Chapter 05: Empirical AUC

Table of contents

Online Supplementary Material

Online Appendix 5.A: Calculating the Wilcoxon statistic

5.1: 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 μ 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, 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 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.

5.2: 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.]

5.2.1: 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.

Table 5.1: On the need for two indices to label cases in an ROC study. The upper row denotes 11 non-diseased cases, labeled N1 – N11, while the lower row denotes seven diseased cases, labeled D1 – D7. To address any case one needs two indices: the row number (t = 1 or t = 2) 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.

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

To address a case one needs *two* indices; the first index is the column number and the second index is the row number, and moreover the total number of columns can, and does in this example, depend on the row number. This means the first index has to be *t*-dependent, i.e., k_t, denoting the column index 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, the author prefers the two-index notation.

5.2.2: An empirical operating point

Let $z_{k_t t}$ represent the (realized) *z*-sample of case $k_t t$. For a given reporting threshold ζ, 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:

$$FPF(\zeta) = \frac{1}{K_1}\sum_{k_1=1}^{K_1} I\left(z_{k_1 1} \geq \zeta\right) \quad . \qquad \textbf{(5.1)}$$

$$TPF(\zeta) = \frac{1}{K_2}\sum_{k_2=1}^{K_2} I\left(z_{k_2 2} \geq \zeta\right) \quad . \qquad \textbf{(5.2)}$$

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

$$\left.\begin{aligned} I(TRUE) &= 1 \\ I(FALSE) &= 0 \end{aligned}\right\} \quad . \qquad \textbf{(5.3)}$$

In Eqn. (5.1) and Eqn. (5.2), the indicator functions act as *counters*, effectively counting instances where the z-sample of a case equals or exceeds ζ, and division by the appropriate denominator yields the desired left hand sides of these equations. The operating point $O(\zeta)$ corresponding to threshold ζ is defined by:

$$O(\zeta) = \left(FPF(\zeta), TPF(\zeta) \right) \quad . \qquad \textbf{(5.4)}$$

The difference between Eqn. (5.1) and Eqn. (5.2) vs. Eqn. (3.22) and Eqn. (3.23) corresponds to that between parametric and non-parametric methods. In **Chapter 03** analytical (or parametric, i.e., depending on model parameters) operating points were obtained. In contrast, in Eqn. (5.1) and Eqn. (5.2), one uses the *observed* ratings to calculate the empirical operating point.

5.3: 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. 4.13, reproduced below:

$$\left.\begin{array}{l} \zeta_r = r + 1 \\ r = 1,..,R-1; \\ \zeta_0 = -\infty;\ \zeta_R = +\infty \end{array}\right\} \quad . \qquad \textbf{(5.5)}$$

The *discrete* operating point O_r is obtained by replacing ζ with ζ_r in Eqn. (5.4), where:

$$O_r \equiv O(\zeta_r) \equiv \left(FPF_r, TPF_r \right) \quad . \qquad \textbf{(5.6)}$$

Its coordinates are defined by:

$$\left.\begin{array}{l} FPF_r \equiv FPF(\zeta_r) = \dfrac{1}{K_1} \displaystyle\sum_{k_1=1}^{K_1} I\left(z_{k_1 1} \geq \zeta_r \right) \\ TPF_r \equiv TPF(\zeta_r) = \dfrac{1}{K_2} \displaystyle\sum_{k_2=1}^{K_2} I\left(z_{k_2 2} \geq \zeta_r \right) \end{array}\right\} \quad . \qquad \textbf{(5.7)}$$

For example,

$$\left.\begin{aligned} FPF_4 &= \frac{1}{K_1}\sum_{k_1=1}^{K_1} I\left(z_{k_1 1} \geq 5\right) = 1/60 = 0.017 \\ TPF_4 &= \frac{1}{K_2}\sum_{k_2=1}^{K_2} I\left(z_{k_2 2} \geq 5\right) = 22/50 = 0.44 \\ O_4 &\equiv \left(FPF_4, TPF_4\right) = \left(0.017, 0.44\right) \end{aligned}\right\} \quad . \qquad \textbf{(5.8)}$$

In Table 4.1, a sample clinical ratings data set was introduced. For convenience it is reproduced below, Table 5.2. In this example, $R = 5$, corresponding to the 5-ratings bins used to acquire the data.

Table 5.2: A typical ROC counts table, reproduced from **Chapter 04**, Table 4.1.

