_1 = 1.461$  for the blue curve. The lower performance represented by the blue curve, based on Youden-index maximization, is due to the adoption of an overly strict threshold.

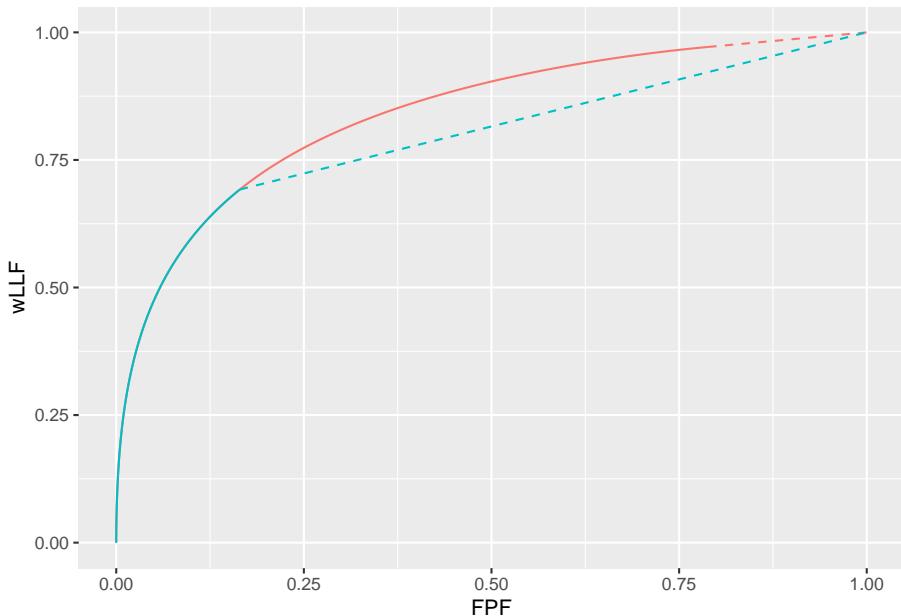


Figure 39.4: wAFROC curves for wAFROC-AUC and Youden-index based optimizations: both curves correspond to  $\mu = 2$ ,  $\lambda = 5$  and  $\nu = 2$ . The optimal reporting threshold  $\zeta_1$  is determined by the selected FOM. The red curve corresponds to FOM = wAFROC-AUC and the blue curve corresponds to FOM = Youden-index. The stricter reporting threshold found by the Youden-index based method sacrifices a considerable amount of area under the wAFROC. The two wAFROC-AUCs are 0.841 and 0.793, respectively.

### 39.3.4 Vary mu

For  $\nu = 1$  and  $\lambda = 1$  wAFROC-AUC and Youden-index based optimizations were performed for 4 values of  $\mu = 0.75, 1, 1.25, 1.5$ . Table 39.2 summarizes the results.

Table 39.3: Summary of optimization results for  $\nu = 1$ ,  $\lambda = 1$  and different values of  $\mu$ . The wAFROC column always displays wAFROC-AUC, even though the optimized quantity may be the Youden-index, as in the last four rows.

FOM	mu	zeta1	wAFROC	ROC	OptOpPt
wAFROC	0.75	1.422	0.518	0.587	(0.103, 0.132)
	1	0.310	0.603	0.745	(0.378, 0.477)
	1.25	-0.132	0.699	0.823	(0.442, 0.654)
	1.5	-0.268	0.777	0.875	(0.404, 0.747)
Youden	0.75	0.367	0.493	0.668	(0.476, 0.343)
	1	0.386	0.603	0.741	(0.350, 0.462)
	1.25	0.461	0.691	0.802	(0.258, 0.560)
	1.5	0.563	0.760	0.850	(0.191, 0.641)

Increasing  $\mu$ , while holding  $\lambda$  and  $\nu$  constant, *simultaneously decreases*  $\lambda'$  and increases  $\mu'$ . As the latter two parameters work in opposite directions (increasing one has a similar effect as decreasing the other) the simultaneous changes result in an amplified effect. The values in the table can be understood from this.

For the wAFROC-AUC based optimizations (first four rows of table), as  $\mu$  increases the reporting threshold  $\zeta_1$  decreases, both wAFROC-AUC and ROC-AUC increase, and the optimal operating point moves to higher LLF values.

For the Youden-index based optimizations (last four rows of table), as  $\mu$  increases the reporting threshold  $\zeta_1$  increases (but the magnitude of the change is smaller than for the first four rows), both wAFROC-AUC and ROC-AUC increase, and the optimal operating point moves to higher LLF values.

The effect of increasing  $\mu$  can be understood as resulting from the competing effects of *greater search performance*, greater numbers of LLs and fewer NLs, both allowing the threshold to be moved down, and *greater classification performance*, allowing the threshold to be moved up (as the separation of two unit

normal distribution increases, the optimal threshold for discriminating between them increases).

Fig. 39.5 shows FROC curves with superimposed optimal operating points.

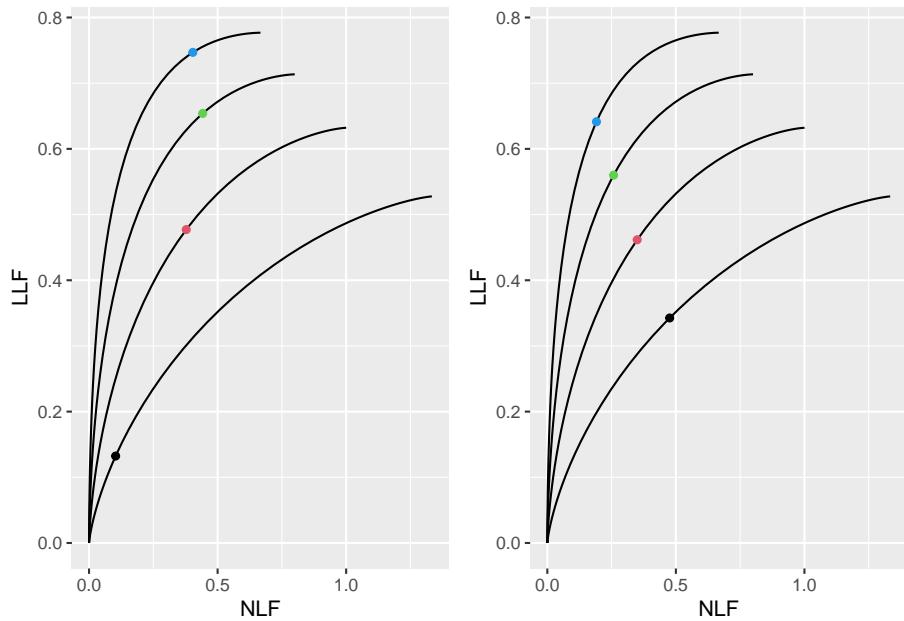


Figure 39.5: Left panel: maximized wAFROC-AUC was used to find optimal  $\zeta_1$ . Right panel: maximized Youden-index was used to find optimal  $\zeta_1$ . Dot colors: black means  $\mu = 0.75$ , red means  $\mu = 1$ , green means  $\lambda = 1.25$  and blue means  $\mu = 1.5$ .

