The RJafroc Froc Book

Dev P. Chakraborty, PhD

2021-12-20

Contents

Preface

- It is intended as an online update to my print 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 online book is in 3 parts.
- This part is devoted to the FROC paradigm.

TBA How much finished

- HMF approximately 30%
- This book is currently (as of December 2021) in preparation.
- Parts labeled TBA and TODOLAST need to be updated on final revision.
- Un-comment links like `\@ref(froc-paradigm-solar-analogy)` etc. Search for '@ref

The pdf file of the book

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

The html version of the book

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

A note on the online distribution mechanism of the book

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

Structure of the book

The book is divided into parts as follows:

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

Contributing to this book

I appreciate constructive feedback on this document. To do this raise an `Issue` on the `GitHub` interface. Click on the `Issues` tab under `dpc10ster/RJafrocFrocBook`, 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

Chapters needing heavy edits

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

Shelved vs. removed vs. parked folders needing heavy edits

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

Coding aids (for me)

- weird error with knitr not responding to changes in Rmd file: traced to upper case lower case confusion: 13A-froc-empirical1.Rmd which should be 13a-froc-empirical1.Rmd

0.0.1 formatting

- sprintf("%.4f", proper formatting of numbers
- OpPtStr(, do:

0.0.2 tables

- https://github.com/haozhu233/kableExtra/issues/624
- kbl(dfA, caption = "....", booktabs = TRUE, escape = FALSE) %>% collapse_rows(columns = c(1, 3), valign = "middle") %>% kable_styling(latex_options = c("basic", "scale_down", "HOLD_position"), row_label_position = "c")
- "'{r, attr.source = ".numberLines"}
- kbl(x12, caption = "Summary of optimization results using wAFROC-AUC.", booktabs = TRUE, escape = FALSE) %>% collapse_rows(columns = c(1), valign = "middle") %>% kable_styling(latex_options = c("basic", "scale_down", "HOLD_position"), row_label_position = "c")
- $\exp(-\lambda')$ space before dollar sign generates a pdf error
- FP errors generated by GitHub actions due to undefined labels: Error: Error: pandoc version 1.12.3 or higher is required and was not found (see the help page ?rmarkdown::pandoc_available). In addition: Warning message: In verify_rstudio_version() : Please install or upgrade Pandoc to at least version 1.17.2; or if you are using RStudio, you can just install RStudio 1.0+. Execution halted

0.0.3 tinytex problems

- dont update in response to messages; breaks everything
- DONT DO THIS: When tinytex::install_tinytex() hangs up try
- DONT DO THIS: tinytex::install_tinytex(repository = "http://mirrors.tuna.tsinghua.edu.cn/CTAN/", version = "latest")
- Getting very long builds: looping certain commands
- First uninstall tinytex then reinstall:

```
#uninstall_tinytex(force = FALSE, dir = tinytex_root())
#tinytex::install_tinytex()
```

- get very long build first time with looping certain commands
- fixed on subsequent pdf builds

Chapter 1

Empirical plots from FROC data

1.1 TBA How much finished

90%

1.2 TBA Introduction

FROC data consists of mark-rating pairs. In this chapter a distinction is made between latent and actual marks. This is followed by a table summarizing FROC notation. This is a key table which will be used in later chapters. Section `\@ref(froc-paradigm-froc-plot)` introduced the empirical FROC plot. This chapter presents mathematical expressions for this and other empirical plots possible with FROC data: the inferred-ROC, the alternative FROC, the weighted alternative FROC, and others. Operating characteristics are *visual* depicters of performance. Scalar quantities, typically area measures derived from operating characteristics, are *quantitative* measures of performance, termed *figures of merit* (FOMs). This chapter defines an area measure for each empirical operating characteristic. An FROC dataset is used to illustrate the plots and area measures. With the exception of the FROC, all empirical plots include a straight line extension from the observed end-point to (1,1). The correlation between the area measures is qualitatively examined via plots. It is shown that for this dataset the FROC area measure correlates poorly with that under the ROC curve, whereas the other measures correlate better. This is explained by the fact that, unlike the other measures, the FROC plot is not contained within the unit square.

1.3 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 and naive 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, and such studies are very difficult to conduct in a clinical setting.

1.3.1 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.

1.3.2 Binning rule

Recall that ROC data modeling requires the existence of a *case-dependent* decision variable, or z-sample z, and case-independent decision thresholds ζ_r, where $r = 0, 1, ..., R_{ROC} - 1$ and R_{ROC} is the number of ROC study bins [1] and a binning rule that if $\zeta_r \leq z < \zeta_{r+1}$ the case is rated $r + 1$. Dummy cutoffs are

[1] The subscript is used to make explicit the paradigm used.

defined as $\zeta_0 = -\infty$ and $\zeta_{R_{ROC}} = \infty$. The z-sample applies to the whole case. To summarize:

$$\left.\begin{array}{r} \text{if } (\zeta_r \le z < \zeta_{r+1}) \Rightarrow \text{rating} = r+1 \\ r = 0, 1, ..., R_{ROC} - 1 \\ \zeta_0 = -\infty \\ \zeta_{R_{ROC}} = \infty \end{array}\right\} \tag{1.1}$$

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 ζ_r, where $r = 1, ..., R_{FROC}$ and R_{FROC} is the number of FROC study bins, and the rule that a latent mark is marked and rated r if $\zeta_r \le z < \zeta_{r+1}$. 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:

$$\left.\begin{array}{r} \text{if } (\zeta_r \le z < \zeta_{r+1}) \Rightarrow \text{rating} = r \\ r = 1, 2, ..., R_{FROC} \\ \zeta_0 = -\infty \\ \zeta_{R_{FROC}+1} = \infty \end{array}\right\} \tag{1.2}$$

1.4 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. The notation also has to account for cases with no marks.

FROC notation is summarized in Table 1.1, in which *marks refer to 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).

1.4.1 Comments on Table 1.1

- 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.

Table 1.1: FROC notation; all marks refer to latent marks.

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	Location-level truth: 1 for NL and 2 for LL
5	$l_s s$	Mark l_s in location-level truth s
6	$N_{k_t t}$	Number of NLs in case $k_t t$
7	$L_{k_2 2}$	Number of lesions in case $k_2 2$
8	$z_{k_t t l_1 1}$	z-sample for case $k_t t$ and mark $l_1 1$
9	$z_{k_2 2 l_2 2}$	z-sample for case $k_2 2$ and mark $l_2 2$
10	R_{FROC}	Number of FROC bins
11	ζ_1	Lowest reporting threshold
12	ζ_r	r = 2, 3, ... the other non-dummy reporting thresholds
13	$\zeta_0, \zeta_{R_{FROC}+1}$	Dummy thresholds
14	$W_{k_2 l_2}$	Weight of lesion $l_2 2$ in case $k_2 2$
15	L_{max}	Maximum number of lesions per case in dataset
16	L_T	Total number of lesions in dataset

- 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*.
- Row 4: s location-level truth state: 1 for non-diseased and 2 for diseased.
- Row 5: Similar to row 3, two indices $l_s s$ are needed to select latent mark l_s in location-level truth state s. As a useful mnemonic, l is for *location*.
- Row 6: $N_{k_t t}$ is the total number of latent NL marks in case $k_t t$.
- Row 7: $L_{k_2 2}$ is the number of lesions in diseased case $k_2 2$.
- Row 8: The z-sample for case $k_t t$ and 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 9: The z-sample of a latent LL is $z_{k_2 2 l_2 2}$. Unmarked lesions are observable events and are therefore assigned negative infinity ratings (the null-set notation is unnecessary for them).
- Row 10: R_{FROC} is the number of bins in the FROC study.
- Rows 11, 12 and 13: The cutoffs in the FROC study. The lowest threshold is ζ_1. The other non-dummy thresholds are ζ_r where $r = 2, 3, ..., R_{FROC}$. The dummy thresholds are $\zeta_0 = -\infty$ and $\zeta_{R_{FROC}+1} = \infty$.

- 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 in a case sum to unity: $\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} = 1$.
- 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.

