The RJafroc Roc Book

Dev P. Chakraborty, PhD

2022-01-04

Contents

Preface

- It is intended as an online update to my "physical" book (Chakraborty, 2017).
- Since its publication in 2017 the `RJafroc` package, on which the `R` code examples in the book depend, has evolved considerably, causing many of the examples to "break".
- This also gives me the opportunity to improve on the book and include additional material.

TBA How much finished

- HMF approximately 70%
- This book is currently (as of December 2021) in preparation.
- Parts labeled TBA and TODOLAST need to be updated on final revision.

The pdf file of the book

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

The html version of the book

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

A note on the online distribution mechanism of the book

- In the hard-copy version of my book (Chakraborty, 2017) the online distribution mechanism was `BitBucket`.

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

Structure of the book

The book is divided into parts as follows:

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

Contributing to this book

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

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

- Diagnostic imaging system evaluation
- Detection
- Detection combined with localization
- Detection combined with localization and classification
- Optimization of Artificial Intelligence (AI) algorithms

- CV
- Alternatives

ToDos TBA

- Check Bamber theorem derivation.

Chapters needing heavy edits

TBA

Shelved vs. removed vs. parked folders needing heavy edits

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

Coding aids

- sprintf("%.4f", proper formatting of numbers
- OpPtStr(, do:
- kbl(dfA, caption = "...", booktabs = TRUE, escape = FALSE) %>% collapse_rows(columns = c(1, 3), valign = "middle") %>% kable_styling(latex_options = c("basic", "scale_down", "HOLD_position"), row_label_position = "c")
- "'{r, attr.source = ".numberLines"}
- kbl(x12, caption = "Summary of optimization results using wAFROC-AUC.", booktabs = TRUE, escape = FALSE) %>% collapse_rows(columns = c(1), valign = "middle") %>% kable_styling(latex_options = c("basic", "scale_down", "HOLD_position"), row_label_position = "c")
- $\exp(-\lambda')$ space before dollar sign generates a pdf error
- FP errors generated by GitHub actions due to undefined labels: Error: Error: pandoc version 1.12.3 or higher is required and was not found (see the help page ?rmarkdown::pandoc_available). In addition: Warning

message: In verify_rstudio_version() : Please install or upgrade Pandoc to at least version 1.17.2; or if you are using RStudio, you can just install RStudio 1.0+. Execution halted

ROC paradigm

Chapter 1

Preliminaries

1.1 TBA How much finished

95%

1.2 Introduction

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

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

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

1.3 Clinical tasks

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

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

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

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

1.3.1 Workflow in an imaging study

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

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

1.3.2 The screening and diagnostic workup tasks

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

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

oped and adopted by radiology departments is the next topic.

1.4 Imaging device development and its clinical deployment

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

1.4.1 Physical measurements

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

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

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

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

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

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

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

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

1.4.2 Quality Control and Image quality optimization

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

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

A good example of QC is the use of the American College of Radiology Mammography Quality Standards Act (ACR-MQSA) phantom to monitor image quality of mammography machines[27-29]. The phantom consists of a (removable) wax insert in an acrylic holder; the latter provides additional absorption and scattering material to more closely match the attenuation and beam hardening of an average breast. Embedded in the wax insert are target objects consisting of 6 fibrils, five groups of microcalcifications, each containing six specks, and five spherical objects of different sizes, called masses. An image of the phantom, Fig. 1.1 (A) is obtained daily, before the first patient is imaged, and is inspected by a technologist, who records the number of target objects of different types that are visible. There is a pass-fail criterion and if the image fails then patients cannot be imaged on that machine until the problem is corrected. At this point, the medical physicist is called in to investigate.