For each of the four values of  $\mu$  the left panel in Fig. 39.5 shows the optimal operating point when wAFROC-AUC is maximized. It shows the FROC curves with optimal operating points superimposed. These illustrate the previous comments, namely, as  $\mu$  increases, *the optimal operating point moves up the FROC curve*.

The right panel in Fig. 39.5 shows the optimal operating point when the Youden-index is maximized.

To visualize true performance one compares wAFROC curves.

Each curve ends at the optimal threshold listed in Table 39.3, namely  $\zeta_1 = -0.268$  for the red curve, and  $\zeta_1 = 0.563$  for the blue curve. The lower performance represented by the blue curve, based on Youden-index maximization, is due to the adoption of an overly strict threshold.

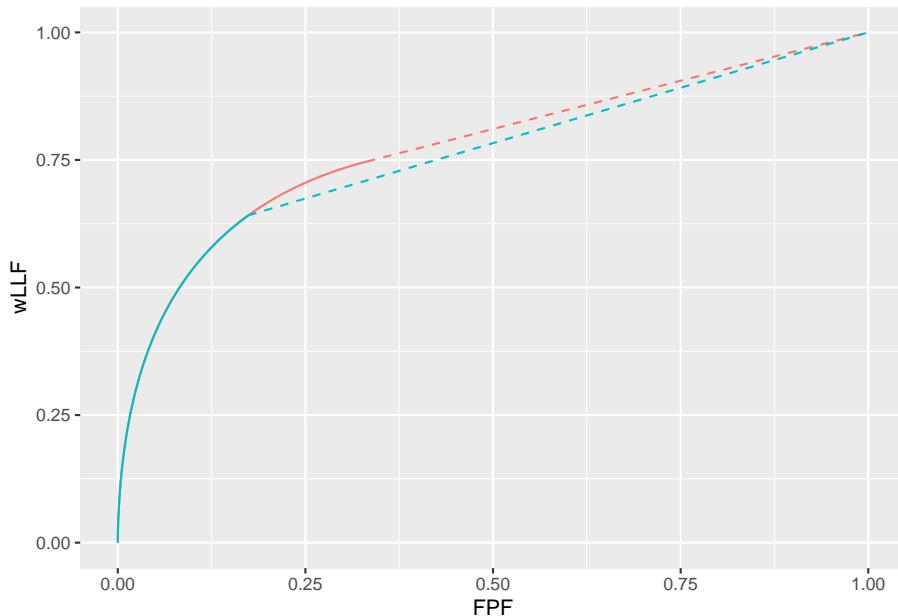


Figure 39.6: wAFROC curves for wAFROC-AUC and Youden-index based optimizations: both curves correspond to  $\lambda = 1$ ,  $\nu = 1$  and  $\mu = 1.5$ . The optimal reporting threshold  $\zeta_1$  is determined by the selected FOM. The red curve corresponds to FOM = wAFROC-AUC and the blue curve corresponds to FOM = Youden-index. The stricter reporting threshold found by the Youden-index based method sacrifices a considerable amount of area under the wAFROC. The two wAFROC-AUCs are 0.777 and 0.760, respectively.

## 39.4 Using the method

Assume that one has designed an algorithmic observer that has been optimized with respect to all other parameters except the reporting threshold. At this point the algorithm reports every suspicious region, no matter how low the malignancy index. The mark-rating pairs are entered into a `RJafroc` format Excel input file. The next step is to read the data file – `DfReadDataFile()` – convert it to an ROC dataset – `DfFroc2Roc()` – and then perform a radiological search model (RSM) fit to the dataset using function `FitRsmRoc()`. This yields the necessary  $\lambda, \mu, \nu$  parameters. These values are used to perform the computations described in the embedded code in this chapter, see for example Section 39.3.2. This determines the optimal reporting threshold. The RSM parameter values and the reporting threshold determine the optimal reporting point on the FROC curve. The designer sets the algorithm to only report marks with confidence levels exceeding this threshold.

## 39.5 An application

The standalone CAD LROC dataset described in (Hupse et al., 2013) was used to create the quasi-FROC ROC-AUC equivalent dataset embedded in `RJafroc` as object `datasetCadSimuFroc`. In the following code the first reader for this dataset, corresponding to CAD, is extracted using `DfExtractDataset` (the other readers, corresponding to radiologists who interpreted the same cases, are not used here). The function `DfFroc2Roc` converts this to an ROC dataset. The function `DfBinDataset` bins the data to about 7 bins. One lesion per abnormal case is assumed: `lesDistr = c(1)`. `FitRsmRoc` fits the binned ROC dataset to the radiological search model RSM. Object `fit` contains all necessary parameters required to perform the optimizations described in previous sections.

```
ds <- datasetCadSimuFroc
dsCad <- DfExtractDataset(ds, rdrs = 1)
dsCadRoc <- DfFroc2Roc(dsCad)
dsCadRocBinned <- DfBinDataset(dsCadRoc, opChType = "ROC")
lesDistr <- c(1)
fit <- FitRsmRoc(dsCadRocBinned, lesDistr)
```

Table 39.4 summarizes the results.

Table 39.4: Summary of optimization results for example FROC dataset. The wAFROC column always displays wAFROC-AUC, even though the optimized quantity may be the Youden-index, as in the last four rows.

FOM	lambda	zeta1	wAFROC	ROC	OptOpPt
wAFROC			1.739	0.774	0.815 (0.278, 0.679)
Youden	18.680		1.982	0.770	0.798 (0.161, 0.627)

The dataset is characterized by a large  $\lambda$  parameter and, consistent with the finding in 39.3.2, the advantage of wAFROC-AUC over Youden-index based optimization, as measured by the difference in corresponding wAFROC-AUCs, is small.

Fig. 39.7 shows FROC curves with superimposed optimal operating points.