	Counts in ratings bins				
	Rating = 1	Rating = 2	Rating = 3	Rating = 4	Rating = 5
$K_1 = 60$	30	19	8	2	1
$K_2 = 50$	5	6	5	12	22
	Operating points				
	≥5	≥4	≥3	≥2	≥1
FPF	0.017	0.050	0.183	0.500	1
TPF	0.440	0.680	0.780	0.900	1

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.

5.3.1 Code listing (partial)

```
rm(list = ls()) # mainEmpRocPlot.R
...
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)
...
grid.draw(p)
```

Line 4 – 6 constructs the counts table shown in Table 5.2 and implements Eqn. (5.7). 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 simple appearing code implements Eqn. (5.7); see §5.3.2.

5.3.2 Code snippets

```
> c(30, 19, 8, 2, 1)
```

```
[1] 30 19  8  2  1
> rev(c(30, 19, 8, 2, 1))
[1]  1  2  8 19 30
> cumsum(rev(c(30, 19, 8, 2, 1)))
[1]  1  3 11 30 60
> c(0, cumsum(rev(c(30, 19, 8, 2, 1))) / K1)
[1] 0.00000000 0.01666667 0.05000000 0.18333333 0.50000000 1.00000000
```

Fig. 5.1 is the empirical ROC plot, produced by sourcing **`mainEmpRocPlot.R`**. It illustrates the convention used to label the operating points is that introduced in §4.3 is, i.e., O_1 is the uppermost non-trivial point, and the labels are incremented by unity as one moves down the plot. By convention, not shown are the trivial operating points $O_0 \equiv \left(FPF_0, TPF_0\right) \equiv \left(1,1\right)$ and $O_R \equiv \left(FPF_R, TPF_R\right) \equiv \left(0,0\right)$.

