Chapter 07: Sources of variability affecting AUC

Table of contents

Online Supplementary Material

7.1: Introduction

In previous chapters, the area AUC under the ROC plot was introduced as the preferred way of summarizing performance in the ROC task, as compared to a pair of sensitivity and specificity values. It can be estimated either non-parametrically, as in **Chapter 05** or parametrically, as in **Chapter 06** and improved ways of estimating it are described in **Chapter 19** and **Chapter 20**.

Irrespective of how it is estimated, AUC is a realization of a random variable, and as such is subject to variability. It is subject to case-sampling variability, due to the finite numbers of cases comprising the dataset. As a general rule, anytime a measurement is based on a *finite* number of samples from a parent population, it is

subject to sampling variability. *This is because no finite sample is unique*: someone else conducting a similar study would, in general, obtain a different sample. Case-sampling variability is estimated using the binormal model in the previous chapter. It is related to the sharpness of the peak of the likelihood function, §6.4.4. The sharper that the peak is, the smaller are the corresponding variances. This chapter focuses on general sources of variability affecting AUC, regardless of how it is estimated, and other (i.e., not binormal model based) ways of estimating it.

Here is an outline of this chapter. The starting point is the identification of different sources of variability affecting AUC estimates. Considered next is dependence of AUC, however estimated, on a case-set index $\{c\}$, $c = 1,2,...,C$, which is largely suppressed in the literature. This can lead to confusion among those with less statistical expertise. Considered next is estimating case-sampling variability of the empirical estimate of AUC by an analytic method. This is followed by descriptions of two resampling-based methods, namely the bootstrap and the jackknife, both of which have wide applicability (i.e., they are not restricted to ROC analysis). The methods are demonstrated using R and the implementation of a calibrated simulator is shown and used to demonstrate their validity, i.e., showing that the different methods of estimating variability agree. The dependence of AUC on reader expertise and modality is considered. An important source of variability, namely the radiologist's choice of internal sensory thresholds, is described. A cautionary comment is made regarding indiscriminate usage of empirical AUC as a measure of performance.

Online Appendix 7.A describes R implementation of the bootstrap method; Online Appendix 7.B is the corresponding implementation of the jackknife method. Online Appendix 7.C describes implementation of the calibrated simulator for single-modality single-reader ROC datasets. Online Appendix 7.D describes code that allows comparison of the different methods of estimating case-sampling variability.

7.2: Three sources of variability

Statistics deals with variability. Understanding sources of variability affecting AUC is critical to an appreciation of ROC analysis. The author's introduction to this subject was through Swets and Pickett's book[1] "Evaluation of Diagnostic System: Methods from Signal Detection Theory". Three sources of variability are identified in that book: *case sampling*, *between-reader* and *within-reader* variability.

1. *Consider a single reader interpreting different case samples*. Case-sampling variability arises from the *finite* number of cases comprising the dataset, compared to the potentially very large *population* of cases. [If one

could sample every case there exists and have them interpreted by the same reader, there would be no case-sampling variability and the poor reader's AUC values (from repeated interpretations of the entire population) would reflect only within reader variability, see #3 below.] Each case-set $\{c\}$, consisting of K_1 non-diseased and K_2 diseased cases, interpreted by the reader yields an AUC value. The notation $\{c\}$ does not mean single cases, rather different *case sets*. Thus $\{c\} = \{1\}, \{2\}$, etc., denote different case sets, each consisting of K_1 non-diseased and K_2 diseased cases.

There is much "data compression" in going from individual case ratings to AUC. For a single reader and given case-set $\{c\}$, the ratings can be converted to an $A_{z\{c\}}$ estimate, Eqn. **(6.49)**. This makes explicit the dependence of the measure on the case-set $\{c\}$. One can conceptualize the distribution of $A_{z\{c\}}$'s over different case-sets, each of the same size $K_1 + K_2$, as a normal distribution, i.e., $A_{z\{c\}} \sim N\left(A_{z\{\bullet\}}, \sigma^2_{cs+wr}\right)$. The dot notation $\{\bullet\}$ denotes an average over all case sets. Thus, $A_{z\{\bullet\}}$ is the *case-sampling mean* of A_z for the single fixed reader and σ^2_{cs+wr} is the *case sampling plus within-reader variance*. The reason for adding the within-reader variance is explained in #3 below. The concept is that a given reader interpreting *different case-sets* effectively samples different parts of the population of cases, resulting in variability in measured A_z. Sometimes easier cases are sampled, and sometimes more difficult ones, etc. This source of variability is expected to decrease with increasing case-set size, i.e., increasing $K_1 + K_2$, which is the reason for seeking large numbers of cases in clinical trials. Case-sampling (and within-reader) variability also decreases the cases become more *homogenous*. An example of a more homogenous case sample would be cases originating from a small geographical region with, for example, limited ethnic variability. This is the reason for seeking multi-institutional clinical trials, because they tend to sample more of the population than patients seen at a single institution.

2. *Consider different readers interpreting a fixed case sample*. Between-reader variability arises from the *finite* number of readers compared to the *population* of readers; the population of readers could be all MQSA certified radiologists interpreting screening mammograms in the US. This time one envisages *different* readers interpreting a fixed case set $\{1\}$. The different reader's $A_{z;j}$ values (j is the reader index, j = 1, 2, ...,

J, where *J* is the total number of readers in the dataset) are distributed $A_{z;j} \sim N\left(\overline{A_{z;\bullet\{1\}}}, \sigma^2_{br+wr}\right)$, where $\overline{A_{z;\bullet\{1\}}}$ is the reader population mean (the dot symbol replacing the reader index, averages over a set of readers, and the grand average, indicated by the bar symbol, obtains the population mean) for the fixed case-set $\{1\}$ and σ^2_{br+wr} is the between-reader plus within-reader variance. The reason for adding the within-reader variance is explained in #3 below The concept is that *different readers interpret the same case set* $\{1\}$, thereby sampling different parts of the reader distribution, causing fluctuations in the measured $A_{z;j\{1\}}$ of the readers. Sometimes better readers are sampled and sometimes not so good ones are sampled. This time there is no "data compression" – each reader in the sample has an associated $A_{z;j}$. However, variability of the *average* $A_{z;j}$ over the *J* readers is expected to decrease with increasing *J*. This is the reason for seeking large reader-samples.

3. *Consider a fixed reader, e.g., j = 1, interpreting a fixed case-sample* $\{1\}$. Within-reader variability is due to variability of the ratings for the same case: the same reader interpreting the same case on different occasions will give different ratings to it, causing fluctuations in the measured AUC. This assumes that memory effects are minimized, for example, by sufficient time between successive interpretations. Since this is an intrinsic source of variability (analogous to the internal noise of a voltmeter) affecting each reader's interpretations, *it cannot be separated from case sampling variability*, i.e., it cannot be "turned off". The last sentence needs further explanation. A measurement of case-sampling variability requires a reader, and the reader comes with an intrinsic source of variability that gets added to the case-sampling variance, so what is measured is the sum of case sampling and within-reader variances, denoted σ^2_{cs+wr}. Likewise, a measurement of between-reader variability requires a fixed case-set interpreted by different readers, each of whom comes with an intrinsic source of variability that gets added to the between-reader variance, denoted σ^2_{br+wr}. To emphasize this point, an estimate of case-sampling variability *always* includes within reader variability, even if the notation does not show this explicitly. Likewise, an estimate of between-reader variability *always* includes within-reader variability, even if the notation does not show this explicitly.

With this background, the purpose of this chapter is to delve into variability in some detail and in particular describe computational methods for estimating them. This chapter introduces the concept of resampling a dataset to estimate variability and the widely used bootstrap and jackknife methods of estimating variance are

described. In a later chapter, these are extended to estimating covariance (essentially a scaled version of the correlation) between two random variables.

The starting point is the simplest scenario: a single reader interpreting a case-set.

7.3: Dependence of AUC on the case sample

Suppose a researcher conducts an ROC study with a single reader. The researcher starts by selecting a case-sample, i.e., a set of proven-truth non-diseased and diseased cases. Another researcher conducting an ROC study at the same institution selects a different case-sample, i.e., a different set of proven-truth non-diseased and diseased cases. The two case-sets contain the same numbers K_1, K_2 of non-diseased and diseased cases, respectively. Even if the same radiologist interprets the two case-sets, and the reader is perfectly reproducible, the AUC values are expected to be different. Therefore, AUC must depend on a *case sample index*, which is denoted $\{c\}$, where c is an integer: c = 1, 2, ..., C, etc.

$$AUC \to AUC_{\{c\}} \quad . \qquad \textbf{(7.1)}$$

Note that $\{c\}$ is not an *individual case* index, rather it is a *case-set* index, i.e., different integer values of c denote different *sets*, or *samples*, *or groups, or collections of cases*. The dependence of AUC on the case sample index is not explicitly shown in the literature.

What does the dependence of AUC on the c index mean? Different case samples differ in their *difficulty* levels. A difficult case set contains a greater fraction of difficult cases than is usual. A difficult diseased case is one where disease is difficult to detect. For example, the lesions could be partly obscured by overlapping normal structures in the patient anatomy; i.e., the lesion does not "stick out". Alternatively, variants of normal anatomy could mimic a lesion, like a blood vessel viewed end on in a chest radiograph, causing the radiologist to miss the real lesion(s) and mistake these blood vessels for lesions. An easy diseased case is one where the disease is easy to detect. For example, the lesion is projected over smooth background tissue, because of which it "sticks out", or is more conspicuous[2]. How does difficulty level affect non-diseased cases? A difficult non-diseased case is one where variants of normal anatomy mimic actual lesions and could cause the radiologist to falsely diagnose the patient as diseased. Conversely, an easy non-diseased case is like a textbook illustration of normal anatomy. Every structure in it is clearly visualized and accounted for by the radiologist's knowledge of the

patient's non-diseased anatomy, and the radiologist is confident that any abnormal structure, *if present*, would be readily seen. The radiologist is unlikely to falsely diagnose the patient as diseased. Difficult cases tend to be rated in the middle of the rating scale, while easy ones tend to be rated at the ends of the rating scale.