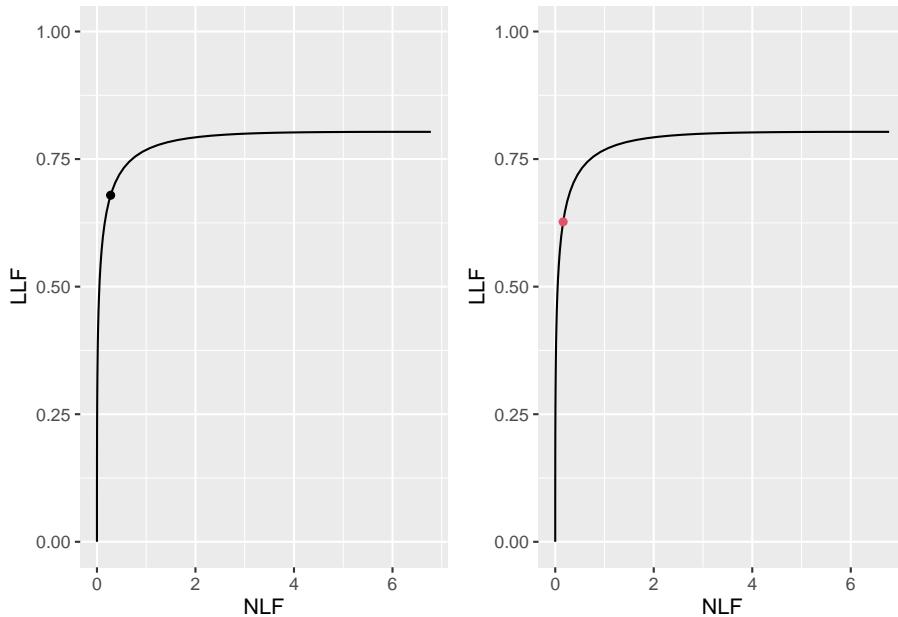


Figure 39.7: Left panel: maximized wAFROC-AUC was used to find optimal  $\zeta_1$ . Right panel: maximized Youden-index was used.

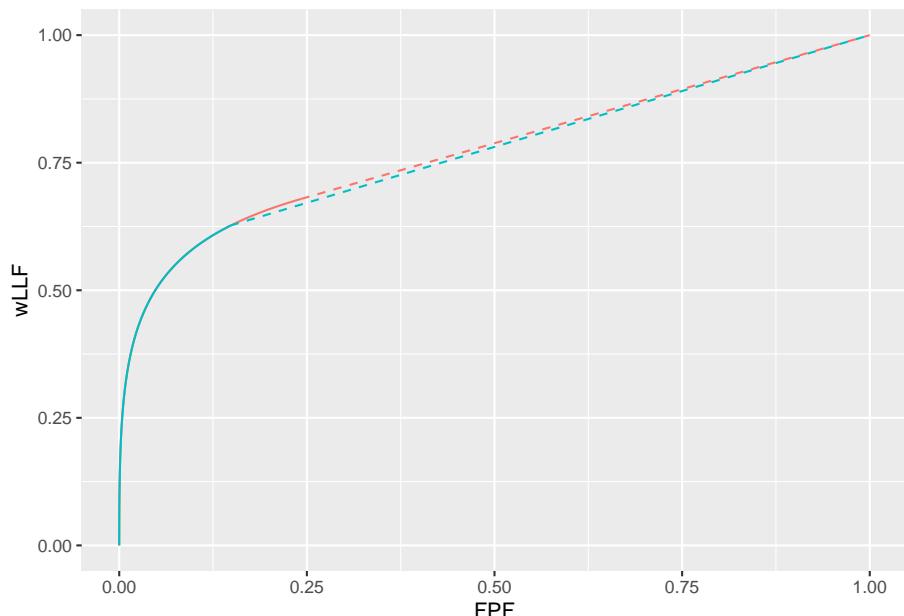


Figure 39.8: Red line and dots: wAFROC-AUC based optimization; blue line and dots: Youden-index based optimization. The two wAFROC-AUCs are 0.774 and 0.770, respectively.

## 39.6 Discussion

Described is a method for finding the optimal operating point on an FROC curve. The method consists of varying the reporting threshold to maximize the area under the wAFROC. An alternate method, based on maximization of the Youden-index, was also tested. Both methods are illustrated using the radiological search model to parameterize the FROC data. In all cases studied the Youden-index based method selected a stricter reporting threshold than optimal, resulting in lower wAFROC-AUC and ROC-AUC as compared to wAFROC-AUC based optimization. The results are illustrated using FROC curves, which are more familiar to CAD designers.

The method was applied to a quasi-FROC dataset created from an originally LROC dataset. For this dataset the optimized wAFROC-AUC was marginally superior to that using the Youden-index.

With increasing  $\lambda$  every case is guaranteed at least one z-sample, and the model becomes more ROC-like.

## 39.7 References



# Chapter 40

## Localization - classification tasks

### 40.1 TBA How much finished

10%

### 40.2 Introduction

TBA: This project is a works-in-progress.

### 40.3 Abbreviations

- Correct-localization correct-classification = **CL-CC**
- Correct-localization incorrect-classification = **CL-IC**
- Incorrect-localization classification not applicable = **IL-NA**

### 40.4 History and basic idea

This project started with a request to extend localization analysis software `RJafroc` to localization-classification tasks. Since this is new research the required data format is not in the `RJafroc` documentation. Some familiarity with basic localization task analysis is assumed.

The basic idea is that spatial localization is a special case of localization-with-classification. **CL-CC** marks are put in TP sheet and other are put in FP sheet.

## 40.5 First example, File1.xlsx

- This example is implemented in file **File1.xlsx**.
- There are four classes of lesions: C1, C2, C3and C4.
- The rating scale is 1 - 10 and positive-directed.
- The dataset has 3 cases: 9, 17 and 19.

### 40.5.1 Truth sheet

This has the ground truth of for cases and lesions, and specifies their class types.

	A	B	C	D	E	F	G	H
1	CaseID	LesionID	Weight	ReaderID	ModalityID	Paradigm	Class	
2	9	1	0	1	1	FROC	C1	
3	9	2	0	1	1	FCTRL	C4	
4	17	1	0	1	1		C1	
5	17	2	0	1	1		C2	
6	17	3	0	1	1		C3	
7	17	4	0	1	1		C4	
8	19	1	0	1	1		C2	
9								
10								
11								
12								
13								
14								
15								
16								
17								
	TP	FP	TRUTH	+				

Figure 40.1: Truth worksheet for File1.xlsx

- Case 9 has two lesions, with classes C1 and C4.
- Case 17 has four lesions, with classes C1, C2, C3and C4.
- Case 19 has one lesion, with class C2.

### 40.5.2 TP sheet

This holds CL-CC marks.

#### 40.5.2.1 Case 9

- Lesion C1, **lesionID = 1**, **CL-CC** mark rated 5.

	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CasedID	LesionID	LL_Rating	Designation	Class		
2	1	1	9	1	5	CL-CC	C1		
3	1	1	17	1	6.1	CL-CC	C1		
4	1	1	17	2	7.1	CL-CC	C2		
5	1	1	17	4	2.3	CL-CC	C4		
6	1	1	19	1	5.7	CL-CC	C2		
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									

Figure 40.2: TP worksheet for File1.xlsx

#### 40.5.2.2 Case 17

- Lesion C1, lesionID = 1, CL-CC mark rated 6.1.
- Lesion C2, lesionID = 2, CL-CC mark rated 7.1.
- Lesion C4, lesionID = 4, CL-CC mark rated 2.3.