1.4.2 A conceptual and notatonal issue

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 ζ_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 [2]. In fact the first model 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 ∞ as does NLF_{max}. However, observed FROC curves end at a finite value of NLF_{max}. This is one reason I introduced the radiological search model (RSM) (Chakraborty, 2006). I will have much more to say about this in Chapter `\@ref(rsm)`, but for now I state one 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 distribution. This means that the actual number of latent NL marks per case can be 0, 1, 2, .., whose average (over all 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., $N_{k_t t} \neq 0$, as in the following:

- $l_1 = \{1, 2, ..., N_{k_t t}\}$ indexes latent NL marks, provided the case has at least one latent NL mark, and otherwise $N_{k_t t} = 0$ and $l_1 = \emptyset$, the null set. The possible values of l_1 are $l_1 = \{\emptyset\} \oplus \left\{1, 2, ...N_{k_t t}\right\}$. The null set applies when the case has no latent NL marks and $\oplus$ is the "exclusive-or"

[2] 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.

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 $1, 2, ...N_{k_t t}$.

- Likewise, $l_2 = \left\{1, 2, ..., L_{k_2 2}\right\}$ indexes latent LL marks. Unmarked LLs are assigned negative infinity ratings. The null set notation is not needed for latent LLs.

1.5 The empirical FROC plot

The FROC, Chapter `\@ref(froc-paradigm-froc-plot)`, is the plot of LLF (along the ordinate) vs. NLF (along the abscissa).

Using the notation of Table 1.1 and assuming binned data[3], then, corresponding to the operating point determined by threshold ζ_r, the FROC abscissa is $\text{NLF}_r \equiv \text{NLF}(\zeta_r)$, the total number of NLs rated $\geq$ threshold ζ_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 ζ_r divided by the total number of lesions:

$$\text{NLF}_r = \frac{n(\text{NLs rated} \geq \zeta_r)}{n(\text{cases})} \tag{1.3}$$

and

$$\text{LLF}_r = \frac{n(\text{LLs rated} \geq \zeta_r)}{n(\text{lesions})} \tag{1.4}$$

The observed operating points correspond to the following values of r:

$$r = 1, 2, ..., R_{FROC} \tag{1.5}$$

Due to the ordering of the thresholds, i.e., $\zeta_1 < \zeta_2 ... < \zeta_{R_{FROC}}$, higher values of r correspond to lower operating points. The uppermost operating point, i.e., that defined by $r = 1$, is referred to the as the *observed end-point*.

Equations (1.3) and (1.4) is are equivalent to:

$$\text{NLF}_r = \frac{1}{K_1 + K_2} \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} \neq 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_r\right) \tag{1.6}$$

[3]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.

and

$$\mathrm{LLF}_r = \frac{1}{L_T} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_r\right) \tag{1.7}$$

Each indicator function, $\mathbb{I}()$, yields unity if the argument is true and zero otherwise.

In Eqn. (1.6) $\mathbb{I}\left(N_{k_t t} \neq 0\right)$ 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 $k_t t$. The term $\mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_r\right)$ 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. (1.7) a summation over t is not needed as only diseased cases contribute to LLF. Analogous to the first indicator function term in Eqn. (1.6), a term like $\mathbb{I}\left(L_{k_2 2} \neq 0\right)$ would be superfluous since $L_{k_2 2} > 0$ as each diseased case must have at least one lesion. The term $\mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_r\right)$ 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.

1.5.1 Definition empirical plot and AUC

The empirical FROC plot connects adjacent operating points $(\mathrm{NLF}_r, \mathrm{LLF}_r)$, 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}.

1.5.2 The origin, a trivial point

Since $\zeta_{R_{FROC}+1} = \infty$ according to Eqn. (1.6) and Eqn. (1.7), $r = R_{FROC} + 1$ yields the trivial operating point (0,0).

1.5.3 The observed end-point and its semi-constrained property

The abscissa of the observed end-point NLF_1, is defined by:

$$\mathrm{NLF}_1 = \frac{1}{K_1 + K_2} \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} \neq 0\right) \sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq \zeta_1\right) \tag{1.8}$$

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:

$$\left.\begin{aligned}\mathrm{LLF}_1 =& \frac{\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq \zeta_1\right)}{L_T} \\ \leq& 1\end{aligned}\right\} \tag{1.9}$$

The numerator is the total number of lesions that were actually marked. The ratio is the fraction of lesions that are marked, which is ≤ 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.

1.5.4 Futility of extrapolation outside the observed end-point

To understand this consider the expression for NLF_0, i.e., using Eqn. (1.6) with $r = 0$:

$$\mathrm{NLF}_0 = \frac{1}{K_1 + K_2}\sum_{t=1}^{2}\sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} \neq 0\right)\sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(z_{k_t t l_1 1} \geq -\infty\right) \tag{1.10}$$

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. (1.8). The 2nd term is:

$$\left.\begin{aligned}\text{2nd term} =& \left(\frac{1}{K_1 + K_2}\right)\sum_{t=1}^{2}\sum_{k_t=1}^{K_t} \mathbb{I}\left(N_{k_t t} \neq 0\right)\sum_{l_1=1}^{N_{k_t t}} \mathbb{I}\left(-\infty \leq z_{k_t t l_1 1} < \zeta_1\right) \\ =& \frac{\text{unknown number}}{K_1 + K_2}\end{aligned}\right\} \tag{1.11}$$

It represents the contribution of unmarked NLs, i.e., latent NLs whose z-samples were below ζ_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:

$$\left.\begin{aligned}\text{LLF}_0 =& \frac{\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(z_{k_2 2 l_2 2} \geq -\infty\right)}{L_T} \\ =& 1\end{aligned}\right\} \tag{1.12}$$

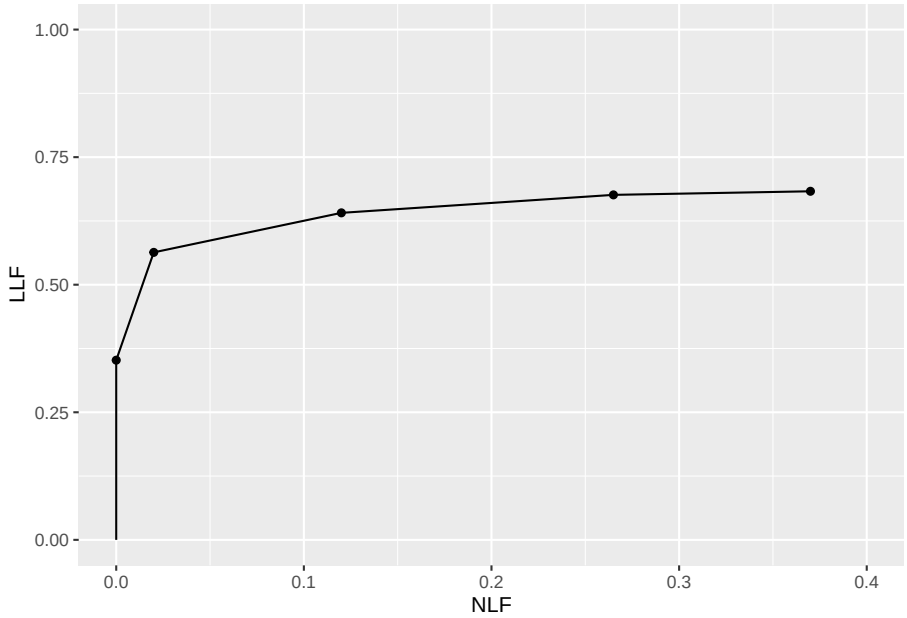
Unlike unmarked latent NLs, **unmarked lesions can safely be assigned the $-\infty$ rating, because an unmarked lesion is an observable event*. The right hand side of Eqn. (1.12) evaluates to unity. However, since the corresponding abscissa NLF_0 is undefined, one cannot plot this point. It follows that 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.

1.5.5 Illustration with a dataset

The following code uses `dataset04` (Zanca et al., 2009) in the `RJafroc` package to illustrate an empirical FROC plot. The dataset has 5-treatments and 4 readers, so in principle one can generate 20 plots. In this example I have selected treatment 1 and reader 1 to produce the plot. The reader should experiment by running `PlotEmpiricalOperatingCharacteristics(dataset04, trts = 1, rdrs = 1, opChType = "FROC")$Plot` with different treatments and readers specified.

```
ret <- PlotEmpiricalOperatingCharacteristics(
  dataset04,
  trts = 1, rdrs = 1, opChType = "FROC")
print(ret$Plot)
```



The study in question was a 5 rating FROC study. The lowest non-trivial point corresponds to the marks rating 5, the next higher one corresponds to marks rated 4 or 5, etc. The FROC plots vary widely but share the common characteristic that the operating points cannot move downward-left as one cumulates lower confidence level marks.