7.3.1: Case sampling induced variability of AUC

An easy case sample will cause AUC to increase over its average value; interpreting many case-sets and averaging the AUCs determines the average value. Conversely, a difficult case sample will cause AUC to decrease. Case sampling variability causes variability in the measured AUC. How does one estimate this essential source of variability? One method, totally impractical in the clinic but easy with simulations, is to have the same radiologist interpret repeated samples of case-sets from the population of cases (i.e., patients), termed *population sampling*, or more viscerally, as the "brute force" method.

Even if one could get a radiologist to interpret different case-sets, it is even more impractical to actually acquire the different case samples of truth-proven cases. Patients do not come conveniently labeled as non-diseased or diseased. Rather, one needs to follow-up on the patients, perhaps do other imaging tests, in order to establish true disease status, or *ground-truth*. In screening mammography, a woman who continues to be diagnosed as non-diseased on successive yearly screening tests in the US, and has no other symptoms of breast disease, is probably disease-free. Likewise, a woman diagnosed as diseased and the diagnosis is confirmed by biopsy (i.e., the biopsy comes back showing a malignancy in the sampled tissues) is known to be diseased. However, not all patients who are diseased are actually diagnosed as diseased: a typical false negative fraction is 20% in screening mammography[3]. This is where follow-up imaging can help determine true disease status at the initial screen. A false negative mistake is unlikely to be repeated at the next screen. After a year, the tumor may have grown, and is more likely to be detected. Having detected the tumor in the most recent screen, radiologists can go back and retrospectively view it in the initial screen, at which it was missed during the "live" interpretation. If one knows *where* to look, the cancer is easier to see. The previous screen images would be an example of a difficult diseased case. In unfortunate instances, the patient may die from the previously undetected cancer, which would establish the truth status at the initial screen, too late to do the patient any good. The process of determining actual truth is often referred to as defining the "gold standard", the "ground truth: or simply "*truthing*". *One can appreciate from this discussion that acquiring independently proven cases, particularly diseased ones, is one of the most difficult aspects of conducting an observer performance study.*

There has to be a better way of estimating case-sampling variability. With a parametric model, the maximum likelihood procedure provides a means of estimating variability of each of the estimated parameters, which can

be used to estimate the variability of A_z, as in **Chapter 06**. The estimate corresponds to case-sampling variability (including an inseparable within-reader variability). If unsure about this point, the reader should run some of the examples in **Chapter 06** with increased numbers of cases. The variability will be seen to decrease.

There are other options available for estimating case-sampling variance of AUC, and this chapter is not intended to be comprehensive. Three commonly used options are described: the DeLong et al method, the bootstrap and the jackknife resampling methods.

7.4: Estimating case-sampling variability using the DeLong method

If the figure-of-merit is the empirical AUC, then a procedure developed by DeLong et al[4] (henceforth abbreviated to DeLong) is applicable that is based on earlier work by Noether[5] and Bamber[6]. The author will not go into details of this procedure (implemented in DeLongVar.R) but limit to showing that it "works". The reader may wish to compare the R implementation with the original publication. However, before one can show that it "works", one needs to know the true value of the variance of empirical AUC. Even if data were simulated using the binormal model, one cannot use the binormal model maximum likelihood estimation (MLE) estimate of variance as it is an estimate, not to be confused with a true value. Estimates are realizations of random numbers and are themselves subject to variability, which decreases with increasing case-set size. Instead, a "brute-force" (i.e., simulated population sampling) approach is adopted to determine the true value of variance of AUC. The simulator provides a means of repeatedly generating case-sets interpreted by the same radiologist, and by sampling it enough time, e.g., $C = 10{,}000$ times, each time calculating AUC, one determines the population mean and standard deviation. The standard deviation determined this way is compared to that yielded by the DeLong method to check if the latter actually works. Open the file **`mainDeLongSd.R`**; a listing follows:

7.4.1: Code Listing

```
rm(list = ls()) # mainDeLongSd.R
source("Wilcoxon.R");source("DeLongVar.R")

seed <- 1;set.seed(seed)
mu <- 1.5;sigma <- 1.3;K1 <- 50;K2 <- 52
cat("seed = ", seed, ", K1 = ", K1, ", K2 = ", K2, ", mu = ", mu, ", sigma = ", sigma, "\n")

# brute force method to find the population mean and stdDev. dev.
empAuc <- array(dim = 10000)
for (i in 1:length(empAuc)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu, sd = sigma)
  empAuc[i] <- Wilcoxon(zk1, zk2)
}
stdDevempAuc  <-  sqrt(var(empAuc))
meanempAuc   <-  mean(empAuc)
cat("population mean empAuc = ", meanempAuc,
    ", population stdDev empAuc = ", stdDevempAuc, "\n")
```

```
# one more trial
zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu, sd = sigma)
empAuc <- Wilcoxon(zk1, zk2)
ret  <- DeLongVar(zk1,zk2)
stdDevDeLong <- sqrt(ret)
cat("1 sample empAuc = ", empAuc,
    ", stdDev DeLong = ", stdDevDeLong, "\n")
```

Line 2 sources the functions needed for this code to work: one calculates the Wilcoxon statistic and the other implements the DeLong method. Line 3 sets the **seed** of the random number generator to 1. The **seed** variable is completely analogous to the case-set index *c*. Keeping **seed** fixed realizes the same random numbers each time the program is run. Different values of **seed** result in different, i.e., statistically independent, random samples. Lines 45 – 6 initialize the values needed by the data simulator: the normal distributions are separated by **mu** = 1.5, the standard deviation of the diseased distribution is **sigma** = 1.3, and there are **K1** = 50 non-diseased and **K2** = 52 diseased cases. Lines 8 -17 implement the "brute force" method of estimating mean and standard deviation of the population distribution of *AUC* and prints the values. The actual data simulation occurs at line 11: the ratings vectors are **zk1** and **zk2**, corresponding to non-diseased and disease cases, respectively. For simplicity, data binning is not employed. Line 12 calculates empirical AUC, using function **Wilcoxon()**, and saves it to the array **empAUC**. Lines 14-15 calculate the mean and standard deviation of the AUC samples: the latter is the "correct" value to which the DeLong standard deviation estimate will be compared. Line 20 generates a fresh ROC dataset to which the DeLong method will be applied. Line 21 calculates the new value of the empirical area *AUC* for this dataset and line 22 applies the DeLong method, which returns the variance of the empirical estimate of AUC, whose square root is the standard deviation. Two runs of this code were made, one with the smaller sample size, and the other with 10 times the sample size (the second run takes much longer). A third run was made with the larger sample size but with a different **seed** value. The results follow:

7.4.2: Code Output

```
> source('~/Desktop/book3/02 A ROC paradigm/A7 Sources of variability in
AUC/software/mainDeLongSd.R')
seed =  1 , K1 =  50 , K2 =  52 , mu =  1.5 , sigma =  1.3
population mean empAuc =  0.819178 , population stdDev empAuc =  0.04176683
1 sample empAuc =  0.8626923 , stdDev DeLong =  0.03804135
> source('~/Desktop/book3/02 A ROC paradigm/A7 Sources of variability in
AUC/software/mainDeLongSd.R')
seed =  1 , K1 =  500 , K2 =  520 , mu =  1.5 , sigma =  1.3
population mean empAuc =  0.8194576 , population stdDev empAuc =  0.01309815
1 sample empAuc =  0.8206962 , stdDev DeLong =  0.01309314
> source('~/Desktop/book3/02 A ROC paradigm/A7 Sources of variability in
AUC/software/mainDeLongSd.R')
seed =  2 , K1 =  500 , K2 =  520 , mu =  1.5 , sigma =  1.3
population mean empAuc =  0.8194988 , population stdDev empAuc =  0.01300203
1 sample empAuc =  0.8047269 , stdDev DeLong =  0.01356696
```

1. An important observation is that as sample-size increases, case-sampling variability decreases: 0.0417 for the smaller sample size vs. 0.01309 for the larger sample size, and the dependence is as the inverse square root of the numbers of cases, e.g., 0.04176683/sqrt(10) = 0.01320783. This is as expected from the central limit theorem[7].
2. With the smaller sample size (K1/K2 = 50/52; the back-slash notation, not to be confused with division, is a convenient way of summarizing the case-sample size) the estimated standard deviation (0.038) is within 10% of that estimated by population sampling (0.042). With the larger sample size, (K1/K2 = 500/520) the two are practically identical (0.013093 vs. 0.01356696 – the latter value is for seed = 2).
3. Notice also that the one sample empirical AUC for the smaller case-size is 0.863, which is less than two standard deviations from the population mean 0.819. The "two standard deviations" comes from rounding up 1.96: as in Eqn. (3.46), where $z_{\alpha/2}$ was defined as the *upper* $\alpha/2$ quantile of the unit normal distribution and $z_{0.025} = 1.96$.
4. To reiterate, with clinical data the DeLong procedure estimates case sampling plus within reader variability. With simulated data as in this example, there is no within-reader variability as the simulator yields identical values for fixed seed.