#### 40.5.2.3 Case 19

- Lesion C2, lesionID = 1, CL-CC mark rated 5.7.

#### 40.5.3 FP sheet

- This holds IL-NA and CL-IC marks.
- ClassTrue is the true class of the lesion.
- ClassDx is the indicated or diagnosed class of the lesion.

	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CasedID	NL_Rating	Designation	ClassTrue	ClassDx		
2	1	1	9	5.5	CL-IC	C4	C3	this misclassification	
3	1	1	9	1.2	IL-NA	NA	NA		
4	1	1	17	7	CL-IC	C3	C2		
5	1	1	17	2.3	IL-NA	NA	NA		
6	1	1	17	2.1	IL-NA	NA	NA		
7	1	1	19	1.4	IL-NA	NA	NA		
8	1	1	19	6.1	CL-IC	C2	C3		
9									
10									
11									
12									
13									
14									
15									
16									

Figure 40.3: FP worksheet for File1.xlsx

#### 40.5.3.1 Case 9

- CL-IC mark rated 5.5, C2 classified as C3.

- **IL-NA** mark rated 1.2.

#### 40.5.3.2 Case 17

- **CL-IC** mark rated 7, C3 classified as C2.
- **IL-NA** mark rated 2.3.
- **IL-NA** mark rated 2.1.

#### 40.5.3.3 Case 19

- **IL-NA** mark rated 1.4.
- **CL-IC** mark rated 6.1, C2 classified as C3.

#### 40.5.4 The two ratings arrays

```
fileName <- "R/CH83-ClassificationTask/File1.xlsx"
x <- DfReadDataFile(fileName = fileName)
x$ratings$NL[1,1,,]
#>      [,1] [,2] [,3]
#> [1,]  5.5  1.2 -Inf
#> [2,]  7.0  2.3  2.1
#> [3,]  1.4  6.1 -Inf
x$ratings$LL[1,1,,]
#>      [,1] [,2] [,3] [,4]
#> [1,]  5.0 -Inf -Inf -Inf
#> [2,]  6.1  7.1 -Inf  2.3
#> [3,]  5.7 -Inf -Inf -Inf
```

The FOM is shown next:

```
print(UtilFigureOfMerit(x, FOM = "wAFROC1"))
#>          rdr1
#> trt1 0.2361111
```

#### 40.6 Second example, File2.xlsx

I increased the LL rating for case 19 to 10; this should increase the FOM. This example is implemented in file **File2.xlsx**.

	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	LesionID	LL_Rating	Designation	Class		
2	1	1	9	1	5	CL-CC	C1		
3	1	1	17	1	6.1	CL-CC	C1		
4	1	1	17	2	7.1	CL-CC	C2		
5	1	1	17	4	2.3	CL-CC	C4		
6	1	1	19	1	10	CL-CC	C2		
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
		TP	FP	TRUTH	+				
		Ready							

Figure 40.4: TP worksheet for File2.xlsx

```
fileName <- "R/CH83-ClassificationTask/File2.xlsx"
x <- DfReadDataFile(fileName = fileName)
print(UtilFigureOfMerit(x, FOM = "wAFROC1"))
#>          rdr1
#> trt1 0.4583333
```

## 40.7 Third example, File3.xlsx

Starting with original file, I transferred a **CL-IC** for case 17 to the TP sheet, where it is a **CL\_CC** mark. This should increase the FOM as credit is given for **CL-CC**. This example is implemented in file **File3.xlsx**.

```
fileName <- "R/CH83-ClassificationTask/File3.xlsx"
x <- DfReadDataFile(fileName = fileName)
print(UtilFigureOfMerit(x, FOM = "wAFROC1"))
#>          rdr1
#> trt1 0.5277778
```

## 40.8 Fourth example, File4.xlsx

So far we have dealt with one modality and one reader.

- Additional algorithmic readers can be added under **readerID**.
- They should not be added as additional treatments (has to do with treatment being regarded as a fixed factor and reader as a random factor in the analysis).
- The starting point is **File3.xlsx**. I duplicated the data from this for two additional readers to create a single-modality three-reader dataset **File4.xlsx**.

The figure displays three Excel worksheets:

- TP Worksheet:** Shows data for ReaderID 1. Columns include ReaderID, ModalityID, CaseID, LesionID, NL\_Rating, Designation, ClassTrue, and ClassDx. A row is highlighted with a green border and labeled "this misclassification".
- FP Worksheet:** Shows data for ReaderID 1. Columns include ReaderID, ModalityID, CaseID, LesionID, NL\_Rating, Designation, ClassTrue, and ClassDx.
- TRUTH Worksheet:** Shows the ground truth for ReaderID 1. Columns include ReaderID, ModalityID, CaseID, LesionID, NL\_Rating, Designation, ClassTrue, and ClassDx.

Figure 40.5: TP and FP worksheets for File3.xlsx

- Shown next are the three worksheets.

The TRUTH worksheet for File4.xlsx contains the following data:

	A	B	C	D	E	F	G	H
1	CaseID	LesionID	Weight	ReaderID	ModalityID	Paradigm	Class	
2	9	1	0	1, 2, 3	1	FROC	C1	
3	9	2	0	1, 2, 3	1	FCTRL	C4	
4	17	1	0	1, 2, 3	1		C1	
5	17	2	0	1, 2, 3	1		C2	
6	17	3	0	1, 2, 3	1		C3	
7	17	4	0	1, 2, 3	1		C4	
8	19	1	0	1, 2, 3	1		C2	

Figure 40.6: Truth worksheet for File4.xlsx

- Shown next are the three FOMs. Note that they are identical.

```
fileName <- "R/CH83-ClassificationTask/File4.xlsx"
x <- DfReadDataFile(fileName = fileName)
print(UtilFigureOfMerit(x, FOM = "wAFROC1"))
#>      rdr1      rdr2      rdr3
#> trt1 0.52777778 0.52777778 0.52777778
```

	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	LesionID	LL_Rating	Designation	Class		
2	1	1	9	1	5	CL-CC	C1		
3	1	1	17	1	6.1	CL-CC	C1		
4	1	1	17	2	7.1	CL-CC	C2		
5	1	1	17	3	7	CL-CC	C3		
6	1	1	17	4	2.3	CL-CC	C4		
7	1	1	19	1	5.7	CL-CC	C2		
8	2	1	9	1	5	CL-CC	C1		
9	2	1	17	1	6.1	CL-CC	C1		
10	2	1	17	2	7.1	CL-CC	C2		
11	2	1	17	3	7	CL-CC	C3		
12	2	1	17	4	2.3	CL-CC	C4		
13	2	1	19	1	5.7	CL-CC	C2		
14	3	1	9	1	5	CL-CC	C1		
15	3	1	17	1	6.1	CL-CC	C1		
16	3	1	17	2	7.1	CL-CC	C2		
17	3	1	17	3	7	CL-CC	C3		
18	3	1	17	4	2.3	CL-CC	C4		