Shown next is calculation of the figure of merit for this dataset. All 20 values are shown. The value for `trt1` and `rdr1` is the area under the FROC plot shown above.

```
UtilFigureOfMerit(dataset04, FOM = "FROC")
#>            rdr1      rdr3      rdr4       rdr5
#> trt1 0.2361972 0.1085035 0.2268486 0.09922535
#> trt2 0.2192077 0.2231338 0.4793310 0.18450704
#> trt3 0.1947359 0.1063028 0.2543662 0.15137324
#> trt4 0.2198768 0.1307394 0.3293662 0.13882042
#> trt5 0.1800528 0.1097535 0.3015141 0.16563380
```

1.6 The inferred-ROC plot

By adopting a rational rule for converting the mark-rating data per case to a single rating per case, and commonly the highest rating rule is used [4], it is

[4] The highest rating method was used in early FROC modeling in (Bunch et al., 1977) and in (Swensson, 1996), the latter in the context of LROC paradigm modeling.

possible to infer ROC data from FROC mark-rating data.

1.6.1 The inferred-ROC rating

The rating of the highest rated mark in 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 ζ divided by total number of non-diseased cases.
- $TPF(\zeta)$ = cumulated inferred TP counts with z-sample $\geq$ threshold ζ 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 end-point to (1,1).*

1.6.2 Inferred FPF

The highest z-sample ROC false positive (FP) rating for non-diseased case $k_1 1$ is defined by:

$$\left.\begin{aligned} FP_{k_1 1} &= \max_{l1} \left(z_{k_1 1 l_1 1} \mid l_1 \neq \emptyset \right) \\ &= -\infty \mid l_1 = \emptyset \end{aligned}\right\} \tag{1.13}$$

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. $FP_{k_1 1}$ is the maximum z-sample over all latent marks occurring on non-diseased case $k_1 1$, or $-\infty$ if the case has no latent marks (this is allowed because a non-diseased case with no marks is an observable event). The corresponding false positive fraction is defined by:

$$\mathrm{FPF}_r \equiv \mathrm{FPF}\left(\zeta_r\right) = \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq \zeta_r \right) \tag{1.14}$$

1.6.3 Inferred TPF

The inferred true positive (TP) z-sample for diseased case $k_2 2$ is defined by:

$$TP_{k_2 2} = \max_{l_1 l_2} \left(z_{k_2 2 l_1 1}, z_{k_2 2 l_2 2} \mid l_1 \neq \emptyset \right) \tag{1.15}$$

or

$$TP_{k_2 2} = \max_{l_2} \left(z_{k_2 2 l_2 2} \mid \left(l_1 = \emptyset \wedge \left(\max_{l_2} \left(z_{k_2 2 l_2 2} \right) \neq -\infty \right) \right) \right) \tag{1.16}$$

or

$$TP_{k_2 2} == -\infty \mid \left(l_1 = \emptyset \wedge \left(\max_{l_2} \left(z_{k_2 2 l_2 2} \right) = -\infty \right) \right) \tag{1.17}$$

Here $\wedge$ is the logical AND operator. An explanation is in order. Consider Eqn. (1.15). There are two z-samples inside the max operator: $z_{k_2 2 l_1 1}, z_{k_2 2 l_2 2}$. The first z-sample is from a NL on a diseased case, as per the $l_1 1$ subscripts, while the second is from a LL on the same diseased case, as per the $l_2 2$ subscripts.

- If $l_1 \neq \emptyset$ then Eqn. (1.15) applies, i.e., one takes the maximum over all z-samples, NLs and LLs, whichever is higher, on the diseased case.
- If $l_1 = \emptyset$ and at least one lesion is marked, then Eqn. (1.16) applies, i.e., one takes the maximum z-sample over all marked LLs.
- If $l_1 = \emptyset$ and no lesions are marked, then Eqn. (1.17) 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:

$$\mathrm{TPF}_r \equiv \mathrm{TPF}(\zeta_r) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \mathbb{I} \left(TP_{k_2 2} \geq \zeta_r \right) \tag{1.18}$$

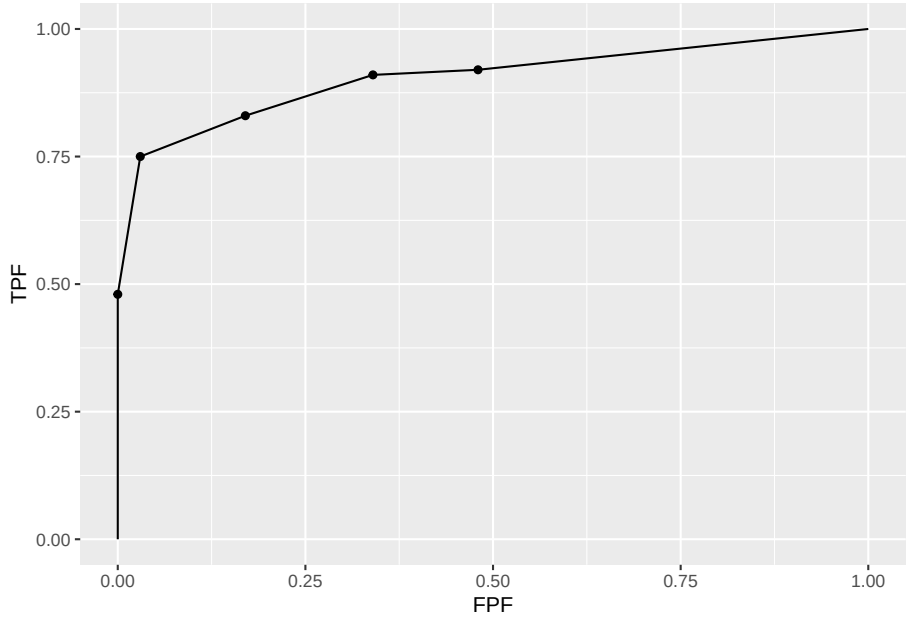
1.6.4 Definition empirical plot and AUC

The inferred empirical ROC plot connects adjacent points $(\mathrm{FPF}_r, \mathrm{TPF}_r)$, 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}.

1.6.5 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical ROC plot. The reader should experiment by running `PlotEmpiricalOperatingCharacteristics(dataset04, trts = 1, rdrs = 1, opChType = "ROC")$Plot` with different treatments and readers specified.

```
ret <- PlotEmpiricalOperatingCharacteristics(
  dataset04,
  trts = 1, rdrs = 1, opChType = "ROC")
print(ret$Plot)
```



Shown next is calculation of the figure of merit for this dataset [5].

```
UtilFigureOfMerit(dataset04, FOM = "HrAuc")
#>         rdr1    rdr3    rdr4    rdr5
#> trt1 0.90425 0.79820 0.81175 0.86645
#> trt2 0.86425 0.84470 0.82050 0.87160
#> trt3 0.81295 0.81635 0.75275 0.85730
#> trt4 0.90235 0.83150 0.78865 0.87980
#> trt5 0.84140 0.77300 0.77115 0.84800
```

[5] In function `UtilFigureOfMerit` the `FOM` argument has to be set to `HrAuc`, which denotes the highest rating inferred-ROC AUC.

1.7 The alternative FROC (AFROC) plot

- Fig. 4 in (Bunch et al., 1977) anticipated another way of visualizing FROC data. I subsequently termed[6] 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. (1.7).
- $\text{FPF}_r \equiv \text{FPF}(\zeta_r)$ was defined in Eqn. (1.14).

1.7.1 Definition empirical plot and AUC

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 AFROC AUC, denoted A_{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*, as detailed next.