This demonstration should convince the reader that one does have recourse other than the "brute force" method, at least when the figure of merit is the empirical area under the ROC. That should come as a relief, as population sampling is impractical in the clinical context. It should also impress the reader, as the DeLong method is able to use information present in a *single* dataset to tease out its *variability*. [This is analogous to the MLE estimate, which is also able to tease out variability based on a parametric fit to a single dataset and examination of the sharpness of the peak of the log-likelihood function, **Chapter 06**.]

Next, two resampling – based methods of estimating case-sampling variance of *AUC* are introduced. The word "resampling" means that the *dataset itself is regarded as containing information regarding its variability*, which can be extracted by sampling from the original data (hence the word "resampling"). These are general and powerful techniques, applicable to any scalar statistic, not just the empirical AUC, which one might be able to use in other contexts[8].

Exercises and proposed projects:

1. Exercise the code with different **`seed`**-values and be convinced that statements in #2 above for the smaller sample size are correct.

a. Specifically, the differences between the "brute force" standard deviation **stdDevempAuc**, the true value, and that yielded by the DeLong method **stdDevDeLong** can be accounted for by sampling variability (statistical statement: the differences are not significant at the 5% level).

b. Note: one does not need to re-establish the true value; with 10,000 samples, variations in **seed** are not expected to alter the "true" values (Try it! The differences are < 1%). One puts lines 20 – 25 inside a **for**-loop, and for iteration of the for-loop, one saves the value at line 23 in a suitably initialized array **stdDevDeLong**. Finally, one then compares the empirical 95% confidence interval for **stdDevDeLong** to the true value **stdDevempAuc**.

2. Extend the simulation model to include the effect of binning; the reader may wish to see examples of how binning is easily accomplished by a function in the **RJafroc** package before returning to this. Try binning the data into five or six bins.

3. Include a model for within-reader variability in the simulation model. Does the DeLong method indeed estimate case sampling plus within-reader variability? Hint: this is a modification of (6.2.1) to account for replications; r is the replication index; $r = 1, 2, 3, ...$

$$\left.\begin{aligned} &Z_{k_t tr} \sim \mu_t + C_{k_t t} + \varepsilon_{k_t tr} \\ &t = 1,2; \mu_1 = 0; \mu_2 = \mu \\ &C_{k_1 1} \sim N(0,1); C_{k_2 2} \sim N\left(0,\sigma_{cs}^2\right); \\ &\varepsilon_{k_t tr} \sim N\left(0,\sigma_{wr}^2\right) \end{aligned}\right\} \quad . \qquad \textbf{(7.2)}$$

The variances in Eqn. (7.2) are for the z-sample model, not to be confused with those introduced in the Introduction, which denote AUC variances. In the design of the simulator, the term σ_{wr}^2 represents the ratings variability of a case repeatedly interpreted by the same observer. The term σ_{cs}^2 represents case-sampling variability, and to correspond to the code in **mainDeLongSd.R** it should be set to **sigma**$^2 = (1.3)^2$.

7.5: Estimating case-sampling variability of AUC using the bootstrap

The simplest resampling method, at least at the conceptual level, is the bootstrap. *The bootstrap method is based on the assumption that it is safe to regard the observed sample as defining the population from which it was sampled*. Since by definition a population cannot be exhausted, the idea is to resample, *with replacement*, from the observed sample. Each resampling step realizes a particular bootstrap sample *set* denoted $\{b\}$, where $b = 1, 2, ..., B$. The curly brackets emphasize that different integer values of b denote different *sets* of cases, not

individual cases. [In contrast, the notation (*k*) will be used to denote *removing* a specific case, *k*, as in the jackknife procedure to be described shortly. The index *b* should not be confused with the index *c*, the case sampling index; the latter denotes repeated sampling from the *population*, which is impractical in real life; the bootstrap index denotes repeated sampling from the *dataset*, which is quite feasible.] The procedure is repeated *B* times, typically *B* can be as small as 200, but to be safe the author generally use about 1000 - 2000 bootstraps. The following example uses Table 4.1 from **Chapter 04** reproduced below, Table 7.1.

Table 7.1: A typical ROC counts table, showing the original data; AUC = 0.870.

	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

For convenience, let us denote cases as follows. The 30 non-diseased cases that received the 1 rating are denoted $k_{1,1}, k_{2,1}, ..., k_{30,1}$. The second index denotes the truth state of the cases. Likewise, the 19 non-diseased cases that received the 2 rating are denoted $k_{31,1}, k_{32,1}, ..., k_{49,1}$ and so on for the remaining non-diseased cases. The 5 diseased cases that received the 1 rating are denoted $k_{1,2}, k_{2,2}, ..., k_{5,2}$, the 6 diseased cases that received the 2 rating are denoted $k_{6,2}, k_{7,2}, ..., k_{11,2}$, and so on. Let us figuratively "put" all non-diseased cases (think of each case as an index card, with the case notation and rating recorded on it) into one hat (the non-diseased hat) and all the diseased cases into another hat (the diseased hat). Next, one randomly picks one case (card) from the non-diseased hat, records it's rating, and puts the case back in the hat, so it is free to be possibly picked again. This is repeated 60 times for the non-diseased hat resulting in 60 ratings from non-diseased cases. A similar procedure is performed using the diseased hat, resulting in 50 ratings from diseased cases. The author has just described, in painful detail (one might say) the realization of the 1st bootstrap sample, denoted $\{b = 1\}$. This is used to construct the 1st bootstrap counts table, Table 7.2.

Table 7.2: The counts table for the 1st bootstrap dataset; AUC = 0.843.

	Counts in ratings bins				
	Rating = 1	Rating = 2	Rating = 3	Rating = 4	Rating = 5
$K_1 = 60$	35	16	9	0	0
$K_2 = 50$	7	9	7	8	19
	Operating points				
	≥5	≥4	≥3	≥2	≥1
FPF	0.000	0.000	0.150	0.417	1
TPF	0.380	0.540	0.680	0.860	1

So what happened? Consider the 35 non-diseased cases with a 1 rating. If each non-diseased case rated 1 in Table 7.2 were picked one time, the total would have been 30, but it is 35. Therefore, some of the original non-diseased cases rated 1 must have been picked multiple times, but one must also make allowance as there is no guarantee that a specific case was picked at all. Still focusing on the 35 non-diseased cases with a 1 rating in the 1st bootstrap sample, the picked labels (reordered after the fact, with respect to the first index) might be:

$$k_{2,1}, k_{2,1}, k_{4,1}, k_{4,1}, k_{4,1}, k_{6,1}, k_{7,1}, k_{7,1}, k_{9,1}, ..., k_{28,1}, k_{28,1}, k_{30,1}, k_{30,1} \quad . \qquad \textbf{(7.3)}$$

In this example, case $k_{1,1}$ was not picked, case $k_{2,1}$ was picked twice, case $k_{3,1}$ was not picked, case $k_{4,1}$ was picked three times, case $k_{5,1}$ was not picked, case $k_{6,1}$ was picked once, etc. The total number of cases in Eqn. (7.3) is 35, and similarly for the other cells in this table. Based on the 1st bootstrapped counts table, one can estimate AUC. Using the website[9] referred to earlier, one gets $AUC = 0.843$. [It is OK to use a parametric FOM since the bootstrap is a general procedure applicable, in principle, to any *FOM*, not just the empirical *AUC*, as is the DeLong method.] The corresponding value for the original data, Table 7.1, was AUC = 0.870. The 1st bootstrapped dataset yielded a smaller value than the original dataset because one happened to have picked an unusually difficult bootstrap sample.

[Notice that in the original data there were 6 + 5 = 11 diseased cases that were rated 1 and 2, but in the bootstrapped dataset there are 7 + 9 = 16 diseased cases that were rated 1 and 2; in other words, the number of *incorrect* decisions on diseased cases went up, which would tend to lower AUC. Counteracting this effect is the increase in number of correct decisions on diseased cases: 8 + 19 = 27 cases rated 4 and 5, as compared to 12 + 22 = 34 in the original dataset. Reinforcing the effect is that increase in the number of *correct* decisions on non-diseased cases, albeit minimally: 35 + 16 = 51 rated 1 and 2 vs. 30 + 19 = 49 in the original dataset, and zero counts rated 4 and 5 in the non-diseased vs. 2 + 1 = 3 in the diseased. The complexity of following this illustrates the difficulty, in fact the futility, of correctly predicting which way performance from an examination of the two ROC counts tables – too many numbers are changing and in the above one did not even consider the

change in counts in the bin labeled 4. Hence, the need for an objective figure of merit, such as the binormal model based *AUC* or the empirical AUC.]

To complete the description of the bootstrap method, one repeats the procedure described in the preceding paragraphs $B = 200$ times, each time running the website calculator (not very practical) and the final result is B values of *AUC*, denoted:

$$AUC_{\{1\}}, AUC_{\{2\}}, ..., AUC_{\{B\}} \quad . \qquad \textbf{(7.4)}$$

where $AUC_{\{1\}}$=0.843, etc. The bootstrap estimate of the variance of AUC is defined by[8]

$$Var_{bs}\left(AUC\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(AUC_{\{b\}} - AUC_{\{\bullet\}}\right)^2 \quad . \qquad \textbf{(7.5)}$$

The right hand side is the traditional definition of (unbiased) variance. *The dot represents the average over the replaced index*. Of course, running the website code 200 times and recording the outputs is not a productive use of time. The following code implements two methods for estimating AUC, the binormal model estimate of AUC, described in **Chapter 06**, and the empirical AUC, described in **Chapter 05**.