Figure 40.7: TP worksheet for File4.xlsx

	A	B	C	D	E	F	G	H	I
1	ReaderID	ModalityID	CaseID	NL_Rating	Designation	ClassTrue	ClassDx	this is classification	
2	1	1	9	5.5	CL-IC	C4	C3		
3	1	1	9	1.2	IL-NA	NA	NA		
4	1	1	17	2.3	IL-NA	NA	NA		
5	1	1	17	2.1	IL-NA	NA	NA		
6	1	1	19	1.4	IL-NA	NA	NA		
7	1	1	19	6.1	CL-IC	C2	C3		
8	2	1	9	5.5	CL-IC	C4	C3		
9	2	1	9	1.2	IL-NA	NA	NA		
10	2	1	17	2.3	IL-NA	NA	NA		
11	2	1	17	2.1	IL-NA	NA	NA		
12	2	1	19	1.4	IL-NA	NA	NA		
13	2	1	19	6.1	CL-IC	C2	C3		
14	3	1	9	5.5	CL-IC	C4	C3		
15	3	1	9	1.2	IL-NA	NA	NA		
16	3	1	17	2.3	IL-NA	NA	NA		

Figure 40.8: FP worksheet for File4.xlsx

## 40.9 Fifth example, File5.xlsx

- Need to add some randomness to the ratings.
- I randomly added to the ratings from a uniform distribution in the range -0.5 to +0.5.
- This is very crude, as in practice the the number of marks will also vary from reader to reader.
- But file File5.xlsx should give one the general idea of how to extend to several algorithmic readers.
- Note that now the FOMs are not identical.

```
fileName <- "R/CH83-ClassificationTask/File5.xlsx"
x <- DfReadDataFile(fileName = fileName)
print(UtilFigureOfMerit(x, FOM = "wAFROC1"))
#>      rdr1      rdr2      rdr3
#> trt1 0.3611111 0.5555556 0.4444444
```

## 40.10 Precautions

- Unlike regular RJafroc analysis, there is no error checking of the classification codes C1, etc. For example, if a lesion with class C1 is recorded in the TP sheet as a **CL-CC** and it is also mistakenly recorded in the FP sheet as a **CL-IC**, the program does not know about the mistake. Multiple FP on the same case are allowed in FROC analysis.
- I suggest that the extra columns in the sample files be recorded for your dataset. This will enable me to subsequently include error-checking code for data entry mistakes.
- For example, the columns `Designation`, `ClassTrue` and `ClassRx` in the FP sheet are currently not read by the software.
- To make further progress you need to drastically reduce the file size (once the new method is fully developed you can always add the remaining cases and readers). The current file size makes it impossible to fully develop the system. Most studies in this field are done with 2-3 modalities and about 100-200 cases.

## 40.11 Discussion

## 40.12 References

# Chapter 41

## Split Plot Study Design

### 41.1 TBA How much finished

10%

### 41.2 Mean Square R(T)

R(T) is read as “reader nested within treatment” (Hillis, 2014).

$$\text{MS}[R(T)] = \frac{1}{I(J-1)} \sum_{i=1}^I \sum_{j=1}^J (\theta_{ij} - \theta_{i\bullet})^2 \quad (41.1)$$

$$\text{MS}[R(T)] = \frac{1}{I} \sum_{i=1}^I \frac{1}{J_i - 1} \sum_{j=1}^{J_i} (\theta_{ij} - \theta_{i\bullet})^2 \quad (41.2)$$

### 41.3 References



# Bibliography

- Alberdi, E., Povyakalo, A. A., Strigini, L., Ayton, P., and Given-Wilson, R. (2008). Cad in mammography: lesion-level versus case-level analysis of the effects of prompts on human decisions. *International Journal of Computer Assisted Radiology and Surgery*, 3(1-2):115–122.
- Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415.
- Barlow, W. E., Chi, C., Carney, P. A., Taplin, S. H., D'Orsi, C., Cutter, G., Hendrick, R. E., and Elmore, J. G. (2004). Accuracy of screening mammography interpretation by characteristics of radiologists. *Journal of the National Cancer Institute*, 96(24):1840–1850.
- Barnes, G., Sabbagh, E., Chakraborty, D., Nath, P., Luna, R., Sanders, C., and Fraser, R. (1989). A comparison of dual-energy digital radiography and screen-film imaging in the detection of subtle interstitial pulmonary disease. *Investigative Radiology*, 24(8):585–591.
- Beam, C. A., Layde, P. M., and Sullivan, D. C. (1996). Variability in the interpretation of screening mammograms by us radiologists. findings from a national sample. *Archives of Internal Medicine*, 156(2):209–13.
- Berbaum, K. S., Dorfman, D. D., Franken, E. A., and Caldwell, R. T. (2002). An empirical comparison of discrete ratings and subjective probability ratings. *Academic Radiology*, 9(7):756–763.
- Black, W. C. (2000). Anatomic extent of disease: A critical variable in reports of diagnostic accuracy. *Radiology*, 217(2):319–320.
- Black, W. C. and Dwyer, A. J. (1990). Local versus global measures of accuracy: An important distinction for diagnostic imaging. *Med Decis Making*, 10(4):266–273.
- Bochud, F., Abbey, C., and Eckstein, M. (1999). Visual signal detection in structured backgrounds iv, calculation of figures of merit for model observers in non-stationary backgrounds. *Journal of the Optical Society of America, A, Optics, Image Science, and Vision*, 17(2):206–17.