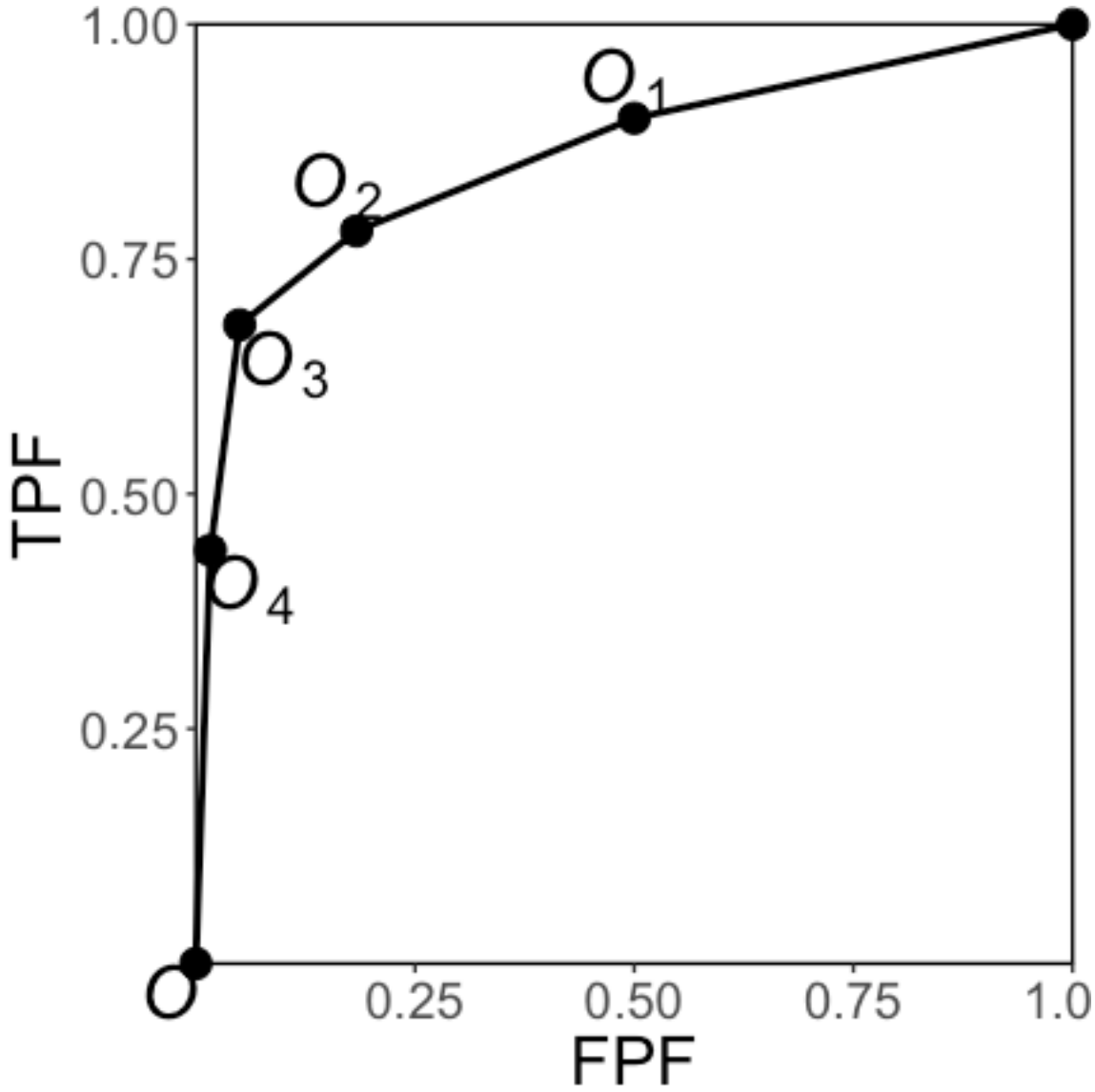

Fig. 5.1: Convention: The operating points are numbered starting with the uppermost non-trivial one, which is O_1, and working down the plot. This figure corresponds to the data in **Error! Reference source not found.**. The plot was produced by **`mainEmpROCPlot.R`**.

5.4: AUC under the empirical ROC plot

Fig. 5.2 shows the empirical plot for the data in Table 5.2, and 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[1,2], 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.