7.5.1: Demonstration of the bootstrap method

Open the project file for this chapter and the file **`mainBootstrapSd.R`**, Online Appendix 7.A. Make sure the seed variable at line 12 is initialized to 1 and source the code, yielding the following output. Think of the seed variable as the case sample index $\{c\}$. Selecting a different seed generates a different case sample. Since the bootstrap method is applicable to any scalar figure of merit, two options are provided in the code, lines 10 - 11; currently FOM <- "Az", which uses the binormal model estimate, but if one reverses the commenting the empirical *AUC* is used. Source the code file. In about 2 seconds on the author's computer yielded the following output:

7.5.1.1: Code Output for seed = 1

```
> source('~/book2/02 A ROC analysis/A7 Sources of variability in AUC/software/mainBootstrapSd.R')
FOM =  Az , seed =  1 , B =  200
OrigAUC =  0.8704519 , meanAUC =  0.8671713 , stdAUC =  0.04380523
```

This shows that the AUC of the original data (i.e., before bootstrapping) is 0.870, the mean AUC of the $B = 200$ bootstrapped datasets is 0.867, and the standard deviation of the 200 bootstraps is 0.0438. Now if one runs the website calculator referenced in the previous chapter on the dataset shown in Table 7.1, one finds that the MLE of the standard deviation of the AUC of the fitted ROC curve is 0.0378. The standard deviation is itself a statistic and there is sampling variability associated with it, i.e., there exists such a beast as a standard deviation of a standard deviation; the bootstrap estimate is near the MLE estimate.

By setting seed to different values, one gets an idea of the variability in the estimate of the standard deviation of AUC (to repeat, seed is like the case sample index $\{c\}$; different values correspond to different case sets). For example, with seed <- 2, one gets:

7.5.1.2: Code Output for seed = 2

```
> source('~/book2/02 A ROC analysis/A7 Sources of variability in AUC/software/mainBootstrapSd.R')
FOM =  Az , seed =  2 , B =  200
OrigAUC =  0.8704519 , meanAUC =  0.8673155 , stdAUC =  0.03815402
```

Note that both the mean of the bootstrap samples and the standard deviation have changed, but both are close to the MLE values. One should experiment with other values of seed. Examined next is the dependence of the estimates on B, the number of bootstraps. With seed <- 1 and B <- 2000 one gets:

7.5.1.3: Code Output for B = 2000

```
> source('~/book2/02 A ROC analysis/A7 Sources of variability in AUC/software/mainBootstrapSd.R')
#boots =  2000 seed =  1 OrigAUC =  0.8704519 meanAUC =  0.8674622 stdAUC =  0.03833508
```

The estimates are evidently rather insensitive to B, but the computation time was longer, ~13 seconds (running MLE 2000 times in 13 seconds is not bad). It is always a good idea to test the stability of the results to different B and seed values. Unlike the DeLong method, the bootstrap is broadly applicable to other figures of merit; specifically, it is not limited to the empirical area under the ROC. However, do beware that it depends on the assumption that the sample itself is representative of the population. With limited numbers of cases, this could be a bad assumption. With small sample sizes, it is relatively easy to enumerate the different outcomes of the sampling process and, more importantly, their respective probabilities, leading to what is termed the *exact* bootstrap. It is exact in the sense that there is no seed variable or number of bootstrap dependence.

7.6: Estimating case-sampling variability of AUC using the jackknife

Attention now turns to the second resampling method, termed the *jackknife*, which is computationally less demanding, but as was seen with the bootstrap, with modern personal computers computational limitations are no longer that important, at least for the types of analyses that this book is concerned with.

In this method, the first case is removed, or *jackknifed*, from the set of cases and the MLE (or empirical estimation) is conducted on the resulting dataset, which has one less case. Let us denote by $AUC_{(1)}$ the resulting value of ROC-AUC. The parentheses around the subscript 1 are meant to emphasize that the AUC value corresponds to that with the first case *removed* from the original dataset. Next, the first case is replaced, and now the second case is removed, the new dataset is analyzed yielding $AUC_{(2)}$, and so on, yielding K (K is the total number of cases; $K = K_1 + K_2$) *jackknife AUC values*:

$$AUC_{(k)}; k = 1, 2, ..., K \quad . \qquad \textbf{(7.6)}$$

The corresponding *jackknife pseudovalues* Y_k are defined by:

$$Y_k = K \bullet AUC - \left(K - 1\right) AUC_{(k)} \quad . \qquad \textbf{(7.7)}$$

Here AUC denotes the estimate using the entire dataset, i.e., not removing any cases. The jackknife pseudovalues will turn out to be of central importance in **Chapter 09**.

The jackknife estimate of the variance is defined by[8]

$$Var_{jk}\left(AUC\right) = \frac{\left(K-1\right)^2}{K} \frac{1}{K-1} \sum_{k=1}^{K} \left(AUC_{(k)} - AUC_{(\bullet)}\right)^2 \quad . \qquad \textbf{(7.8)}$$

Since variance of K scalars is defined by:

$$Var\left(x\right) = \frac{1}{K-1} \sum_{k=1}^{K} \left(x_k - x_{(\bullet)}\right)^2 \quad . \qquad \textbf{(7.9)}$$

$$Var_{jk}\left(AUC\right)=\frac{\left(K-1\right)^2}{K}Var\left(AUC_{(k)}\right) \quad . \qquad \textbf{(7.10)}$$

In Eqn. (7.8) the author has deliberately *not* simplified the right hand side by cancelling out *K*-1. The purpose is to show, Eqn.(7.10), that the usual expression for the variance needs to be multiplied by a *variance inflation factor* $\frac{\left(K-1\right)^2}{K}$, which is approximately equal to K, in order to obtain the correct jackknife estimate of variance of *AUC*. This factor was not necessary when one used the bootstrap method. That is because the bootstrap samples are more representative of the actual spread in the data. The jackknife samples are more restricted than the bootstrap samples, so the spread of the data is smaller; hence the need for the variance inflation factor[8].

Source the file **mainJackknifeSd.R**, Online Appendix 7.B, after ensuring that **Az** is selected as the **FOM**, yielding the following results:

7.6.1: Code Output

```
> source('~/book2/02 A ROC analysis/A7 Sources of variability in AUC/software/mainJackknifeSd.R')
FOM =  Az
OrigAUC =  0.8704519 jackknifeMeanAuc =  0.8704304 stdAUC =  0.03861591
```

Notice that the code does not use a set.seed() statement, as no random number generator is needed in the jackknife method (systematically removing and replacing each case in sequence, one at a time, is not random sampling, which should further explain the need for the variance inflation factor in Eqn. (7.10)). The bootstrap and jackknife methods are broadly applicable to other figures of merit; specifically, they are not limited to the empirical area under the ROC.

7.7: Estimating case-sampling variability of AUC using a calibrated simulator

In real life one does have the luxury of sampling from the population of cases, but with a simulator almost anything is possible. The population sampling method used previously, §7.4.2, to compare the DeLong method to a known standard used *arbitrarily* set simulator values (mu = 1.5 and sigma = 1.3 at line 5 of **mainDeLongSd.R**). One does not know if these values are actually representative of real clinical data. In this section a simple method of implementing population sampling using a *calibrated* simulator is described. The code is in **mainCalSimulator.R** in Online Appendix 7.C. A *calibrated simulator* is one whose parameters are chosen to match those of an actual clinical dataset, so the simulator is calibrated to the specific one-and-only-one dataset. Why might one wish to do that? Rather than "fish in the dark" and set arbitrary values for the

simulator parameters, one needs to find realistic values that match an actual clinical dataset. This way one has at least some assurance that the simulator is realistic and therefore its verdict on a proposed method or analysis is more likely to be correct.

As an example, consider a real clinical dataset, such as in Table 7.1. This data set analyzed by MLE yielded model parameters, *a*, *b* and the 4 thresholds $\zeta_1, \zeta_2, \zeta_3, \zeta_4$. The specific values were (in the same order): 1.320453, 0.607497, 0.007675259, 0.8962713, 1.515645 and 2.39671 (listed in 7.7.1: Code Output below). On each pass through the simulator one samples 60 values from the non-diseased distribution and 50 values from the diseased distribution, implemented in SimulateRocCountsTable.R, Online Appendix 7.C, which returns a simulated ROC counts table like Table 7.1. MLE on the ROC counts table yields A_z. The process is repeated $P = 2000$ (p is the population sampling index, ranging from 1 to P) and finally one calculates the mean and standard deviation of the 2000 A_z values. Open the file **mainCalSimulator.R**, Online Appendix 7.C, confirm that FOM is set to "Az" at line 11 (i.e., it is not commented out) and source it. Shown are results of two runs with different values for the seed (namely, 1 and 2):

7.7.1: Code Output

```
> source('~/book2/02 A ROC analysis/A7 Sources of variability in AUC/software/mainCalSimulator.R')
seed =  1 , FOM =  Az , P =  2000
Calibrated simulator values: a, b, zetas:
 1.320453 0.607497 0.007675259 0.8962713 1.515645 2.39671
seed =  1 OrigAUC =  0.8704519 meanAUC =  0.8676727 stdAUC =  0.04033307

> source('~/book2/02 A ROC analysis/A7 Sources of variability in AUC/software/mainCalSimulator.R')
seed =  2 , FOM =  Az , P =  2000
Calibrated simulator values: a, b, zetas:
 1.320453 0.607497 0.007675259 0.8962713 1.515645 2.39671
seed =  2 OrigAUC =  0.8704519 meanAUC =  0.8681855 stdAUC =  0.04055164
```

The seed = 1 estimate of standard deviation of AUC (0.0403) is recorded in Table 7.3, row A, sub-row "Population". The entry for sub-row "MLE" was obtained using the ROCFIT equivalent Eng's JAVA program[9], §6.2.6. The DeLong method entry for row A was obtained using **mainDeLongSd.R** with FOM set to " Wilcoxon", as indicated by the asterisk; see §7.4.2. The bootstrap entry was obtained using **mainBootstrapSd.R**, and the jackknife entry was obtained using **mainJackknifeSd.R**; in both cases FOM was set to "Az". Note that the four estimates are close to each other, around 0.04. This confirms the validity of the different approaches to estimating the case sampling standard deviation, and is a self-consistency check on the calibration process.