1.7.2 The constrained observed end-point of the AFROC

Since $\zeta_{R_{FROC}+1} = \infty$, according to Eqn. (1.7) and Eqn. (1.14), $r = R_{FROC} + 1$ yields the trivial operating point (0,0). Likewise, since $\zeta_0 = -\infty$, $r = 0$ yields the trivial point (1,1):

$$\left.\begin{aligned} \text{FPF}_{R_{FROC}+1} =& \frac{1}{K_1}\sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq \infty\right) \\ =& 0 \\ \text{LLF}_{R_{FROC}+1} =& \frac{1}{L_T}\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(LL_{k_2 2 l_2 2} \geq \infty\right) \\ =& 0 \end{aligned}\right\} \tag{1.19}$$

[6] 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

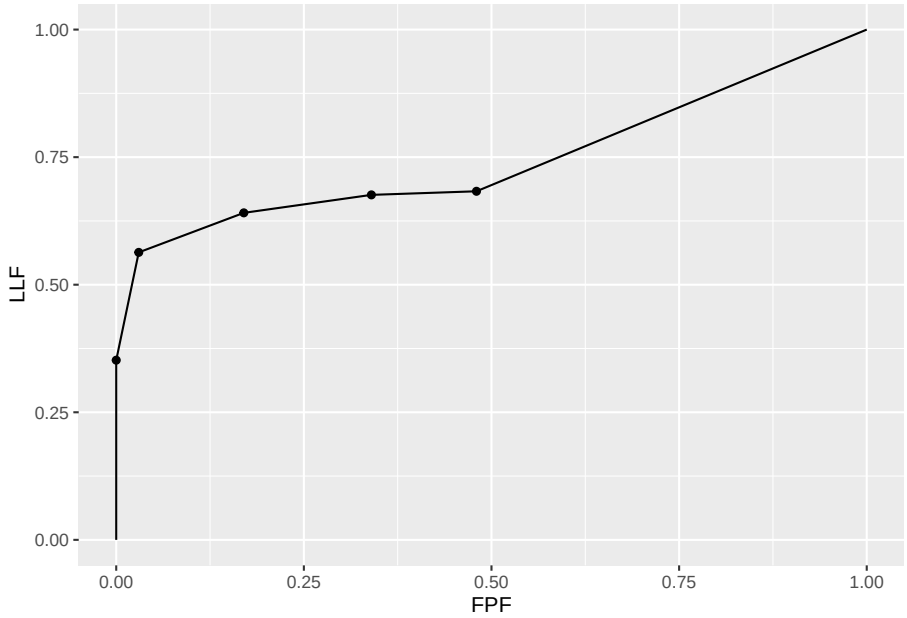
$$\left.\begin{aligned} \mathrm{FPF}_0 =& \frac{1}{K_1}\sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq -\infty\right) \\ =& 1 \\ \mathrm{LLF}_0 =& \frac{1}{L_T}\sum_{k_2=1}^{K_2}\sum_{l_2=1}^{L_{k_2 2}} \mathbb{I}\left(LL_{k_2 2 l_2 2} \geq -\infty\right) \\ =& 1 \end{aligned}\right\} \tag{1.20}$$

Because every non-diseased case is assigned a rating, and is therefore counted, the right hand side of the first equation in (1.20) 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. (1.20) is true even when the inferred-FP rating is $-\infty$).

1.7.3 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical AFROC plot. The reader should experiment by running `PlotEmpiricalOperatingCharacteristics(dataset04, trts = 1, rdrs = 1, opChType = "AFROC")$Plot` with different treatments and readers specified.

```
ret <- PlotEmpiricalOperatingCharacteristics(
  dataset04,
  trts = 1, rdrs = 1, opChType = "AFROC")
print(ret$Plot)
```



Shown next is calculation of the figure of merit for this dataset.

```
UtilFigureOfMerit(dataset04, FOM = "AFROC")
#>           rdr1      rdr3      rdr4      rdr5
#> trt1 0.7427113 0.7104930 0.7003169 0.7909859
#> trt2 0.7586972 0.7161620 0.7225352 0.7927465
#> trt3 0.6983451 0.6955282 0.6777817 0.7547535
#> trt4 0.7817606 0.7234507 0.7132746 0.8136268
#> trt5 0.7169718 0.6690845 0.6587324 0.7682042
```

1.8 The weighted-AFROC plot (wAFROC) plot

The AFROC ordinate defined in Eqn. (1.7) gives equal importance to every lesion in 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. [7]

[7] Historical note: I became aware of how serious this issue could be when a researcher contacted me 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!

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 $W_{k_2 l_2} \geq 0$ denote the *weight* (i.e., short for 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., the notation $W_{k_2 2 l_2 2}$ would be superfluous). For each diseased case $k_2 2$ the weights are subject to the constraint:

$$\sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} = 1 \tag{1.21}$$

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):

$$\text{wLLF}_r \equiv \text{wLLF}\left(\zeta_r\right) = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2 2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 l_2 2} \geq \zeta_r\right) \tag{1.22}$$

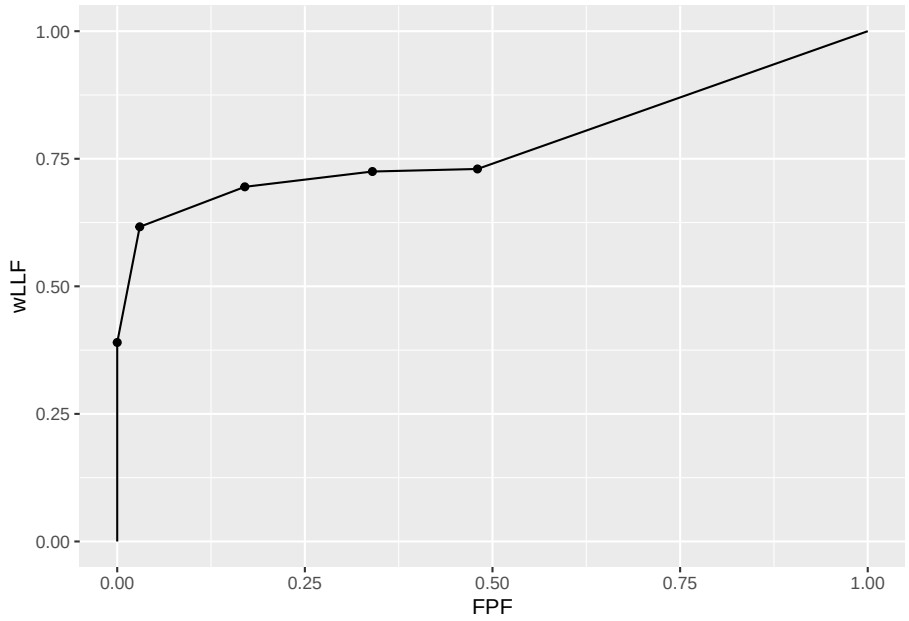
1.8.1 Definition empirical plot and AUC

The empirical wAFROC plot connects adjacent operating points $(\text{FPF}_r, \text{wLLF}_r)$, 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}.

1.8.2 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical ROC plot. The reader should experiment by running `PlotEmpiricalOperatingCharacteristics(dataset04, trts = 1, rdrs = 1, opChType = "wAFROC")$Plot` with different treatments and readers specified.

```
ret <- PlotEmpiricalOperatingCharacteristics(
  dataset04, trts = 1, rdrs = 1, opChType = "wAFROC")
print(ret$Plot)
```



Shown next is calculation of the figure of merit for this dataset.

```
UtilFigureOfMerit(dataset04, FOM = "wAFROC")
#>           rdr1      rdr3      rdr4      rdr5
#> trt1 0.7792667 0.7248917 0.7036250 0.8050917
#> trt2 0.7870000 0.7269000 0.7226167 0.8037833
#> trt3 0.7296917 0.7157583 0.6723083 0.7726583
#> trt4 0.8101333 0.7431167 0.6943583 0.8294083
#> trt5 0.7488000 0.6822750 0.6551750 0.7712500
```

1.9 AFROC vs. wAFROC

The wAFROC and AFROC, particularly the concept that the wAFROC gives equal importance to each diseased case while the AFROC gives more important to diseased cases with more lesions, are perhaps best illustrated with a numerical example. The dataset consists of $K_1 = 4$ non-diseased and $K_2 = 4$ diseased cases. The first two diseased cases have one lesion each, and the remaining two have two lesions each.