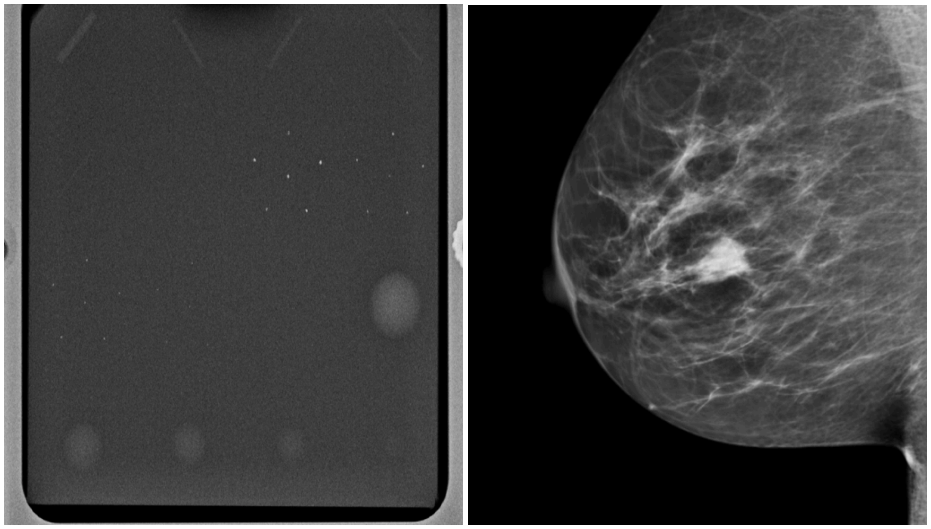


Figure 1.1: (A) Image of an ACR phantom, (B) Clinical image.

Fig. 1.1 (A – B): (A) Image of an American College of Radiology mammography accreditation phantom. The phantom contains target objects consisting of six fibrils, five groups of microcalcifications, and five nodule-like objects. An image of the phantom is obtained daily, before the first patient is imaged, and is inspected by a technologist, who records the number of target objects of different types that are visible. On his 27" iMac monitor, I see four fibrils, three speck groups and four masses, which would be graded as a "pass". This is greatly simplified version of the test. The scoring accounts for irregular fibril or partially visible masses borders, etc., all of which is intended to get more objectivity out of the measurement. (B) A breast image showing an invasive cancer, located roughly in the middle of the image. Note the lack of similarity between the two images (A) and (B). The breast image is much more complex and there is more information, and therefore more to go wrong than with the phantom image. Moreover, there is variability between patients in contrast to the fixed image in (A). In my clinical experience, the phantom images interpreted visually are a poor predictor of clinical image quality.

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

Fig. 1.1 (B) shows a mammogram with a mass-like cancer visible near it center. It is characterized by complex anatomical background, quite unlike the uniform background in the phantom image in Fig. 1.1 (A). In mammography 30% of retrospectively visible lesions are missed at initial screening and radiologist variability can be as large as 40% [35]. QC machine parameters (e.g., kVp, the kilovoltage accuracy) are usually measured to 1% accuracy. It is ironic that the weak link, in the sense of greatest variability, is the radiologist but quality control and much effort is primarily focused on measuring/improving the physical parameters of the machine. This comment is meant to motivate clinical medical physicists, most of who are focused on QC, to become more aware about observer performance methods, where achieving better than 5% accuracy is quite feasible[36]. The author believes there should be greater focus on improving radiologist performance, particularly those with marginal performance. Efforts in this direction, using ROC methods, are underway in the UK [37, 38] by Prof Alistair Gale and colleagues.

1.5 Image quality vs. task performance

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

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

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

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

1.6 Why physical measures of image quality are not enough

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

The problem with physical measures of image quality lies in relating them to clinical performance. Phantom images have little resemblance to clinical images, compare Fig. 1.1 (A) and (B). X-ray machines generally have automatic exposure control: the machines use a brief exposure to automatically sense the thickness of the patient from the detected x-rays. Based on this, the machine chooses the best combinations of technical factors (kVp and tube charge) and image processing. The machine has to be put in a special manual override mode to obtain reasonable images of phantoms, as otherwise the exposure control algorithm, which expects patient anatomy, is misled by the atypical nature of the "patient", compared to typical patient anatomy, into producing very poor phantom images. This type of problem makes it difficult to reproduce problems encountered using clinical images with phantom images. It has been my general experience that QC failures often lag clinical image quality reported problems: more often than not, clinical image quality problems are reported before QC measurements indicate a problem. This is not surprising since clinical images, e.g., Fig. 1.1 (B) are more complex and have more information[41], both in the clinical and in the information theoretic sense[42], than the much simpler phantom image shown in Fig. 1.1 (A), so there is more that can go wrong with clinical images than with phantom images. Manufacturers now design an-

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

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

1.7 Model observers

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

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

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

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

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

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

1.8 Measuring observer performance: four paradigms

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

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

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

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

1.8.1 Basic approach to the analysis

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

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

1.8.2 Historical notes

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

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

Table 1.1: FrybackThornbury hierarchy of efficacies.

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

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

1.9 Hierarchy of assessment methods

The methods described in this book need to be placed in context of a six-level hierarchy of assessment methods[7, 60]. The cited paper by Fryback and Thornbury on "The Efficacy of Diagnostic Imaging" is a highly readable account, which also gives a more complete overview of this field, including key contributions by Yerushalmy[61] and Lusted[62]. The term efficacy is defined generically as "the probability of benefit to individuals in a defined population from a medical technology applied for a given medical problem under ideal conditions of use". Demonstration of efficacy at each lower level is a necessary but not sufficient condition to assure efficacy at higher level. The different assessment methods are, in increasing order of efficacy : technical, diagnostic accuracy, diagnostic thinking, therapeutic, patient outcome and societal, Table 1.1.