Row B repeats the values in row A, except that this time the empirical AUC is being used as the figure of merit. The flexibility afforded by the calibrated simulator is that using it one can test various ideas. For example, what

happens if the number of cases is increased? One expects the standard deviations in the last column of Table 7.3 to decrease, but by how much. Row B uses datasets generated by the simulator calibrated to the data in Table 7.3. Since the numbers of cases has not changed, the values are similar to those in row A. In row-C the number of cases has been inflated by a factor of 10, and the standard deviations decrease by about a factor of square root of 10. [Since rows B, C and D use empirical AUC, MLE estimates are inapplicable.]

Exercise: Use the calibrated simulator to test the effect of changing simulator parameters, particularly mu. As mu increases, the standard deviation is expected to decrease, because there is "less room" for AUC to vary, since it is constrained to be ≤1.

Table 7.3: Comparison of different estimates of the standard deviation of AUC, namely MLE, the DeLong method, bootstrap, jackknife and population sampling. MLE = maximum likelihood estimate; shown are results for a real dataset (A, B) and two simulated datasets (C and D) and two choices for estimating AUC: parametric and empirical. * The entry for the DeLong method was obtained using the empirical AUC. (P = 2000) (B = 2000)

	Dataset	AUC Estimate	Var. estimation method	$\sigma(A_z)$
A	$\overrightarrow{K_1}=(30,19,8,2,1)$ $\overrightarrow{K_2}=(5,6,5,12,22)$	Parametric AUC	MLE	0.0378
			DeLong	*0.0380
			Bootstrap	0.0438
			Jackknife	0.0386
			Population	0.0403
B	$\overrightarrow{K_1}=(30,19,8,2,1)$ $\overrightarrow{K_2}=(5,6,5,12,22)$	Empirical AUC	MLE	NA
			DeLong	0.0380
			Bootstrap	0.0413
			Jackknife	0.0369
			Population	0.0369
C	Calibrated Simulator K1 = 60, K2 = 50		MLE	NA
			DeLong	0.0333
			Bootstrap	0.0366
			Jackknife	0.0335
			Population	0.0359
D	Calibrated Simulator K1 = 600, K2 = 500		MLE	NA
			DeLong	0.0113
			Bootstrap	0.0110
			Jackknife	0.0113
			Population	0.0113

7.8: Dependence of AUC on reader expertise

Suppose one conducts an ROC study with *J* readers where typically *J* is about 5 but can be as low as 3 and as high as 20 (the wide variability reflects, in the author's opinion, lack of understanding of the factors affecting the optimal choice of *J* and the related issue of statistical power). Each reader interprets the *same case sample*,

i.e., the same set of cases, but because they have different expertise levels and for other reasons (see below), the observed ROC counts tables will not be identical. The variance of the observed values is an empirical estimate of between-reader variance (including the inseparable within-reader component). Here is an example, in file **MainBetweenReaderSd.R**. This file loads the Van Dyke[10] dataset, consisting of two modalities and five readers. Source the code file to get:

7.8.1: Code Output

```
> source('~/book2/02 A ROC analysis/A7 Sources of variability in
AUC/software/mainBetweenReaderSd.R')
between-reader variance in modality 1 = 0.003082629
between-reader variance in modality 2 = 0.001304602
avg. between-reader variance in both modalities = 0.002193615
```

Notice that the between-reader (including, as always, within-reader) variance appears to be modality dependent. Determining if the difference is significant requires more analysis. For now one simply averages the two estimates.

How can one handle between-reader variability in the notation? Each reader's interpretation can be analyzed by MLE to get the corresponding AUC value. The notation for the observed AUC values is:

$$AUC_{j\{c\}}, j = 1,2...,J \quad . \qquad \textbf{(7.11)}$$

How does one conceptualize reader variability? As stated before, it is due to differences in expertise levels, but there is more to it. Since the single reader is characterized by parameters $\mu, \sigma, \zeta_2, \zeta_3, .., \zeta_{R-1}$ (R is the number of ratings bins; it is assumed that all readers employ the same number of bins, although they may employ it in different ways, i.e., the values of the thresholds may be different). While the non-diseased distribution for each reader could have mean different from 0 and variance different from unity, one can always translate it to zero and scale it to assure that the non-diseased distribution is the unit normal distribution. However, one cannot be assured that the separation and the width of the diseased distribution, and the thresholds, will not depend on the reader. Therefore, the most general way of thinking of reader variability is to put a j subscript on each of the model parameters, yielding $\mu_j, \sigma_j, \zeta_{1,j}, \zeta_{2,j}, .., \zeta_{R-1,j}$. Now the first two of these define the population ROC curve for reader j, and the corresponding AUC value is (this equation was derived in **Chapter 06**, Eqn. (6.92.24)):

$$AUC_{j\{c\}} = \Phi\left(\frac{\mu_j}{\sqrt{1+\sigma_j^2}}\right) \quad . \qquad \textbf{(7.12)}$$

All else being equal, readers with larger μ_j will perform better because they are better able to separate the non-diseased and diseased cases in z-space than their fellow readers. It is difficult and possibly misleading to try to estimate the differences directly from the observed ROC counts tables, but in general better readers will yield counts more skewed towards the low end of the rating scale on non-diseased cases and more skewed towards the high end of the rating scale for diseased cases. The ideal reader would rate all diseased cases one value (e.g. 5) and all non-diseased cases a smaller fixed value (e.g., 1, 2, 3, or 4), resulting in unit AUC, i.e., perfect performance. According to Eqn. (7.12), a reader with smaller σ_j will also perform better. As noted before, typically the σ parameter is greater than unity. *The reasons for this general finding will be discussed later, but accept the author's word for now that the best the reader can is to reduce this parameter to unity.* See Summary of **Chapter 06** for reasons for the observation that generally the variance of the diseased distribution is larger than one – it has to do with the inhomogeneity of the distribution of diseased cases and the possibility that a mixture distribution is involved. As regards thresholds, while the population based performance for a particular reader does not depend on thresholds, the thresholds determine the ROC counts table, so differences in usage of the thresholds will translate to differences in estimates of $AUC_{j\{c\}}$, but this is expected to be a smaller effect compared to the dependence on μ_j & σ_j. To summarize, variability of readers can be attributed to variability in the binormal model parameters and, to a lesser extent, to variability in adopted thresholds.

7.9: Dependence of AUC on modality

Suppose one conducts an ROC study with *j* (*j* =1,2,…*J*) readers but there are *I* (*i*=1,2,…*I*) modalities. This is frequently referred to as the multiple reader multiple case (MRMC) paradigm. Each reader interprets the *same case sample*, i.e., the same set of cases, in two or more modalities. Here is an example, in file **`MainModalityEffect.R`**. This file loads the Van Dyke dataset, consisting of two modalities and five readers. Source the code file to get:

7.9.1: Code Output

```
> source('~/book2/02 A ROC analysis/A7 Sources of variability in
AUC/software/mainModalityEffect.R')
reader-average FOM in modality 1 = 0.897037 reader-average FOM in modality 2 = 0.9408374 , effect
size, i.e., fom modality 1 minus modality 2 = -0.04380032
```

Notice that the second modality has a higher FOM. Determining if the difference is significant requires more analysis as described in **Chapter 09**. The difference between the reader-averaged FOMs is referred to as the *observed effect size*.

How does on handle modality dependence of the FOM in the notation? If *K* is the total number of cases, the total number of interpretations involved is *IJK*, each of which results in a rating. MLE analysis yields *IJ* values for AUC, one for each modality-reader combination. The appropriate notation is

$$AUC_{ij\{c\}} \quad . \qquad \textbf{(7.13)}$$

The most general way of thinking of reader and modality variability is to put *ij* subscripts on each of the model parameters, yielding $\mu_{ij}, \sigma_{ij}, \zeta_{2,ij}, \zeta_{3,ij}, .., \zeta_{R-1,ij}$. For a particular combination of modality and reader, the population ROC curve as fitted by the binormal model, yields the area under the ROC curve:

$$AUC_{ij} = \Phi\left(\frac{\mu_{ij}}{\sqrt{1+\sigma_{ij}^2}}\right) \quad . \qquad \textbf{(7.14)}$$

Given an MRMC dataset, using MLE one can estimate the parameters $\mu_{ij}, \sigma_{ij}, \zeta_{2,ij}, \zeta_{3,ij}, .., \zeta_{R-1,ij}$ for each modality-reader combination, and this could be used to design a simulator that is calibrated to the specific clinical dataset, which in turn can be used to illustrate the ideas and to test any proposed method of analyzing the data. However, the problem is more complex; the procedure needs to also account for the correlations arising from the large number of pairings inherent is such a dataset (e.g., reader 1 in modality 1 vs. reader 2 in modality 2, since both interpret a common dataset). Designing a MRMC calibrated simulator was until recently, an unsolved problem, which necessitated recent work[11] by the author and Mr. Xuetong Zhai. **Chapter 23** describes recent progress towards this end.