Shown next is the NL ratings array which has 8 rows, corresponding to the total number of cases, and 2 columns, corresponding to the maximum number of NLs and LLs per case (notice there are two entries for case #3). The negative infinities represent unmarked locations.

```
#>
#> NL ratings:
#>      [,1] [,2]
#> [1,] -Inf -Inf
#> [2,]  0.5 -Inf
#> [3,]  0.7  0.6
#> [4,] -0.3 -Inf
#> [5,]  1.5 -Inf
#> [6,] -Inf -Inf
#> [7,] -Inf -Inf
#> [8,] -Inf -Inf
```

Shown next is the FP ratings array. Since FPs are only possible on non-diseased cases, this is a length 4 row-vector. Each value is the maximum of the two NL ratings for the corresponding non-diseased case. As an example, for case #3 the maximum of the two NL values is 0.7.

```
#>
#> FP ratings:
#> [1] -Inf  0.5  0.7 -0.3
```

Show next is the sorted FP ratings.

```
#> [1] -Inf -0.3  0.5  0.7
```

The sorting makes it easy to construct the FPF values, shown next.

```
#>
#> FPF values:
#>  0.000 0.000 0.000 0.000 0.000 0.250 0.500 0.500 0.750 1.000
```

The first non-zero FPF value is 0.25, which occurs when the sliding threshold is lowered past the highest FP value, namely 0.7. The 0.25 comes from 1 FP case divided by 4 non-diseased cases. The next FPF value is 0.5, which occurs when the sliding threshold is lowered past the next-highest FP value, namely 0.5. The 0.5 comes from 2 divided by 4. The next FPF value is 0.75 and the last FPF value is unity.

Shown next is the LL ratings array:

```
#>
#> LL ratings:
#>      [,1] [,2]
#> [1,]  0.9 -Inf
#> [2,] -0.2 -Inf
#> [3,]  1.6 -Inf
#> [4,]  3.0    2
```

Show next is the sorted LL ratings.

```
#> [1] -Inf -Inf -Inf -0.2  0.9  1.6  2.0  3.0
```

The sorting makes it easy to construct the LLF values, shown next.

```
#>
#> LLF values:
#>  0.000 0.167 0.333 0.500 0.667 0.667 0.667 0.833 0.833 1.000
```

The first non-zero LLF value is 0.167, which occurs when the sliding threshold is lowered past the highest LL value, namely 3. The 0.167 comes from 1 LL divided by 6 lesions. The next LLF value is 0.333, which occurs when the sliding threshold is lowered past the next-highest LL value, namely 2 ($2/6 = 0.333$). The next LLF value is 0.5, which occurs when the sliding threshold is lowered past 1.6 ($3/6 = 0.5$), and so on.

Show next is the lesion weights array:

```
#>
#> lesion weights:
#>      [,1] [,2]
#> [1,]  1.0 -Inf
#> [2,]  1.0 -Inf
#> [3,]  0.1  0.9
#> [4,]  0.9  0.1
```

Since the first two diseased cases have one lesion each, the [1,1] and [2,1] elements of the the weight array are each equal to unity and the [1,2] and [2,2] elements are each equal to negative infinity, which is being used as a missing value. For diseased case #3 the first lesion has weight 0.1 while the second lesion has weight 0.9 (the weights must sum to unity). For diseased case #4 the weights are reversed.

The sorted LL ratings array and the weights are used to construct the wLLF values shown next.

```
#>
#> wLLF values:
#>  0.000 0.225 0.250 0.275 0.525 0.525 0.525 0.775 0.775 1.000
```

The first non-zero `wLLF` value is 0.225, which occurs when the sliding threshold is lowered past the highest `LL` value, namely 3. Since this comes from the first lesion on diseased case #4, whose weight is 0.9, the corresponding incremental vertical jump is 1 divided by 4 (*sic*) times 0.9. Notice that we are dividing by 4, the total number of diseased cases, not 6 as in the `LLF` example.

The next `wLLF` value is 0.25, which occurs when the sliding threshold is lowered past the next-highest `LL` value, namely 2, which comes from the 2nd lesion on the fourth diseased case with weight 0.1. The incremental jump in `wLLF` is 1 divided by 4 times 0.1, which is 0.025. The net `wLLF` value corresponding to these two lesions is $1/4 * 0.9 + 1/4 * 0.1 = 1/4 = 0.25$. The two lesions on diseased case #4 contribute 0.25 in `wLLF` increment. In contrast, they contribute $2/6 = 0.333$ in `LLF` increment, showing explicitly that the AFROC gives greater importance to diseased cases with more lesions while the wAFROC does not.

The next `wLLF` value is 0.275, which occurs when the sliding threshold is lowered past 1.6, which ratings comes from the first lesion on diseased case #3, with weight 0.1, $1/4 * 0.9 + 1/4 * 0.1 + 1/4 * 0.1 = 0.275$, and so on.

The reader should complete these hand-calculations to reproduce all of the `wLLF` values shown above.

Shown in Fig. 1.1 are the empirical AFROC and wAFROC plots.

The operating points can be used to numerically calculate the AUCs under the empirical AFROC and wAFROC plots, as done in the following code:

```
afroc_auc <- 0.5 * 4 / 6 + 0.25 * 5 /6 +
 5 /6 * 0.25 +  (1 - 5 /6) * 0.25 /2

wafroc_auc <- 0.5 * 0.525 + 0.25 * 0.775 +
  0.775 * 0.25 +  (1 - 0.775) * 0.25 /2

afroc_auc
#> [1] 0.7708333
wafroc_auc
#> [1] 0.678125
```

The same AUC results are obtained using the function `UtilFigureOfMerit`

```
cat("\nAFROC AUC = ", as.numeric(UtilFigureOfMerit(frocData, FOM = "AFROC")),"\n")
#>
#> AFROC AUC =  0.7708333
cat("wAFROC AUC = ", as.numeric(UtilFigureOfMerit(frocData, FOM = "wAFROC")),"\n")
#> wAFROC AUC =  0.678125
```

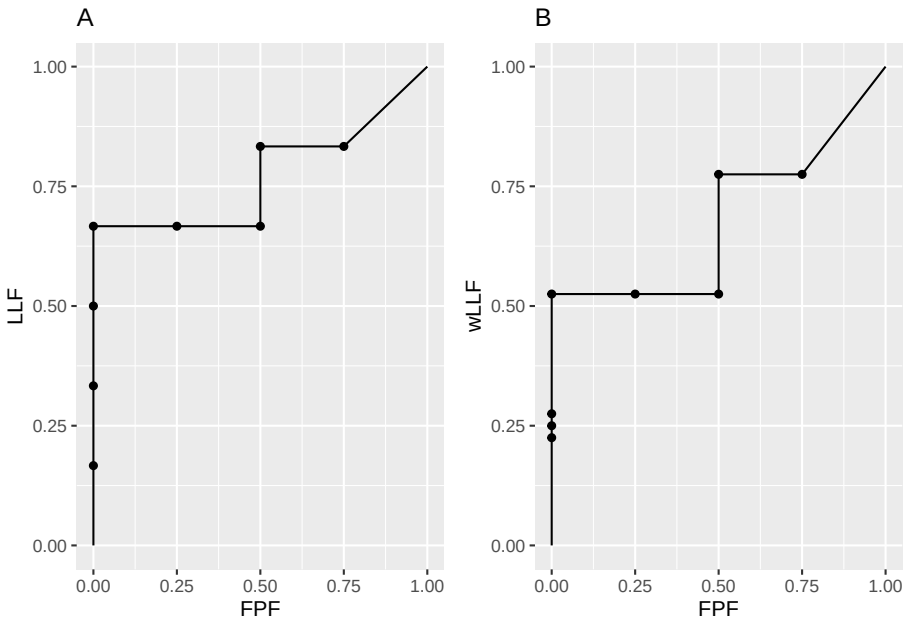


Figure 1.1: Left: AFROC plot; Right: corresponding wAFROC plot.

1.10 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. (1.13)]:

$$\left.\begin{aligned} FP^1_{k_t t} &= \max_{l_1}\left(z_{k_t t l_1 1} \mid l_1 \neq \emptyset\right) \\ &= -\infty \mid l_1 = \emptyset \end{aligned}\right\} \tag{1.23}$$

$FP^1_{k_t t}$ is the maximum over all latent NL marks, labeled by the location index l_1, occurring in case $k_t t$, or $-\infty$ if $l_1 = \emptyset$. The corresponding false positive fraction FPF^1_r is defined by [the "1" superscript is necessary to distinguish it from Eqn. (1.14)]:

$$FPF_r^1 \equiv FPF_r^1(\zeta_r) = \frac{1}{K_1 + K_2} \sum_{t=1}^{2} \sum_{k_t=1}^{K_t} \mathbb{I}\left(FP_{k_t t}^1 \geq \zeta_r\right) \tag{1.24}$$

Note the subtle differences between Eqn. (1.14) and Eqn. (1.24). The latter counts "FPs" on non-diseased and diseased cases while Eqn. (1.14) 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 or almost all cases are diseased.