Table 1.1: Fryback and Thornbury proposed hierarchy of assessment methods. Demonstration of efficacy at each lower level is a necessary but not sufficient condition to assure efficacy at higher level. [MTF = modulation transfer function; NPS(f) = noise power spectra as a function of spatial frequency f; DQE(f) = detective quantum efficiency]

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

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

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

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

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

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

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

1.10 Overview of the book and how to use it

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

1.10.1 Overview of the book

1.10.1.1 Part A: The ROC paradigm

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

1.10.1.2 Part B: The statistics of ROC analysis

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

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

1.10.1.3 Part C: The FROC paradigm

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

1.10.1.4 Part D: Advanced topics

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

1.10.1.5 Part E: Appendices (TBA)

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

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

1.10.2 How to use the book

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

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

This concludes the introduction of the book.

1.11 Summary

1.12 Discussion

1.13 References

Chapter 2

The Binary Task

2.1 TBA How much finished

85%

2.2 Introduction

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

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

Table 2.1: Truth Table.

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

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

2.3 The fundamental 2x2 table

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

Let D represent the radiologist's decision, with $D = 1$ representing the decision "case is non-diseased" and $D = 2$ representing the decision "case is diseased". Let T denote the truth with $T = 1$ representing "case is actually non-diseased" and $T = 2$ representing "case is actually diseased". Each decision, one of two values, will be associated with one of two truth states, resulting in an entry in one of 4 cells arranged in a 2 x 2 layout, termed the decision vs. truth table, Table 2.1, which is of fundamental importance in observer performance. The cells are labeled as follows. The abbreviation TN, for true negative, represents a $D = 1$ decision on a $T = 1$ case. FN, for false negative, represents a $D = 1$ decision on a $T = 2$ case (also termed a "miss"). FP, for false positive, represents a $D = 2$ decision on a $T = 1$ case (a "false-alarm") and TP, for true positive, represents a $D = 2$ decision on a $T = 2$ case (a "hit").

Table 2.2 shows the numbers of decisions in each of the four categories defined in Table 2.1. Specifically, $n(TN)$ is the number of true negative decisions, $n(FN)$ is the number of false negative decisions, etc. The last row is the sum of the corresponding columns. The sum of the number of true negative decisions $n(TN)$ and the number of false positive decisions $n(FP)$ must equal the total number of non-diseased cases, denoted K_1. Likewise, the sum of the number of false negative decisions $n(FN)$ and the number of true positive decisions $n(TP)$ must equal the total number of diseased cases, denoted K_2. The last column is the sum of the corresponding rows. The sum of the number of true negative $n(TN)$ and false negative $n(FN)$ decisions is the total number of negative decisions, denoted $n(N)$. Likewise, the sum of the number of false positive $n(FP)$ and true positive $n(TP)$ decisions is the total number of positive decisions, denoted $n(P)$. Since each case yields a decision, the bottom-right corner cell is

Table 2.2: Cell counts.

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

$n(N) + n(P)$, which must also equal $K_1 + K_2$, the total number of cases K. These statements are summarized in Eqn. (2.1).

$$\left.\begin{aligned} K_1 &= n(TN) + n(FP) \\ K_2 &= n(FN) + n(TN) \\ n(N) &= n(TN) + n(FN) \\ n(P) &= n(TP) + n(FP) \\ K = K_1 + K_2 &= n(N) + n(P) \end{aligned}\right\} \tag{2.1}$$

2.4 Sensitivity and specificity

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

$$P(D|T) = P(\text{diagnosis is D}|\text{truth is T}) \tag{2.2}$$

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

$$P(TP) = P(\text{D} = 2|\text{T} = 2) \tag{2.3}$$

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

$$P(FN) = P(\text{D} = 1|\text{T} = 2) \tag{2.4}$$

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

$$\left.\begin{aligned} P(TN) &= P(D=1|T=1) \\ P(FP) &= P(D=2|T=1) \end{aligned}\right\} \tag{2.5}$$

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

$$\left.\begin{aligned} P(D=1|T=1) + P(D=2|T=1) &= 1 \\ P(D=1|T=2) + P(D=2|T=2) &= 1 \end{aligned}\right\} \tag{2.6}$$