7.10: Effect on empirical AUC of variations in thresholds and numbers of bins

There are actually two effects. (1) The *empirical* AUC will tend to be smaller than the *true* AUC. If there are few operating points, and they are clustered together, the difference may be large, Fig. 7.1 (A).

7.10.1: Code listing

```
rm( list = ls())#mainEmpVsFit.R # freeze lines
```

```
library(RJafroc);library(ggplot2)

seed <- 10;set.seed(seed)
mu <- 2; sigma <- 1.5; cat("Population AUC = ", pnorm(mu/sqrt(1+sigma^2)), "\n")
K1 <- 500; K2 <- 500
fp <- rnorm(K1);tp <- rnorm(K2, mu, sigma)
zetas <- c(-Inf, 1.5, 2, 2.5, 3, 4, Inf)
fp1 <- as.numeric(cut(fp, zetas));tp1 <- as.numeric(cut(tp, zetas))
rocData1 <- DfToRJafrocDataset(fp1, tp1, "ROC")
plotEmp1 <- PlotEmpiricaOperatingCharacteristics(rocData1, 1, 1, lgdPos = "NULL")
print(plotEmp1$ROCPlot); empAuc1 <- GetFigureOfMerit(rocData1, fom = "Wilcoxon")
cat("Emp. AUC = ", empAuc1, "\n")
Fit1 <- FitCbmRoc(rocData1);print(Fit1$fittedPlot);print(as.numeric(Fit1$AUC))

zetas <- c(-Inf, -0.5, 0, 1, 1.5, 2, Inf)
fp2 <- as.numeric(cut(fp, zetas));tp2 <- as.numeric(cut(tp, zetas))
rocData2 <- DfToRJafrocDataset(fp2, tp2, "ROC")
plotEmp2 <- PlotEmpiricaOperatingCharacteristics(rocData2, 1, 1, lgdPos = "NULL")
print(plotEmp2$ROCPlot); empAuc2 <- GetFigureOfMerit(rocData2, fom = "Wilcoxon")
cat("Emp. AUC = ", empAuc2, "\n")
Fit2 <- FitCbmRoc(rocData2);print(Fit2$fittedPlot);print(as.numeric(Fit2$AUC))
```

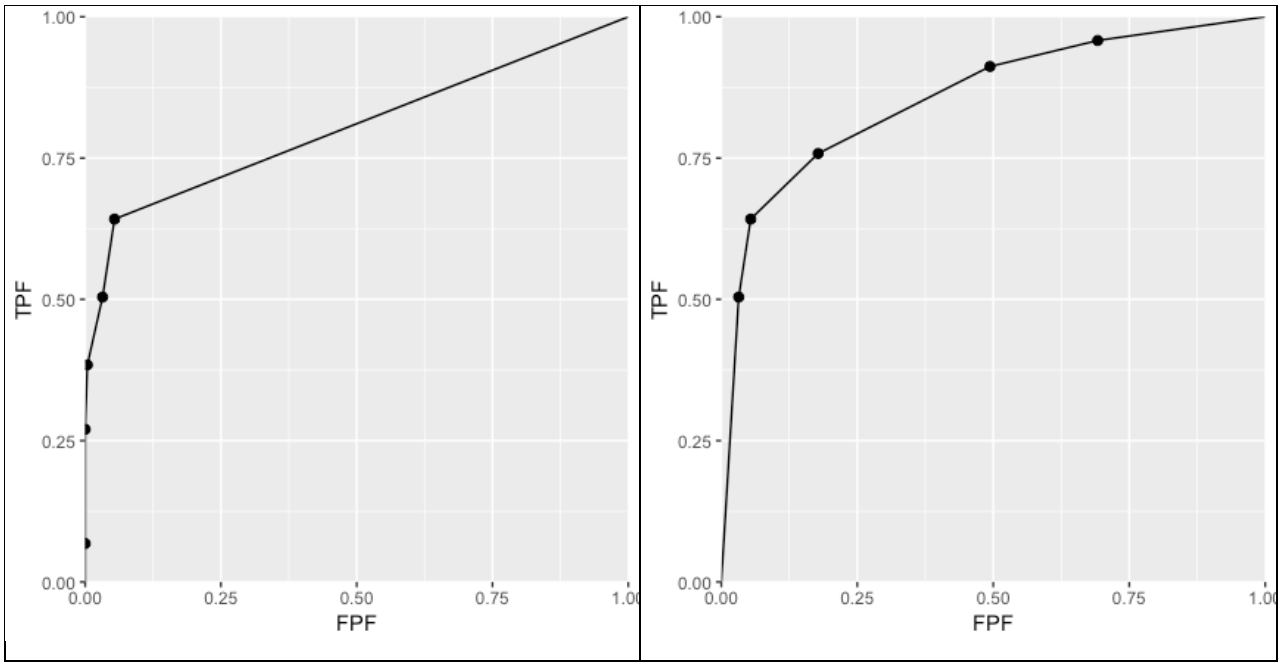
This figure was generated by a binormal model simulator, with thresholds chosen to exaggerate the effect, line 4 - 7. The true or population AUC is 0.8664, while the empirical AUC is 0.8030, line 13. However, since interest is in *differences* in AUCs, e.g., between two modalities, and the underestimates may tend to cancel, this may not be a serious issue. However, an effect that may be problematical is that the operating points for a given reader may not span the same FPF ranges in the two modalities, in which case the empirical AUCs will be different, as depicted in Fig. 7.1 (A - B). The AUC in modality (B), where the operating points span the entire range, is 0.8580, line 21, which is closer to the population value. *Since the usage of the bins is not under the researcher's control, this effect cannot be ruled out*. Fitted AUCs are expected to be less sensitive, but not immune, to this effect. Fig. 7.1 (C) is a contaminated binormal model (CBM) fitted curve, line 14, to the same data as in (A), fitted AUC = 0.892, while Fig. 7.1 (D) is a CBM fitted curve, line 22, to the same data as in (B), fitted AUC = 0.867. The difference in AUCs between (A) and (B) is -0.055, while that between (C) and (D) is 0.024. The consequences of these effects on the validity of analyses using the empirical AUC have not been studied. [The parameters of the model were $a = 1.33$ and $b = 0.667$, which yields the quoted value of the population AUC. The population value is that predicted by the parameters; it has zero sampling variability. The fitted curves are those predicted by the CBM, discussed in **Chapter 20**.]

(2) The second effect is varying numbers of thresholds or bins between the readers. One could be a radiologist, capable of maintaining at most about 6 bins, and the other an algorithmic observer, such as CAD, capable of maintaining more bins. Moreover, if the radiologist is an expert, the data points will tend to cluster near the initial near vertical part of the ROC (see **Chapter 17** for explanation). This is illustrated using code in file **`mainBinVariability.R`**. Sourcing this code yields Fig. 7.1 (E – F) and the AUC values shown in these plots.

7.10.2: Code listing

```
rm( list = ls())#mainBinVariability.R # Freeze line numbers
library(caTools);library(ggplot2);source("rocY.R")

mu <- 2;sigma <- 1.5 # you should experiment with other values
a <- mu/sigma; b <- 1/sigma #famous a and b parameters of binormal model: Dorfman&Alf
cat("true AUC = ", pnorm(mu/sqrt(1+sigma^2)), "\n")

x <- seq(0.0, 1, 0.01)

zeta <- c(3, 2.5, 2)
FPF <- pnorm(-zeta);FPF <- c(0,FPF,1)
TPF <- pnorm((mu-zeta)/sigma);TPF <- c(0,TPF,1)
pointsData <- data.frame(FPF = FPF, TPF = TPF)
AUC <- trapz(FPF,TPF)
cat("empirical AUC, sparse points = ", AUC, "\n")
rocPlot1 <- ggplot(mapping = aes(x = FPF, y = TPF)) +
  geom_line(data = pointsData) +
  geom_point(data = pointsData)

print(rocPlot1)

zeta <- seq(3, -2, -0.5)
FPF <- pnorm(-zeta);FPF <- c(0,FPF,1)
TPF <- pnorm((mu-zeta)/sigma);TPF <- c(0,TPF,1)
pointsData <- data.frame(FPF = FPF, TPF = TPF)
AUC <- trapz(FPF,TPF)
cat("empirical AUC, dense point = ", AUC, "\n")
rocPlot2 <- ggplot(mapping = aes(x = FPF, y = TPF)) +
  geom_line(data = pointsData) +
  geom_point(data = pointsData)
print(rocPlot2)
```

In Fig. 7.1(E) and Fig. 7.1(F) the effect is dramatic. The expert radiologist trapezoidal AUC is 0.7418, while that for CAD is 0.8632; the latter is close to the population value. It is left as an exercise for the reader to demonstrate that using CBM one can avoid the severe underestimate of performance that occurs in plot (E).