- Bolker, B. and R Development Core Team (2020). *bbmle: Tools for General Maximum Likelihood Estimation*. R package version 1.0.23.1.
- Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90.
- Bunch, P. C., Hamilton, J., Sanderson, G., and Simmons, A. (1977a). A free-response approach to the measurement and characterization of radiographic-observer performance. *Proc. SPIE*, 127:124–135.
- Bunch, P. C., Hamilton, J. F., Sanderson, G. K., and Simmons, A. H. (1977b). 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.
- Burgess, A. E. (2011). Visual perception studies and observer models in medical imaging. In *Seminars in nuclear medicine*, volume 41, pages 419–436. Elsevier.
- Chakraborty, D. (1997a). Comparison of computer analysis of mammography phantom images (campi) with perceived image quality of phantom targets in the acr phantom. In Kundel, H. L., editor, *Proc. SPIE Medical Imaging 1997: Image Perception*, volume 3036, pages 160–167. SPIE.
- Chakraborty, D. (1997b). Computer analysis of mammography phantom images (campi): An application to the measurement of microcalcification image quality of directly acquired digital images. *Medical Physics*, 24(8):1269–1277.
- Chakraborty, D., Breathnach, E., Yester, M., Soto, B., Barnes, G., and Fraser, R. (1986). Digital and conventional chest imaging: A modified ROC study of observer performance using simulated nodules. *Radiology*, 158:35–39.
- Chakraborty, D. and Fatouros, P. P. (1998). Application of computer analysis of mammography phantom images (campi) methodology to the comparison of two digital biopsy machines. In James T. Dobbins III, J. M. B., editor, *Proc SPIE Medical Imaging 1998: Physics of Medical Imaging*, volume 3336, pages 618–628. SPIE.
- Chakraborty, D., Phillips, P., and Zhai, X. (2020a). *RJafroc: Analyzing Diagnostic Observer Performance Studies*. R package version 1.3.2.9000.
- Chakraborty, D., Phillips, P., and Zhai, X. (2020b). *RJafroc: Artificial Intelligence Systems and Observer Performance*. R package version 2.0.1.9000.
- Chakraborty, D. P. (1989). Maximum likelihood analysis of free-response receiver operating characteristic (froc) data. *Medical physics*, 16(4):561–568.
- 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. (2006a). An alternate method for using a visual discrimination model (vdm) to optimize softcopy display image quality. *Journal of the Society for Information Display*, 14(10):921–926.
- Chakraborty, D. P. (2006b). A search model and figure of merit for observer data acquired according to the free-response paradigm. *Phys. Med. Biol.*, 51:3449–3462.
- Chakraborty, D. P. (2008). Validation and statistical power comparison of methods for analyzing free-response observer performance studies. *Acad Radiol*, 15(12):1554–1566.
- 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.
- Chakraborty, D. P. and Berbaum, K. S. (2004). Observer studies involving detection and localization: Modeling, analysis and validation. *Med Phys*, 31(8):2313–2330.
- Chakraborty, D. P., Sivarudrappa, M., and Roehrig, H. (1999). Computerized measurement of mammographic display image quality. In John M. Boone; James T. Dobbins III, J. M. B., editor, *Proc SPIE Medical Imaging 1999: Physics of Medical Imaging*, volume 3659, pages 131–141. SPIE.
- Chakraborty, D. P. and Svahn, T. (2011). Estimating the parameters of a model of visual search from ROC data: an alternate method for fitting proper ROC curves. *Proc. SPIE* 7966, 7966.
- Chakraborty, D. P. and Yoon, H. J. (2008). Operating characteristics predicted by models for diagnostic tasks involving lesion localization. *Medical Physics*, 35(2):435–445.
- Chakraborty, D. P. and Yoon, H. J. (2009). JAFROC analysis revisited: figure-of-merit considerations for human observer studies. *Proc. SPIE Medical Imaging: Image Perception, Observer Performance, and Technology Assessment*, 7263:72630T.
- Chakraborty, D. P. and Zhai, X. (2016). On the meaning of the weighted alternative free-response operating characteristic figure of merit. *Medical physics*, 43(5):2548–2557.
- 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.
- Daly, S. (1993). *The visible differences predictor: an algorithm for the assessment of image fidelity*, pages 179–206. MIT Press, Cambridge, Mass.
- De Boo, D. W., Uffmann, M., Weber, M., Bipat, S., Boorsma, E. F., Scheerder, M. J., Freling, N. J., and Schaefer-Prokop, C. M. (2011). Computer-aided detection of small pulmonary nodules in chest radiographs: an observer study. *Academic radiology*, 18(12):1507–1514.
- 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.
- DeSantis, C., Siegel, R., Bandi, P., and Jemal, A. (2011). Breast cancer statistics, 2011. *CA: a cancer journal for clinicians*, 61(6):408–418.
- Dobbins III, J. T., McAdams, H. P., Sabol, J. M., Chakraborty, D. P., Kazerrooni, E. A., Reddy, G. P., Vikgren, J., and Båth, M. (2016). Multi-institutional evaluation of digital tomosynthesis, dual-energy radiography, and conventional chest radiography for the detection and management of pulmonary nodules. *Radiology*, 282(1):236–250.
- Dorfman, D. and Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals - rating-method data. *Journal of Mathematical Psychology*, 6:487–496.
- Dorfman, D. and Berbaum, K. (2000). A contaminated binormal model for ROC data: Part ii. a formal model. *Acad Radiol.*, 7(6):427–37.
- Dorfman, D., Berbaum, K., and Metz, C. (1992a). ROC characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. *Invest. Radiol.*, 27(9):723–731.
- Dorfman, D., Berbaum, K., Metz, C., Lenth, R., Hanley, J., and Abu Dagga, H. (1997). Proper receiving operating characteristic analysis: The bigamma model. *Acad. Radiol.*, 4(2):138–149.
- 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.
- Dorfman, D. D., Berbaum, K. S., and Metz, C. E. (1992b). Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. *Investigative radiology*, 27(9):723–731.
- Duchowski, A. T. (2002). *Eye Tracking Methodology: Theory and Practice*. Clemson University, Clemson, SC.

- Edwards, D. C., Kupinski, M. A., Metz, C. E., and Nishikawa, R. M. (2002). Maximum likelihood fitting of froc curves under an initial-detection-and-candidate-analysis model. *Medical physics*, 29(12):2861–2870.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, Boca Raton.
- Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.
- Egan, J., Greenburg, G., and Schulman, A. (1961). Operating characteristics, signal detectability and the method of free response. *J Acoust Soc Am.*, 33:993–1007.
- Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. Academic Press Series in Cognition and Perception. Academic Press, Inc., New York, first edition.
- Ernster, V. L. (1981). The epidemiology of benign breast disease. *Epidemiologic reviews*, 3(1):184–202.
- Fenton, J. (2015). Is it time to stop paying for computer-aided mammography? *JAMA Intern Med*.
- Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D’Orsi, C., Berns, E. A., Cutter, G., Hendrick, R. E., Barlow, W. E., and Elmore, J. G. (2007). Influence of computer-aided detection on performance of screening mammography. *N Engl J Med*, 356(14):1399–1409.
- Fisher, R. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest and smallest member of a sample. *Proc. Cambridge Phil. Society*, 24:180–190.
- Fletcher, R. (1970). A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322.
- Fletcher, R. (2013). *Practical methods of optimization*. John Wiley and Sons.
- Franken, Edmund A., J., Berbaum, K. S., Marley, S. M., Smith, W. L., Sato, Y., Kao, S. C. S., and Milam, S. G. (1992). Evaluation of a digital workstation for interpreting neonatal examinations: A receiver operating characteristic study. *Investigative Radiology*, 27(9):732–737.
- 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.

- Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109):23–26.
- Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley and Sons, New York.
- Gur, D., Bandos, A. I., Cohen, C. S., Hakim, C. M., Hardesty, L. A., Ganott, M. A., Perrin, R. L., Poller, W. R., Shah, R., Sumkin, J. H., Wallace, L. P., and Rockette, H. E. (2008). The "laboratory" effect: Comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology*, 249(1):47–53.
- 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.
- Halpern, S. D., Karlawish, J. H., and Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Jama*, 288(3):358–362.
- Hanley, J. A. (1988). The robustness of the "binormal" assumptions used in fitting ROC curves. *Med. Decis. Making*, 8(3):197–203.
- Hanley, J. A. and Hajian-Tilaki, K. O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update. *Academic Radiology*, 4(1):49–58.
- Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.
- Hartmann, L. C., Sellers, T. A., Frost, M. H., Lingle, W. L., Degnim, A. C., Ghosh, K., Vierkant, R. A., Maloney, S. D., Pankratz, V. S., Hillman, D. W., et al. (2005). Benign breast disease and the risk of breast cancer. *New England Journal of Medicine*, 353(3):229–237.
- Hein, P. A., Krug, L. D., Romano, V. C., Kandel, S., Hamm, B., and Rogalla, P. (2010). Computer-aided detection in computed tomography colonography with full fecal tagging: comparison of standalone performance of 3 automated polyp detection systems. *Canadian Association of Radiologists Journal*, 61(2):102–108.
- Hillis, S., Obuchowski, N., Schartz, K., and Berbaum, K. (2005a). 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. (2007a). A comparison of denominator degrees of freedom methods for multiple observer (ROC) analysis. *Statistics in medicine*, 26(3):596–619.
- Hillis, S. L. (2007b). 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 multi-reader multicase radiological imaging data. *Statistics in Medicine*, 33(2):330–360.
- Hillis, S. L., Berbaum, K., and Metz, C. (2008a). 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., Berbaum, K. S., and Metz, C. E. (2008b). Recent developments in the dorfman-berbaum-metz procedure for multireader roc study analysis. *Academic radiology*, 15(5):647–661.
- 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.
- Hillis, S. L., Obuchowski, N. A., Schartz, K. M., and Berbaum, K. S. (2005b). A comparison of the dorfman–berbaum–metz and obuchowski–rockette methods for receiver operating characteristic (ROC) data. *Statistics in medicine*, 24(10):1579–1607.
- Hupse, R., Samulski, M., Lobbes, M., Heeten, A., Imhof-Tas, M., Beijerinck, D., Pijnappel, R., Boetes, C., and Karssemeijer, N. (2013). Standalone computer-aided detection compared to radiologists' performance for the detection of mammographic masses. *European Radiology*, 23(1):93–100.
- 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.
- Jiang, Y. and Metz, C. E. (2010). BI-RADS data should not be used to estimate ROC curves. *Radiology*, 256(1):29–31.
- Kooi, T., Gubern-Merida, A., Mordang, J.-J., Mann, R., Pijnappel, R., Schuur, K., den Heeten, A., and Karssemeijer, N. (2016). A comparison between a deep convolutional neural network and radiologists for classifying regions of interest in mammography. In *International Workshop on Breast Imaging*, pages 51–56. Springer.
- Kundel, H., Berbaum, K., Dorfman, D., Gur, D., Metz, C. E., and Swensson, R. G. (2008). Receiver operating characteristic analysis in medical imaging (icru report 79). Report, International Commission on Radiation Units and Measurements.

- Kundel, H. L., Nodine, C. F., Conant, E. F., and Weinstein, S. P. (2007). Holistic component of image perception in mammogram interpretation: Gaze-tracking study. *Radiology*, 242(2):396–402.
- 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.
- Lubin, J. (1995). *A visual discrimination model for imaging system design and evaluation*. Visual Models for Target Detection and Recognition. World Scientific Publishers, Singapore.
- Lusted, L. B. (1971). Signal detectability and medical decision making. *Science*, 171:1217–1219.
- Macmillan, N. and Creelman, C. (1991). *Detection Theory: A User's Guide*. Cambridge University Press, New York.
- Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.
- Metz, C. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.
- Metz, C. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigative Radiology*, 24:234–245.
- Metz, C. and Pan, X. (1999). Proper binormal ROC curves: Theory and maximum-likelihood estimation. *J Math Psychol*, 43(1):1–33.
- Metz, C. E. (1986). ROC methodology in radiologic imaging. *Investigative Radiology*, 21(9):720–733.
- Metz, C. E., Starr, S. J., and Lusted, L. B. (1976). Observer performance in detecting multiple radiographic signals: prediction and analysis using a generalized roc approach. *Radiology*, 121(2):337–347.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97.
- Miller, H. (1969). The FROC curve: a representation of the observer's performance for the method of free response. *The Journal of the Acoustical Society of America*, 46(6(2)):1473–1476.

- 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.
- Nishikawa, R. (2012). Estimating sensitivity and specificity in an ROC experiment. *Breast Imaging*, pages 690–696.
- Nishikawa, R. M. and Pesce, L. L. (2011). Fundamental limitations in developing computer-aided detection for mammography. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 648:S251–S254.
- Noether, G. E. (1967). Elements of nonparametric statistics. Report, Wiley and Sons.
- Obuchowski, N. A. (1998). Sample size calculations in studies of test accuracy. *Statistical Methods in Medical Research*, 7(4):371–392.
- Obuchowski, N. A. (2000). Sample size tables for receiver operating characteristic studies. *Am. J. Roentgenol.*, 175(3):603–608.
- Obuchowski, N. A., Lieber, M. L., and Powell, K. A. (2000). Data analysis for detection and localization of multiple abnormalities with application to mammography. *Acad. Radiol.*, 7(7):516–525.
- Obuchowski, N. A. and Rockette, H. (1995a). 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.
- Obuchowski, N. A. and Rockette, H. E. (1995b). Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations. *Communications in Statistics-simulation and Computation*, 24(2):285–308.
- Pan, X. and Metz, C. E. (1997). The “proper” binormal model: parametric receiver operating characteristic curve estimation with degenerate data. *Academic radiology*, 4(5):380–389.
- Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.
- Penedo, M., Souto, M., Tahoces, P. G., Carreira, J. M., Villalon, J., Porto, G., Seoane, C., Vidal, J. J., Berbaum, K. S., Chakraborty, D. P., and Fajardo, L. L. (2005). Free-response receiver operating characteristic evaluation of lossy jpeg2000 and object-based set partitioning in hierarchical trees compression of digitized mammograms. *Radiology*, 237(2):450–457.