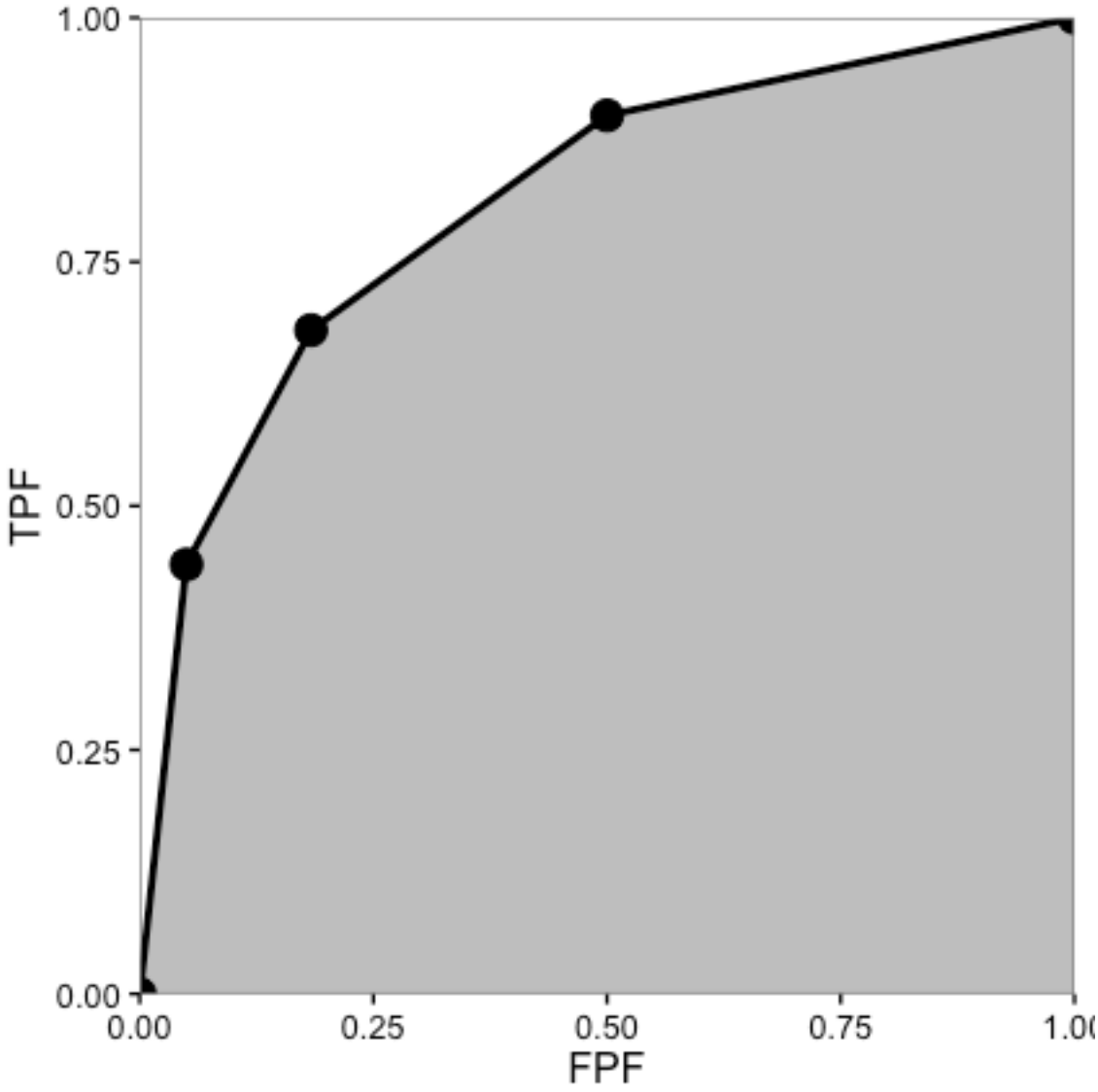

Fig. 5.2: The empirical ROC plot corresponding to Table 5.2; the shaded area is the area under this plot, a widely used figure of merit in non-parametric ROC analysis. The plot was produced by the code in **`mainEmpiricalAUC.R`**.

5.5: The Wilcoxon statistic

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

$$W = \frac{1}{K_1 K_2} \sum_{k_1}^{K_1} \sum_{k_2}^{K_2} \psi\left(z_{k_1 1}, z_{k_2 2}\right) \quad . \qquad \textbf{(5.9)}$$

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

$$\begin{aligned} \psi(x,y) &= 1 && x < y \\ \psi(x,y) &= 0.5 && x = y \\ \psi(x,y) &= 0 && x > y \end{aligned} \quad . \qquad \textbf{(5.10)}$$

The function ψ 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. (5.9) the two summations and division by the total number of comparisons yields the observed or *empirical* probability that diseased cases are rated higher than non-diseased ones. Since it is a probability, it can theoretically range from zero to one. However, if the observer has any discrimination ability, one expects diseased cases to be rated equal or greater than non-diseased ones, so in practice one expects:

$$0.5 \le W \le 1 \quad . \qquad \textbf{(5.11)}$$

The limit 0.5 corresponds to a guessing observer, whose operating point moves along the chance diagonal of the ROC plot.

5.6: Bamber's Equivalence theorem

Here is the result: the Wilcoxon statistic W equals the area AUC under the *empirical* ROC plot:

$$AUC = W \quad . \qquad \textbf{(5.12)}$$

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. (5.12) and compare it to the result of a direct integration of the area under the empirical ROC curve (i.e., adding the area of a triangle and several trapezoids). **R** provides a function that does just that. It is part of package **caTools** and is called **trapz(x,y)**. 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 empirical area AUC are implemented in Online Appendix 5.A, in file **mainWilcoxon.R**. **Source** this file to get the following output.

5.6.1: Code Output

```
> source('~/book2/02 A ROC analysis/A5 Empirical AUC/software/mainWilcoxon.R')
```

```
The wilcoxon statistic is =  0.8606667
direct integration yields AUC =  0.8606667
```

Note the equality of the two estimates.

Proof: The following proof is adapted from a paper by Bamber [3] 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 (floating point variable) scale. The reason is as follows: 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 – this is the reason why the proof is quite general*.

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., $P\left(Z_2 = \zeta_i\right)$ is *not zero*, as would be the case for continuous ratings.