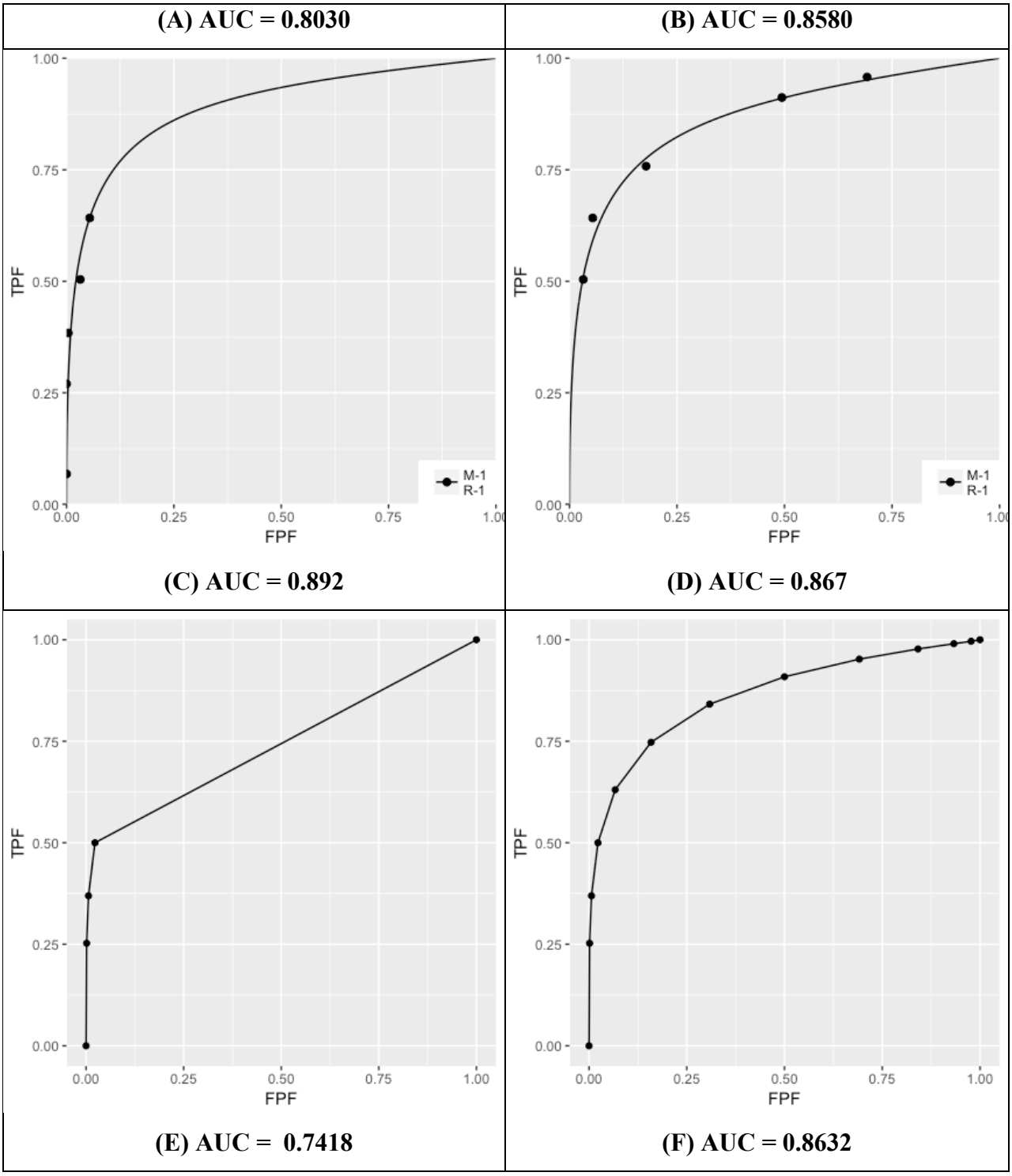


Fig. 7.1 (A-D): Plots (A - B) depict empirical plots for two simulated datasets for the same model, i.e., same continuous ratings, using different thresholds. In (A) the thresholds are clustered at low FPF values, while in (B) they are more evenly spaced. Empirical AUCs for the plots are 0.803 for (A) and 0.858 for (B). The clustering in (A) leads to a low estimate of AUC. Plots (C) and (D) are fitted curves corresponding to the same data as in (A) and (B), respectively. For each plot, the population AUC is 0.866. The fitted curves are less sensitive, but not immune, to the data clustering variations. With a large number of evenly spaced points, the empirical AUC is close to that of the fitted curve. This effect is demonstrated in plots (E) and (F). The plots were generated by **`mainEmpVsFit.R`** and **`mainBinVariability.R`**.

7.11: Empirical vs. fitted AUCs

There is a preference with some researchers to using the empirical AUC as a figure of merit. Its usage enables analyses[12-16] variously referred to as the "probabilistic", mechanistic" or "first-principles" approach and the "one-shot" approach[17] to multiple reader multiple case analysis. The author is aware of some statisticians who distrust parametric modeling and the associate normality assumptions (the author trusts that the demonstrations in §6.2.2 may assuage the concerns). In addition, empirical AUC frees the researcher from problems with binormal model based fitting, e.g., handling degenerate datasets (these problems go away with two of the fitting methods described in later chapters). The fact that the empirical AUC can always be calculated, even, for example, with a single operating point, can make the analyst blissfully unaware of anomalous data structures. In contrast, the binormal curve-fitting method in **Chapter 06** will complain when the ratings bins are not well populated, e.g., by failing to converge. This at least alerts the analyst that conditions are not optimal, and prompt data visualization and consideration of alternate fitting methods.

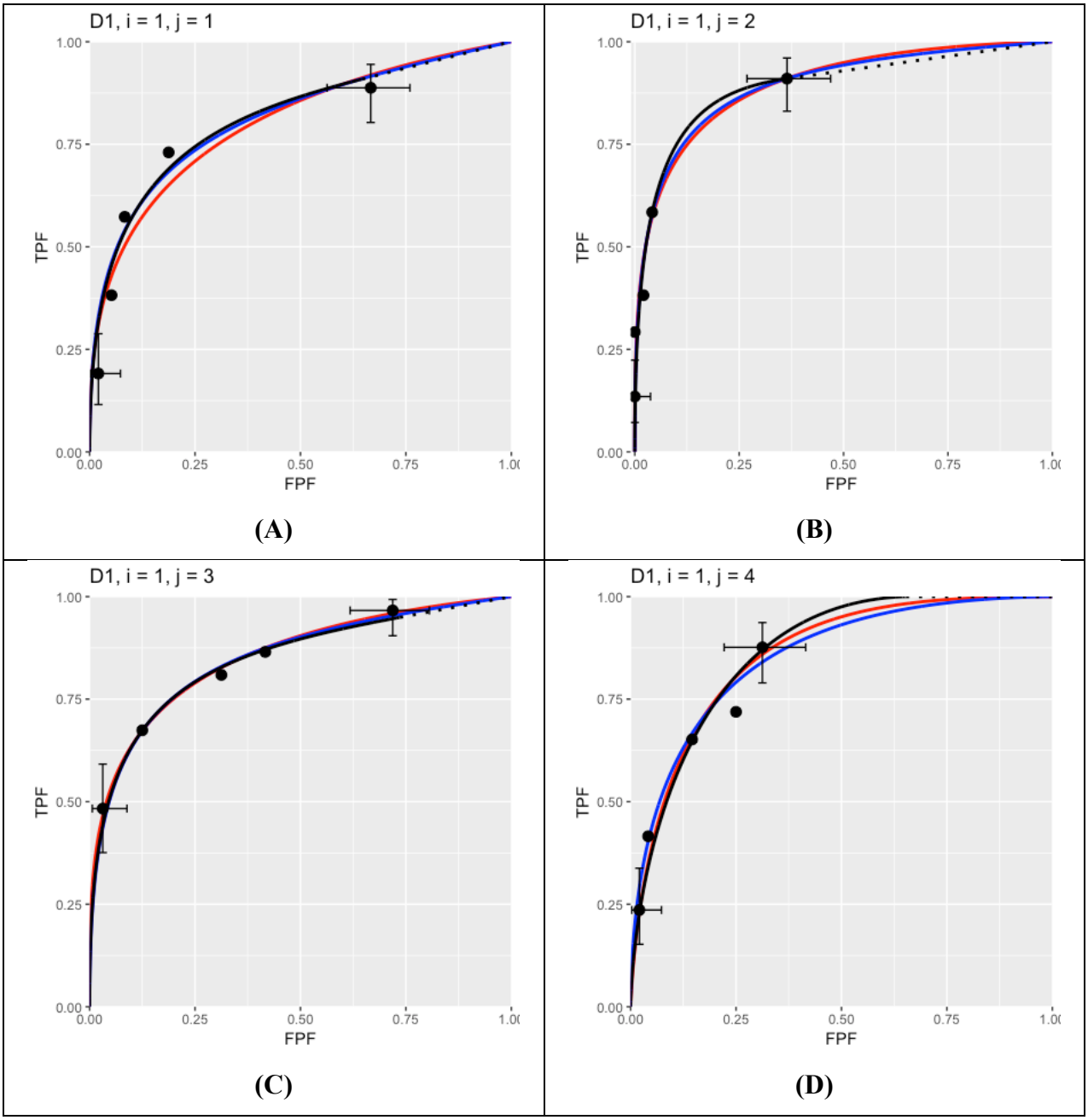
If empirical AUC is defined by a large number of operating points, such as with continuous ratings obtained with algorithmic observers, then empirical AUC will be nearly equal to the true AUC, to within sampling error. However, with human observers one rarely gets more than about 6 distinct ratings. The researcher has no control over the internal sensory thresholds used by the radiologist to bin the data, and these could depend on the modality. As demonstrated in the previous section, the empirical AUC is sensitive to the choice of thresholds, especially when the number of thresholds is small, as is usually the case with radiologists, and when the operating points are clustered on the initial near vertical section of the plot, as is also the case with experts.

Chapter 19 describes applications of three advanced methods of fitting ROC curves. The methods were applied to fourteen (14) datasets comprising 43 modalities, 80 readers and 2012 cases. The binormal model would fail on most of the datasets. The fitted data and the operating points are shown in an online file **`RSM Vs. Others.docx`** corresponding to the cited chapter. It contains 236 plots, each with three fits and operating point shown. A sampling of these plots for a single dataset, a single modality and five readers, is shown in Fig. 7.2 (A - E). One can judge from these plots whether threshold variability effect exist. The author believes they cannot be ruled out.