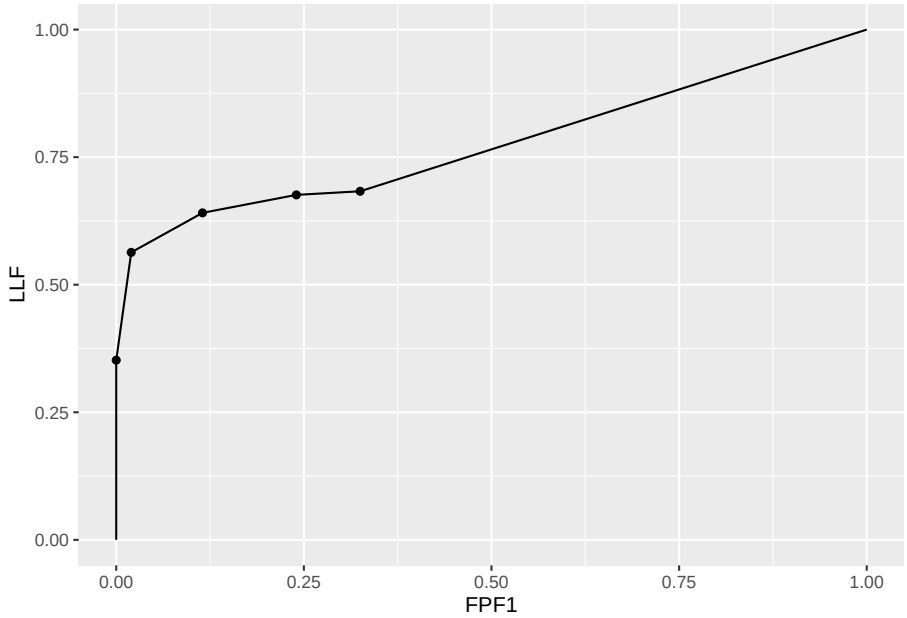
1.10.1 Definition empirical plot and AUC

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_{AFROC1}.

1.10.2 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical ROC plot. The reader should experiment by running `PlotEmpiricalOperatingCharacteristics(dataset04, trts = 1, rdrs = 1, opChType = "AFROC1")$Plot` with different treatments and readers specified.

```
ret <- PlotEmpiricalOperatingCharacteristics(
  dataset04,
  trts = 1, rdrs = 1, opChType = "AFROC1")
print(ret$Plot)
```



Shown next is calculation of the figure of merit for this dataset.

```
UtilFigureOfMerit(dataset04, FOM = "AFROC1")
#>            rdr1      rdr3      rdr4      rdr5
#> trt1 0.7744718 0.7157218 0.7229225 0.7913908
#> trt2 0.7826585 0.7278169 0.7364437 0.7897887
#> trt3 0.7412852 0.6868310 0.6946303 0.7573415
#> trt4 0.8087852 0.7346831 0.7343486 0.8155634
#> trt5 0.7580810 0.6825704 0.6643662 0.7742782
```

1.11 The weighted-AFROC1 (wAFROC1) plot

1.11.1 Definition empirical plot and AUC

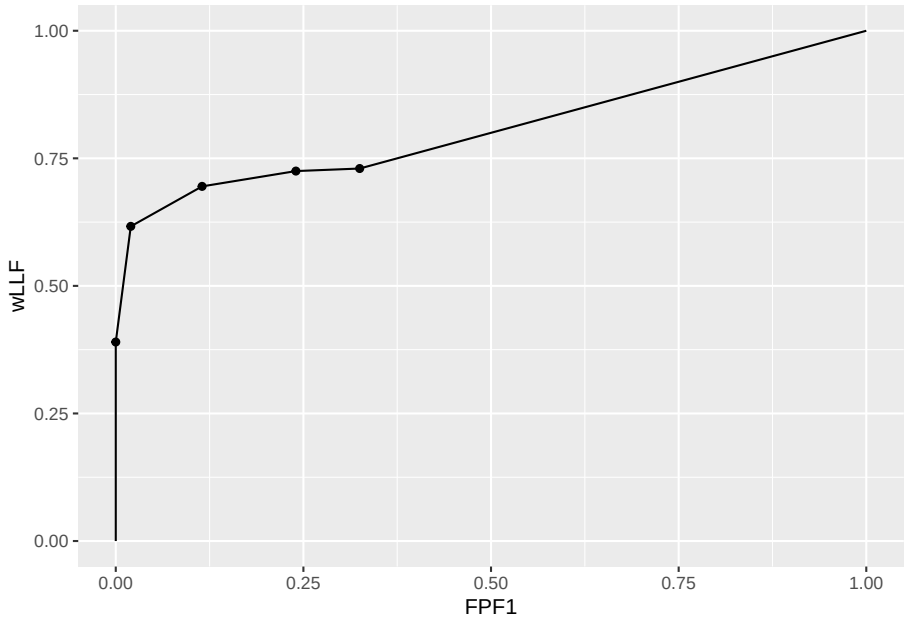
The empirical weighted-AFROC1 (wAFROC1) plot connects adjacent operating points (FPF_r^1, wLLF_r), including the origin (0,0) and (1,1), with straight lines. The only difference between it and the wAFROC plot is in the x-axis. The area under this plot is the empirical weighted-AFROC AUC, denoted A_{wAFROC1}.

1.11.2 Illustration with a dataset

The following code uses `dataset04` to illustrate an empirical wAFROC plot1.
The reader should experiment by running `PlotEmpiricalOperatingCharacteristics(dataset04,`

`trts = 1, rdrs = 1, opChType = wAFROC1")$Plot` with different treatments and readers specified.

```
ret <- PlotEmpiricalOperatingCharacteristics(
  dataset04,
  trts = 1, rdrs = 1, opChType = "wAFROC1")
print(ret$Plot)
```



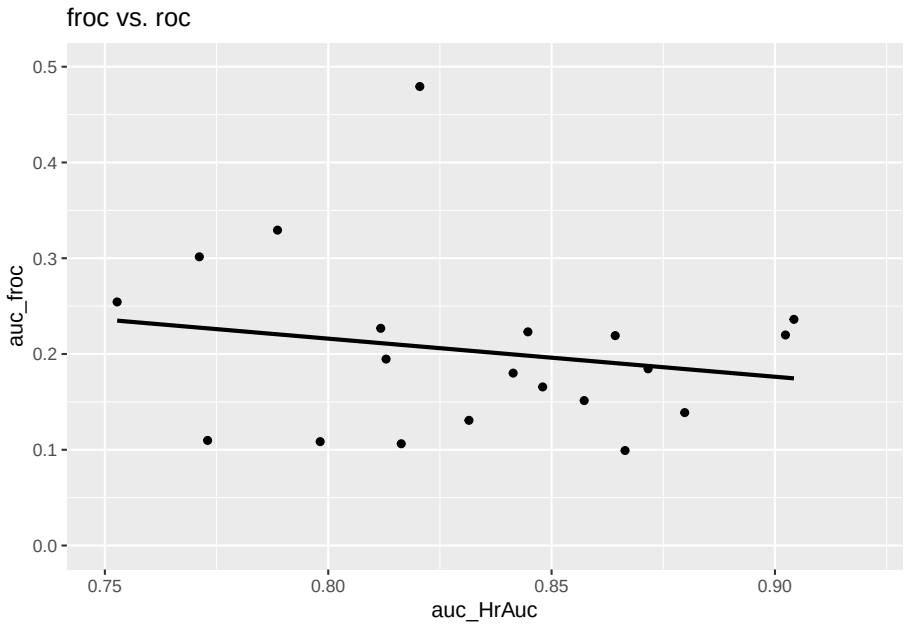
Shown next is calculation of the figure of merit for this dataset.

```
UtilFigureOfMerit(dataset04, FOM = "wAFROC1")
#>            rdr1      rdr3      rdr4      rdr5
#> trt1 0.8068333 0.7298917 0.7262042 0.8058542
#> trt2 0.8084625 0.7379917 0.7363083 0.8010167
#> trt3 0.7680875 0.7075583 0.6890208 0.7743875
#> trt4 0.8348750 0.7533917 0.7160250 0.8308333
#> trt5 0.7857708 0.6953292 0.6605167 0.7774000
```

1.12 Plots of FROC, AFROC and wAFROC AUCs vs. ROC AUC

Plots of A_{FROC}, A_{AFROC} and A_{wAFROC} vs. A_{ROC} were generated for the dataset used in the previous illustrations.