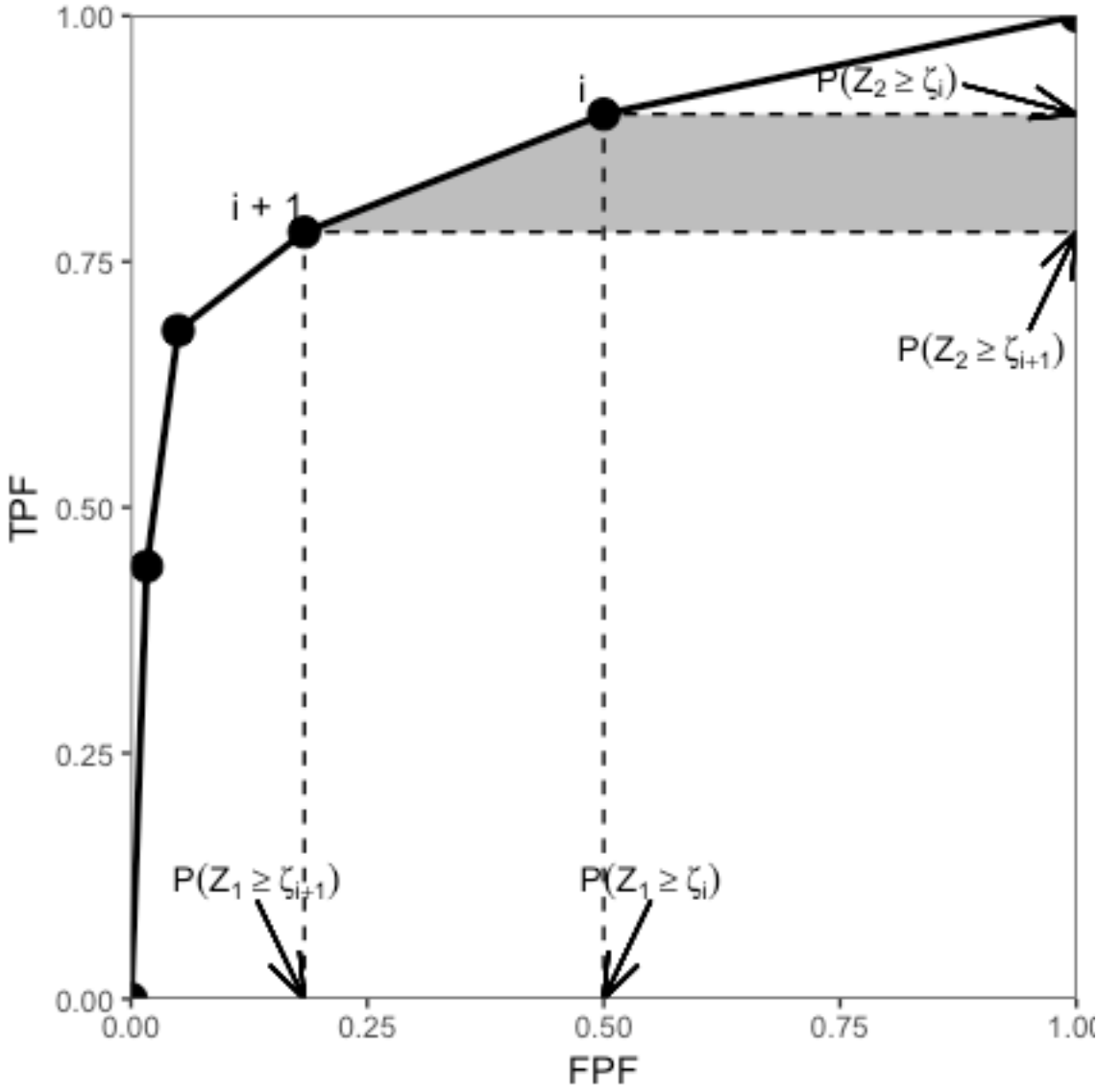

Fig. 5.3: Illustration of the derivation of Bamber's equivalence theorem. Shows an empirical ROC plot for R = 5; the shaded area is due to points labeled i and i + 1. This figure was created by the code in **`MainBamberTheorem.R`**.

The abscissa of the operating point i is $P\left(Z_1 \geq \zeta_i\right)$ and the corresponding ordinate is $P\left(Z_2 \geq \zeta_i\right)$. 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. 5.3, created by the code in **`MainBamberTheorem.R`**, see Online Appendix 5.A, has height:

$$P\left(Z_2 \geq \zeta_i\right) - P\left(Z_2 \geq \zeta_{i+1}\right) = P\left(Z_2 = \zeta_i\right). \qquad \textbf{(5.13)}$$

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

$$P\left(Z_2 \geq \zeta_i\right) = P\left(Z_2 = \zeta_i\right) + P\left(Z_2 \geq \zeta_{i+1}\right).$$