Over-dependence on the empirical AUC can lead to a false-sense of security regarding the validity of the analysis and avoidance of deeper issues affecting radiologist performance. The critique will become clearer when one of the newer fitting methods is described in **Chapter 19**. Given that methods, to be described, do

exist that fit any dataset, the author's advice to users is to consider using them to calculate AUC, instead of indiscriminately using empirical AUC.

A factor arguing in favor of usage of empirical AUC is that some of the newer significance testing procedures cited above, applicable to empirical AUC might be on firmer theoretical grounds than the ones to be proposed in the following chapters. As an example, they explicitly allow for truth state dependences of some of the variability components, which is reasonable. Careful simulation work, so far lacking, needs to be conducted to determine if these advantages counterbalance some the issues with empirical AUC identified in this chapter.


D1, i = 1, j = 1
TPF
FPF
(A)
D1, i = 1, j = 2
TPF
FPF
(B)
D1, i = 1, j = 3
TPF
FPF
(C)
D1, i = 1, j = 4
TPF
FPF
(D)

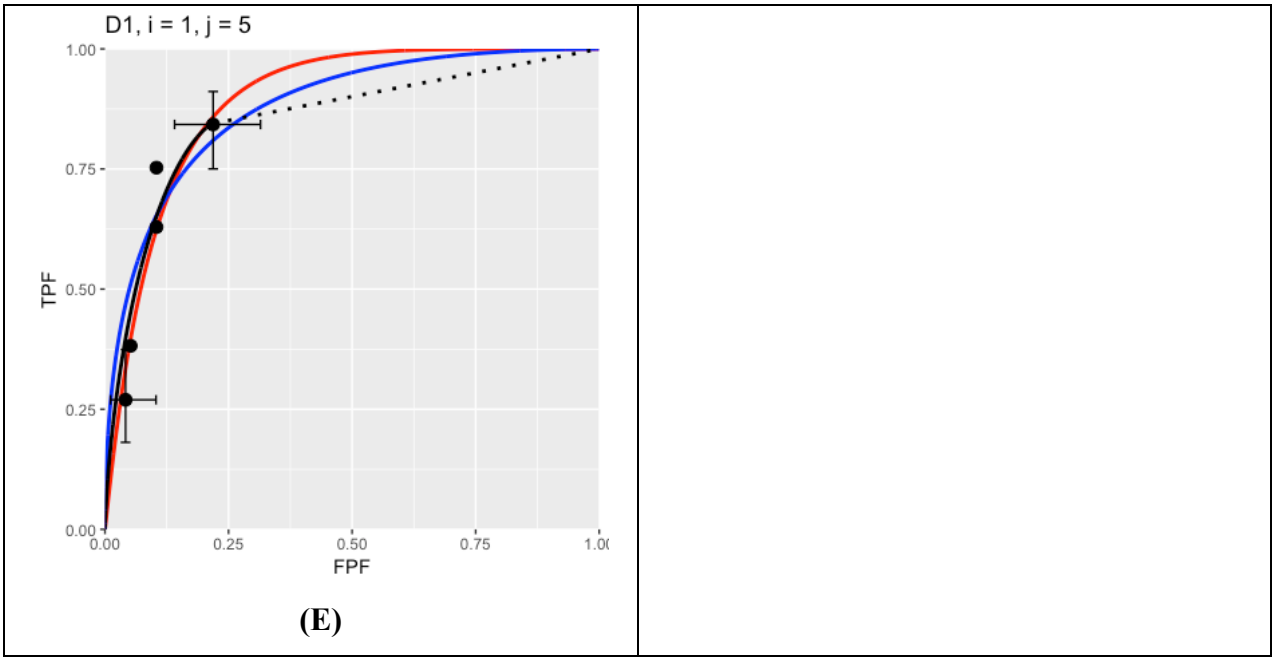


Fig. 7.2 (A-E): This figure shows a small sample of the 236 viewable plots in the cited online document. In this figure, each panel corresponds to a different reader (the *j*-index in the labels). The modality is the same (the *i*-index) and the dataset is labeled D1. The three curves correspond to different advanced method of fitting ROC data. The interest in this chapter is on the positions of the operating points. Reader (C) traverses more of the FPF range than does reader (E). Empirical AUC may result in a greater error for reader (E) than for reader (C). An explanation of the three fits is deferred to **Chapter 19**.

7.12: Discussion/Summary

This chapter focused on the factors affecting variability of AUC, namely case-sampling and between-reader variability, each of which contain an inseparable within-reader contribution. The only way to get an estimate of within-reader variability is to have the same reader re-interpret the same case-set on multiple occasions (with sufficient time delay to minimize memory effects). This is rarely done and is unnecessary, in the ROC context, to sound experimental design and analysis. Some early publications have suggested that such re-interpretations are needed to estimate the within-reader component, but modern analysis, described in the next part of the book, does not require re-interpretations. Indeed, it is a waste of precious reader-time resources. Rather than have the same readers re-interpret the same case-set on multiple occasions, it makes much more sense to recruit more readers and/or collect more cases, guided by a systematic sample size estimation method. Another reason the author is not in favor of re-interpretations is that the within-reader variance is usually smaller than case-sampling and between-reader variances. Re-interpretations would minimize a quantity that is already small, which is not good science.

In the author's judgment, current literature on this topic lacks notational clarity, particularly when it comes to case sampling. An important part of this chapter is the explicit usage of the case-set index $\{c\}$ to describe a key-factor, namely the case-sample, on which AUC depends. This index is assumed in the literature, which can lead to confusion; especially understanding one the methods used to analyze ROC MRMC data, see **Chapter 10**. Different simulated datasets correspond to different values of $\{c\}$. This indexing leads to a natural, in the author's opinion, understanding of the bootstrap method; one simply replaces $\{c\}$ with $\{b\}$, the bootstrap case-set index.

The bootstrap and jackknife methods described in this chapter have wide applicability. Later they will be extended to estimating the covariance (essentially a scaled correlation) between two random variables. Also described was the DeLong method, applicable to the empirical AUC. Using a real dataset and simulators, all methods were shown to agree with each other, especially when the numbers of cases is large, Table 7.3 (row-D).

The concept of a calibrated simulator was introduced as a way of "anchoring" a simulator to a real dataset. While relatively easy for a single dataset, the concept has yet to be extended to where it would be useful, namely designing a simulator calibrated to a dataset consisting of interpretations by multiple readers in multiple modalities of a common dataset. Just as a calibrated simulator allowed comparison of the different variance estimation methods to a known standard, obtained by population sampling, a more general calibrated simulator would allow better testing the validity of the analysis described in the next few chapters.

A source of variability not generally considered, namely threshold variability, is introduced, and a cautionary note is struck with respect to indiscriminate usage of the empirical AUC. Finally, the author wishes to reemphasize the importance of viewing ROC plots, to detect anomalous conditions that might be otherwise overlooked.

This concludes **Part A** of this book. The next chapter begins **Part B**, namely the statistical analysis of multiple-reader multiple-case (MRMC) ROC datasets.

7.13: References

1. Swets JA, Pickett RM. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* first ed. New York: Academic Press; 1982.
2. Kundel HL, Revesz G. Lesion Conspicuity, Structured Noise, and Film Reader Error. *American Journal of Roentgenology.* 1976;126:1233-1238.
3. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. *Archives of Internal Medicine.* 1996;156(2):209-213.
4. 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.
5. Noether GE. *Elements of nonparametric statistics.* Wiley & Sons;1967.
6. 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.
7. Casella G, Berger RL. *Statistical inference.* Vol 2: Duxbury Pacific Grove, CA; 2002.
8. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap.* Vol 57. Boca Raton: Chapman & Hall/CRC; 1993.
9. Eng J. ROC analysis: web-based calculator for ROC curves, http://www.jrocfit.org. 2006.
10. Van Dyke CW, White RD, Obuchowski NA, Geisinger MA, Lorig RJ, Meziane MA. Cine MRI in the diagnosis of thoracic aortic dissection. *79th RSNA Meetings.* 1993.
11. Zhai X, Chakraborty DP. A bivariate contaminated binormal model for robust fitting of proper ROC curves to a pair of correlated, possibly degenerate, ROC datasets. *Medical Physics.* 2017:n/a-n/a.
12. Gallas BD, Pennello Ga, Myers KJ. Multireader multicase variance analysis for binary data. *Journal of the Optical Society of America A, Optics, image science, and vision.* 2007;24(12):70-80.
13. Kupinski MA, Clarkson E, Barrett HH. A Probabilistic Model for the MRMC Method, Part 2: Validation and Applications. *Academic Radiology.* 2006;13(11):1422-1430.
14. Clarkson E, Kupinski MA, Barrett HH. A Probabilistic Model for the MRMC Method, Part 1: Theoretical Development. *Academic Radiology.* 2006;13(11):1410-1421.
15. Clarkson E, Kupinski MA, Barrett HH. A Model for MRMC AUC Measurements: Theory and Simulations. *SPIE Proc.* 2005.
16. Barrett HH, Kupinski MA, Clarkson E. Probabilistic foundations of the MRMC method. *Progress in Biomedical Optics and Imaging - Proceedings of SPIE.* Vol 57492005:21-31.

17. Gallas BD. One-Shot Estimate of MRMC Variance: AUC. *Academic Radiology.* 2006;13(3):353-362.