The following is the plot of A_{FROC} vs. A_{ROC}. There are 20 points on the plot corresponding to 5 treatments and 4 readers. The straight line is a least squares fit. A_{ROC} is assumed to be the gold standard. Note the poor correlation between A_{FROC} and A_{ROC}. The slope is negative and there is much scatter.

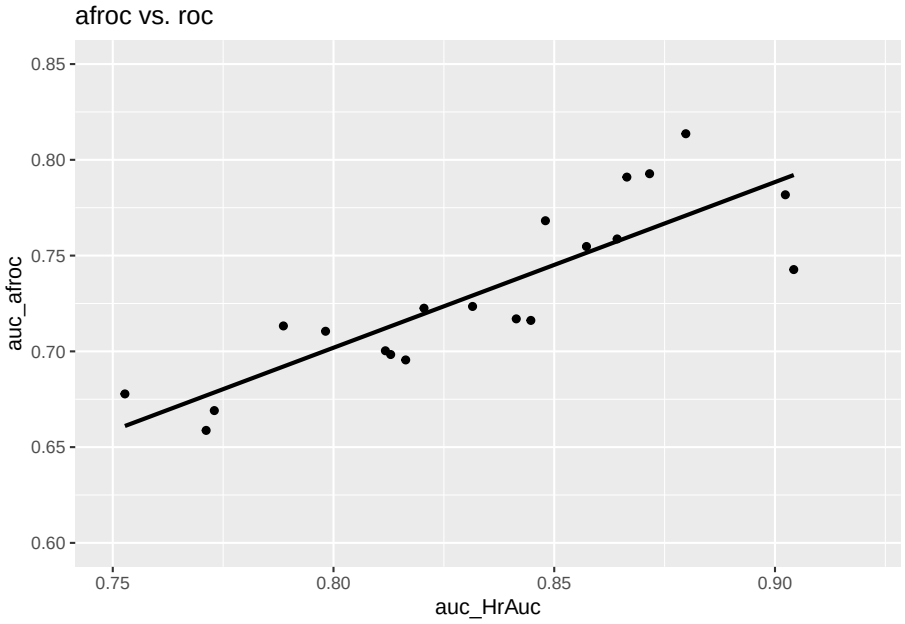


Evidently A_{FROC} does not correlate well with A_{ROC}. The reason should be fairly obvious. The FROC is unconstrained in the NLF direction and the area under the plot rewards an observer who generates more and more NLs while not generating more LLs, i.e., as the operating point moves further to the right along the flat part of the plot. In fact the perfect observer, Section `\@ref(froc-paradigm-solar-analogy)`, whose FROC plot is the vertical line connecting (0,0) and (0,1) has zero A_{FROC}! One can try ot avoid this problem by limiting the area under the FROC to that between NLF = 0 and NLF = x where x is some arbitrarily chosen fixed value – indeed this procedure has been used by many CAD algorithm designers. Since the choice of x is arbitrary the procedure would be subjective and totally dependent on the algorithm designer. Moreover it would not solve the problem that the perfect observer would still yield $A_{\mathrm{FROC}} = 0$. The perfect observer problem is not academic as the method would fail for any observer with $\mathrm{NLF}_{max} < x$. For such an observer the partial area would be undefined. This would force the algorithm designer to chose x as the minimum of all NLF_{max} values over all observers and treatments, which would exclude a lot of data from the analysis leading to a severe statistical power penalty.

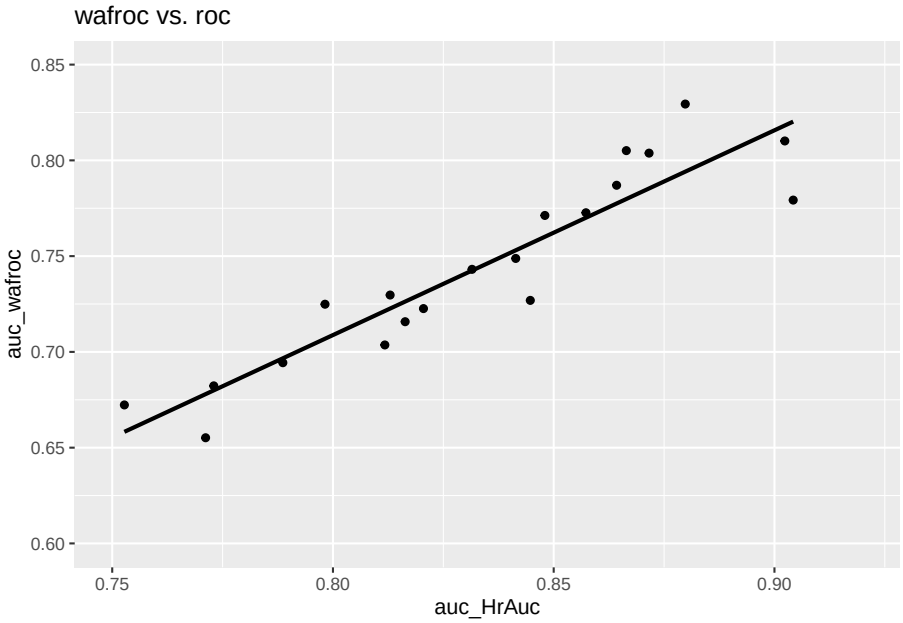
The basic problem is that the FROC plot is unconstrained in the NLF direction.

The following is the plot of A_{AFROC} vs. A_{ROC}. This time there is a strong

positive correlation between the two. The reason is that the AFROC is fully contained in the unit square. An observer who generates more NL marks would in fact yield smaller A_{AFROC} - this is explained in a later section TBA .



The following is the plot of A_{wAFROC} vs. A_{ROC}. Again, there is a strong positive correlation between the two. The wAFROC is also fully contained in the unit square.



1.13 TBA 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.

FROC data consists of mark-rating pairs. In this chapter a distinction is made between latent and actual marks. This is followed by a table summarizing FROC notation. This is a key table which will be used in later chapters. Section `\@ref(froc-paradigm-froc-plot)` introduced the empirical FROC plot. This chapter presents mathematical expressions for this and other empirical plots possible with FROC data: the inferred-ROC, the alternative FROC, the weighted alternative FROC, and others. Operating characteristics are *visual* depicters of performance. Scalar quantities, typically area measures derived from operating characteristics, are *quantitative* measures of performance, termed *figures of merit* (FOMs). This chapter defines an area measure for each empirical operating characteristic. An FROC dataset is used to illustrate the plots and area measures. With the exception of the FROC, all empirical plots include a straight line extension from the observed end-point to (1,1). The correlation between the area measures is qualitatively examined via plots. It is shown that

for this dataset the FROC area measure correlates poorly with that under the ROC curve, whereas the other measures correlate better. This is explained by the fact that, unlike the other measures, the FROC plot is not contained within the unit square.

1.14 References

Chapter 2

Meanings of FROC figures of merit

2.1 TBA How much finished

50%

2.2 Introduction

Chapter TBA `\@ref(froc-empirical)` 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 formulae for calculating them directly from ratings. In this sense this chapter is the extension to the FROC paradigm of Chapter TBA (empirical-auc) 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 θ, 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 θ_{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 θ_{AFROC} and the empirical weighted-AFROC FOM-statistic θ_{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 θ_{wAFROC} and the area A_{wAFROC} under the empirical wAFROC plot. [A similar equality applies to the empirical AFROC FOM-statistic θ_{wAFROC} and the area A_{wAFROC} 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.

2.3 Empirical AFROC FOM-statistic

A_{wAFROC} was defined in TBA `\@ref(froc-empirical-wAFROC)` as the area under the empirical AFROC. The corresponding FOM-statistic θ_{wAFROC} is defined as follows: one calculates the rating of the highest rated NL mark $\mathrm{FP}_{k_1 1}$ 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. TBA (eq:empirical-auc-PsiFunction) [1]. A summation is performed over all cases and all lesions.