- Petrick, N. and Pastel, M. (2018). Guidance for industry and fda staff clinical performance assessment: considerations for computer-assisted detection devices applied to radiology images and radiology device data—premarket approval (pma) and premarket notification [510 (k)] submission.
- Philpotts, L. E. (2009). Can computer-aided detection be detrimental to mammographic interpretation? *Radiology*, 253(1):17–22.
- Pisano, E., Gatsonis, C., Hendrick, E., Yaffe, M., Baum, J., Acharyya, S., Conant, E., Fajardo, L., Bassett, L., D'Orsi, C., Jong, R., and Rebner, M. (2005). Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med*, 353(17):1773–1783.
- Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America*, 24(6):745–749.
- Pollack, I. (1953). The information of elementary auditory displays. ii. *The Journal of the Acoustical Society of America*, 25(4):765–769.
- Popescu, L. M. (2011). Nonparametric signal detectability evaluation using an exponential transformation of the FROC curve. *Medical physics*, 38(10):5690–5702.
- Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (2007). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge, 3 edition.
- Rao, V. M., Levin, D. C., Parker, L., Cavanaugh, B., Frangos, A. J., and Sunshine, J. H. (2010). How widely is computer-aided detection used in screening and diagnostic mammography? *Journal of the American College of Radiology*, 7(10):802–805.
- Rockette, H., Gur, D., and Metz, C. (1992). The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. *Investigative Radiology*, 27:169–172.
- 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.
- Ruschin, M., Timberg, P., Bath, M., Hemdal, B., Svahn, T., Saunders, R., Samei, E., Andersson, I., Mattsson, S., Chakraborty, D. P., and Tingberg, A. (2007). Dose dependence of mass and microcalcification detection in digital mammography: free response human observer studies. *Medical Physics*, 34:400 – 407.
- 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.
- Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656.
- Shanno, D. F. and Kettler, P. C. (1970). Optimal conditioning of quasi-newton methods. *Mathematics of Computation*, 24(111):657–664.
- Siddiqui, K. M., Johnson, J. P., Reiner, B. I., and Siegel, E. L. (2005). Discrete cosine transform jpeg compression vs. 2d jpeg2000 compression: Jndmetrix visual discrimination model image quality analysis. In *Medical Imaging 2005: PACS and Imaging Informatics*, volume 5748, pages 202–207. International Society for Optics and Photonics.
- Skaane, P., Bandos, A. I., Gullien, R., Eben, E. B., Ekseth, U., Haakenaasen, U., Izadi, M., Jebsen, I. N., Jahr, G., and Krager, M. (2013). Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology*, 267(1):47–56.
- Soh, B. P., Lee, W., McEntee, M. F., Kench, P. L., Reed, W. M., Heard, R., Chakraborty, D. P., and Brennan, P. C. (2013). Screening mammography: test set data can reasonably describe actual clinical reporting. *Radiology*, 268(1):46–53.
- Starr, S., Metz, C., and Lusted, L. (1977). Comments on generalization of receiver operating characteristic analysis to detection and localization tasks. *Phys. Med. Biol.*, 22:376–379.
- 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.
- Stein, S. K. and Barcellos, A. (1992). *Calculus and analytic geometry*. McGraw-Hill Companies, 5 edition.
- Summers, R. M., Handwerker, L. R., Pickhardt, P. J., Van Uitert, R. L., Deshpande, K. K., Yeshwant, S., Yao, J., and Franaszek, M. (2008). Performance of a previously validated ct colonography computer-aided detection system in a new patient population. *American Journal of Roentgenology*, 191(1):168–174.
- Swensson, R. G. (1996a). Unified measurement of observer performance in detecting and localizing target objects on images. *Medical Physics*, 23(10):1709–1725.
- Swensson, R. G. (1996b). 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.
- Tan, T., Platel, B., Huisman, H., Sánchez, C., Mus, R., and Karssemeijer, N. (2012). Computer-aided lesion diagnosis in automated 3-d breast ultrasound using coronal spiculation. *Medical Imaging, IEEE Transactions on*, 31(5):1034–1042.
- Taylor, S. A., Halligan, S., Burling, D., Roddie, M. E., Honeyfield, L., McQuillan, J., Amin, H., and Dehmeshki, J. (2006). Computer-assisted reader software versus expert reviewers for polyp detection on ct colonography. *American Journal of Roentgenology*, 186(3):696–702.
- Thompson, J. D., Hogg, P., Manning, D. J., Szczepura, K., and Chakraborty, D. P. (2014). A free-response evaluation determining value in the computed tomography attenuation correction image for revealing pulmonary incidental findings: a phantom study. *Academic radiology*, 21(4):538–545.
- Thompson, M. L. and Zucchini, W. (1989). On the statistical analysis of roc curves. *Statistics in medicine*, 8(10):1277–1290.
- 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.
- USAirForce, R. (1947). A statistical theory of target detection by pulsed radar.
- Van den Branden Lambrecht, C. J. and Verschueren, O. (1996). Perceptual quality measure using a spatiotemporal model of the human visual system. In *Digital Video Compression: Algorithms and Technologies 1996*, volume 2668, pages 450–461. International Society for Optics and Photonics.
- 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*.
- Vikgren, J., Zachrisson, S., Svalkvist, A., Johnsson, A. A., Boijsen, M., Flinck, A., Kheddache, S., and Bath, M. (2008). Comparison of chest tomosynthesis and chest radiography for detection of pulmonary nodules: Human observer study of clinical cases. *Radiology*, 249(3):1034–1041.
- Wagner, R. F., Beiden, S. V., and Metz, C. E. (2001). Continuous versus categorical data for ROC analysis: Some quantitative considerations. *Academic Radiology*, 8(4):328–334.

- Warren, L. M., Given-Wilson, R. M., Wallis, M. G., Cooke, J., Halling-Brown, M. D., Mackenzie, A., Chakraborty, D. P., Bosmans, H., Dance, D. R., and Young, K. C. (2014). The effect of image processing on the detection of cancers in digital mammography. *American Journal of Roentgenology*, 203(2):387–393.
- Wilcoxon, F. (1945). Individual comparison by ranking methods. *Biometrics*, 1:80–83.
- Yoon, H. J., Zheng, B., Sahiner, B., and Chakraborty, D. P. (2007). Evaluating computer-aided detection algorithms. *Medical Physics*, 34(6):2024–2038.
- Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.
- Zanca, F., Hillis, S. L., Claus, F., Van Ongeval, C., Celis, V., Provoost, V., Yoon, H.-J., and Bosmans, H. (2012). Correlation of free-response and receiver-operating-characteristic area-under-the-curve estimates: Results from independently conducted FROC/ROC studies in mammography. *Med Phys*, 39(10):5917–5929.
- 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.
- Zhou, X.-H., Obuchowski, N. A., and McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine*. John Wiley and Sons, New York.