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

$$1 - P\left(Z_1 \geq \zeta_i\right) = P\left(Z_1 < \zeta_i\right),$$

and

$$1 - P\left(Z_1 \geq \zeta_{i+1}\right) = P\left(Z_1 < \zeta_{i+1}\right).$$

The area A_i of the shaded trapezoid in Fig. 5.3 is (the steps are shown explicitly):

$$\left.\begin{aligned} A_i &= \frac{1}{2} P\left(Z_2 = \zeta_i\right)\left[P\left(Z_1 < \zeta_i\right) + P\left(Z_1 < \zeta_{i+1}\right)\right] \\ A_i &= P\left(Z_2 = \zeta_i\right)\left[\frac{1}{2} P\left(Z_1 < \zeta_i\right) + \frac{1}{2}\left(P\left(Z_1 = \zeta_i\right) + P\left(Z_1 < \zeta_i\right)\right)\right] \\ A_i &= P\left(Z_2 = \zeta_i\right)\left[\frac{1}{2} P\left(Z_1 = \zeta_i\right) + P\left(Z_1 < \zeta_i\right)\right] \end{aligned}\right\} . \qquad \textbf{(5.14)}$$

In going from the first to the second line of Eqn. (5.14), use has been made of the last relation, below, derived from Eqn. (5.13) after replacing 2 with 1 and expressing the two probabilities on the left hand side of Eqn. (5.13) in terms of their complementary probabilities:

$$\left.\begin{aligned} 1 - P\left(Z_1 < \zeta_i\right) - \left(1 - P\left(Z_1 < \zeta_{i+1}\right)\right) &= P\left(Z_1 = \zeta_i\right) \\ P\left(Z_1 < \zeta_{i+1}\right) - P\left(Z_1 < \zeta_i\right) &= P\left(Z_1 = \zeta_i\right) \\ P\left(Z_1 < \zeta_{i+1}\right) &= P\left(Z_1 = \zeta_i\right) + P\left(Z_1 < \zeta_i\right) \end{aligned}\right\} . \qquad \textbf{(5.15)}$$

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

$$AUC = \sum_{i=0}^{R-1} A_i = \frac{1}{2}\sum_{i=0}^{R-1} P\left(Z_2 = \zeta_i\right)P\left(Z_1 = \zeta_i\right) + \sum_{i=0}^{R-1} P\left(Z_2 = \zeta_i\right)P\left(Z_1 < \zeta_i\right) \quad . \qquad \textbf{(5.16)}$$

It is shown in Appendix 5.A that the term A_0 corresponds to the triangle at the upper right corner of Fig. 5.3. Likewise, the term A_4 corresponds to the horizontal trapezoid defined by the lowest non-trivial operating point.

Eqn. (5.16) can be restated as:

$$AUC = \frac{1}{2}P\left(Z_2 = Z_1\right) + P\left(Z_1 < Z_2\right) \quad . \qquad \textbf{(5.17)}$$

The Wilcoxon statistic was defined in Eqn. (5.9). 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 AUC in Eqn. (5.17). Therefore,

$$AUC = W \quad . \qquad \textbf{(5.18)}$$

■

5.7: The Importance of Bamber's theorem

The equivalence theorem is the starting point for all non-parametric methods of analyzing ROC plots, e.g., Refs. 4,5. 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[6] 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.

5.8: 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[5]. Bamber's theorem, namely the equivalence between the empirical AUC and the Wilcoxon statistic has been derived and demonstrated. More recently, a first principle derivation and generalization of the variance-component modeling approach to analyzing multiple-reader multiple-case (MRMC) datasets, has been described[7,8].

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).

Details of Wilcoxon theorem

5.A.1: Upper triangle

For $i = 0$, Eqn. (5.14) implies:

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

The base of the triangle is:

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

The height of the triangle is:

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

■

5.A.2: Lowest trapezoid

For $i = 4$, Eqn. (5.14) implies:

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

The upper side of the trapezoid is

$$1 - P\left(Z_1 \geq 5\right) = P\left(Z_1 < 5\right)$$

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

$$\frac{1 + P\left(Z_1 < 5\right)}{2}$$

The height is

$$P\left(Z_2 \geq 5\right) = P\left(Z_2 = 5\right)$$

Multiplication of the last two expressions yields A_4.

■

5.9: References

1. Wilcoxon F. Individual Comparison by Ranking Methods. *Biometrics.* 1945;1:80-83.
2. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics.* 1947;18:50−60.
3. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology.* 1975;12(4):387-415.
4. Hanley JA, Hajian-Tilaki KO. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update. *Academic Radiology.* 1997;4(1):49-58.
5. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics.* 1988;44:837-845.
6. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 1982;143(1):29-36.
7. Clarkson E, Kupinski MA, Barrett HH. A Probabilistic Model for the MRMC Method, Part 1: Theoretical Development. *Academic Radiology.* 2006;13(11):1410-1421.
8. Gallas BD. One-Shot Estimate of MRMC Variance: AUC. *Academic Radiology.* 2006;13(3):353-362.