The highest rating $\mathrm{FP}_{k_1 1}$ on non-diseased case $k_1 1$ is defined as:

$$\left.\begin{aligned} \mathrm{FP}_{k_1 1} &= \max_{l_1} \left(z_{k_1 1 l_1 1} \mid l_1 \neq \emptyset \right) \\ \mathrm{FP}_{k_1 1} &= -\infty \mid l_1 = \emptyset \end{aligned}\right\} \tag{2.1}$$

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:

$$\theta_{wAFROC} = \frac{1}{K_1 L_T} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} \psi \left(\mathrm{FP}_{k_1 1}, z_{k_2 2 l_2 2} \right) \tag{2.2}$$

Since every lesion is assigned a rating, albeit negative infinity for an unmarked lesion, the null set conditioning is not needed.

2.3.1 Upper limit for AFROC FOM-statistic

The FOM-statistic θ_{wAFROC} achieves its highest value, unity, if and only if every lesion is rated higher than any mark on non-diseased cases, for then the ψ function always yields unity, and the summations yield :

[1] The kernel function comparison yields 1 if the LL rating is higher, 0.5 if the ratings are identical and zero otherwise.

$$\left.\begin{aligned}\theta_{wAFROC} =& \frac{1}{K_1 \sum_{k_2=1}^{K_2} L_{k_2}} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} 1 \\ =& \frac{1}{K_1 \sum_{k_2=1}^{K_2} L_{k_2}} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} L_{k_2} \\ =& \frac{1}{K_1} \sum_{k_1=1}^{K_1} 1 \\ =& 1\end{aligned}\right\} \tag{2.3}$$

2.3.2 Range of AFROC FOM-statistic

If, on the other hand, every lesion is rated lower than every mark on every non-diseased case, the ψ function always yields zero, and the FOM-statistic is zero. Therefore,

$$0 \leq \theta_{wAFROC} \leq 1 \tag{2.4}$$

Eqn. (2.4) shows that θ_{wAFROC} behaves like a probability but its range is *twice* that of θ_{ROC}; recall that $0.5 \leq \theta_{ROC} \leq 1$ (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 θ_{wAFROC} (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. (2.4) 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.

2.4 Empirical weighted-AFROC FOM-statistic

The empirical weighted-AFROC plot and lesion weights were defined in SectionTBA `\@ref(froc-empirical-wAFROC)`. 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):

$$\theta_{\text{wAFROC}} = \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} \psi\left(\text{FP}_{k_1 1}, z_{k_2 2 l_2 2}\right) \tag{2.5}$$

The weights obey the constraint:

$$\sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} = 1 \tag{2.6}$$

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, θ_{wAFROC} and θ_{wAFROC} are identical. For equally weighted lesions,

$$W_{k_2 l_2} = \frac{1}{L_{k_2}} \tag{2.7}$$

For example, for equally weighted lesions and a case with three lesions, each weight equals one-third (1/3) [2].

2.5 Two Theorems

The area A_{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. 2.1.

The operating point labeled i has coordinates $(\text{FPF}_i, \text{wLLF}_i)$ given by Eqn. TBA `\@ref(eq:froc-empirical-FPF)` and Eqn. TBA `\@ref(eq:froc-empirical-wLLFr)`, respectively, reproduced here for convenience:

$$FPF_i \equiv FPF\left(\zeta_i\right) = \frac{1}{K_1} \sum_{k_1=1}^{K_1} \mathbb{I}\left(FP_{k_1 1} \geq \zeta_i\right) \tag{2.8}$$

$$\text{wLLF}_i \equiv \text{wLLF}_{\zeta_i} = \frac{1}{K_2} \sum_{k_2=1}^{K_2} \sum_{l_2=1}^{L_{k_2}} W_{k_2 l_2} \mathbb{I}\left(z_{k_2 l_2 2} \geq \zeta_i\right) \tag{2.9}$$

TBA Online Appendix 14.A proves the following theorems:

[2] 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.

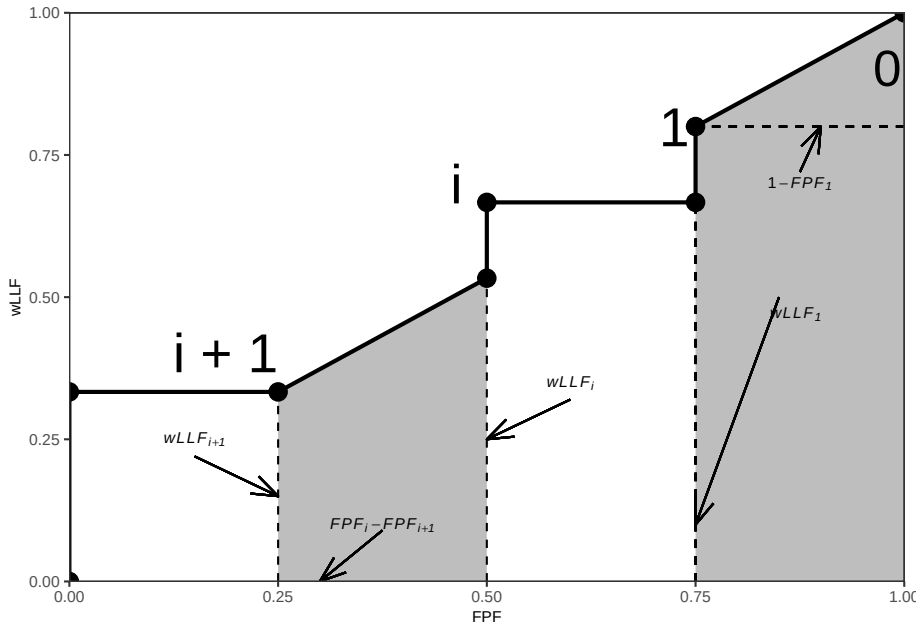


Figure 2.1: An example wAFROC plot; from left to right, the two shaded areas correspond to A_i and A_0, respectively, defined below.

2.5.1 Theorem 1

The area A_{wAFROC} under the empirical wAFROC plot equals the weighted-AFROC FOM-statistic θ_{wAFROC} defined by Eqn. (2.5):

$$\theta_{\text{wAFROC}} = A_{\text{wAFROC}} \tag{2.10}$$

This is the FROC counterpart of Bamber's Wilcoxon vs. empirical ROC area equivalence theorem (Bamber, 1975), derived in Section TBA (empirical-auc-wilcoxon-bamber-theorem).

2.5.2 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:

$$A_0 = \frac{(1 - \text{FPF}_1)(1 + \text{wLLF}_1)}{2} \tag{2.11}$$

According to Eqn. (2.11), A_0 increases as FPF_1 decreases, i.e., as more non-diseased cases are *not marked* and as 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. (2.11), the area under the straight line extension is $A_0 = 1$.

2.6 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.

2.6.1 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 x 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.

2.6.2 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 0.9 x 4 + 0.1 x 4, i.e., 4 (compared to 8 in the preceding example). The third diseased case contributes 0.1 x 4 + 0.9 x 0.5, i.e., 2.6 (compared to 4.5 in the preceding example). The second diseased case contributes 1 x 2 = 2 (compared to 2 in the preceding example), and the first diseased case contributes 1 x 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.

2.7 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 statistic5.

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"6 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 available7-9. 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 anal-

ysis 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 al10 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 Popescu11 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 curves12-14 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.

2.8 References

Bibliography

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.

Bunch, P. C., Hamilton, J. F., Sanderson, G. K., and Simmons, A. H. (1977). A free response approach to the measurement and characterization of radiographic observer performance. In *Application of Optical Instrumentation in Medicine VI*, volume 127, pages 124–135. International Society for Optics and Photonics.

Chakraborty, D. P. (1989). Maximum likelihood analysis of free-response receiver operating characteristic (froc) data. *Medical physics*, 16(4):561–568.

Chakraborty, D. P. (2006). A search model and figure of merit for observer data acquired according to the free-response paradigm. *Physics in Medicine & Biology*, 51(14):3449.

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. 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.

DeSantis, C., Siegel, R., Bandi, P., and Jemal, A. (2011). Breast cancer statistics, 2011. *CA: a cancer journal for clinicians*, 61(6):408–418.

Duchowski, A. T. (2002). *Eye Tracking Methodology: Theory and Practice*. Clemson University, Clemson, SC.

Swensson, R. G. (1996). Unified measurement of observer performance in detecting and localizing target objects on images. *Medical physics*, 23(10):1709–1725.

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.