Equivalently, these equations can be written:

$$\left.\begin{aligned} P(TN) + P(FP) &= 1 \\ P(FN) + P(TP) &= 1 \end{aligned}\right\} \tag{2.7}$$

Comments:

- An easy way to remember Eqn. (2.7) is to start by writing down the probability of one of the four probabilities, e.g., $P(TN)$, and "reversing" both terms inside the parentheses, i.e., $T \Rightarrow F$, and $N \Rightarrow P$. This yields the term $P(FP)$ which when added to the previous probability, $P(TN)$, yields unity, i.e., the 1st equation in Eqn. (2.7).
- Because there are two equations in four unknowns, only two of the four probabilities, one per equation, are independent. By tradition these are chosen to be $P(D = 1|T = 1)$ and $P(D = 2|T = 2)$, i.e., $P(TN)$ and $P(TP)$, which happen to be the probabilities of correct decisions on non-diseased and diseased cases, respectively. The two basic probabilities are so important that they have names: $P(D = 2|T = 2) = P(TP)$ is termed sensitivity (Se) and $P(D = 1|T = 1) = P(TN)$ is termed specificity (Sp):

$$\left.\begin{aligned} \text{Se} = P(TP) &= P(D=2|T=2) \\ \text{Sp} = P(TN) &= P(D=1|T=1) \end{aligned}\right\} \tag{2.8}$$

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

2.4.1 Reasons for the names sensitivity and specificity

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

2.4.2 Estimating sensitivity and specificity

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

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

$$\left.\begin{aligned} \widehat{\mathrm{Se}} &= \widehat{P(TP)} = \tfrac{n(TP)}{K_2} \\ \widehat{\mathrm{Sp}} &= \widehat{P(TN)} = \tfrac{n(TN)}{K_1} \end{aligned}\right\} \tag{2.9}$$

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

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

2.5 Disease prevalence

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

$$\left.\begin{aligned} P(D|\text{pop}) &= P(T=2|\text{pop}) \\ P(D|\text{lab}) &= P(T=2|\text{lab}) \end{aligned}\right\} \tag{2.10}$$

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

$$\left.\begin{aligned} P(T=1|\text{pop}) + P(T=2|\text{pop}) &= 1 \\ P(T=1|\text{lab}) + P(T=2|\text{lab}) &= 1 \end{aligned}\right\} \tag{2.11}$$

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

$$\widehat{P(D|\text{pop})} = \frac{K_2}{K_1+K_2}\Big|_{pop} \tag{2.12}$$

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

$$\left.\begin{aligned} \widehat{P(D|\text{pop})} &\sim 0.005 \\ \widehat{P(D|\text{lab})} &\sim 0.5 \gg \widehat{P(D|\text{pop})} \end{aligned}\right\} \tag{2.13}$$

2.6 Accuracy

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

$$\widehat{Ac} = \frac{n(TN) + n(TP)}{n(TN) + n(TP) + n(FP) + n(FN)} \tag{2.14}$$

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

$$\widehat{Ac} = \widehat{Sp}\widehat{P(!D)} + \widehat{Se}\widehat{P(D)} \tag{2.15}$$

The exclamation mark symbol is used to denote the "not" or negation operator. For example, $P(!D)$ means the probability that the patient is not diseased. Eqn. (2.15) applies equally to laboratory or population studies, *provided sensitivity and specificity are estimated consistently.* One cannot combine a population estimate of prevalence with a laboratory measurement of sensitivity and / or specificity.

Eqn. (2.15) can be understood from the following argument. $\widehat{Sp}$ is the fraction of correct (i.e., negative) decisions on non-diseased cases. Multiplying this by $\widehat{P(!D)}$ yields $\widehat{Sp}\widehat{P(!D)}$, the fraction of correct negative decisions on all cases. Similarly, $\widehat{Sp}$ is the fraction of correct positive decisions on all cases. Therefore, their sum is the fraction of (all, i.e., negative and positive) correct decisions on all cases. A formal mathematical derivation follows. The terms on the right hand side of Eqn. (2.9) can be "turned around" yielding:

$$\left.\begin{aligned} n(TP) &= K_2\widehat{Se} \\ n(TN) &= K_1\widehat{Sp} \end{aligned}\right\} \tag{2.16}$$

Therefore,

$$\begin{aligned} \widehat{Ac} &= \frac{n(TN) + n(TP)}{K} \\ &= \frac{K_1\widehat{Sp} + K_2\widehat{Se}}{K} \\ &= \widehat{Sp}\widehat{P(!D)} + \widehat{Se}\widehat{P(D)} \end{aligned} \tag{2.17}$$

2.7 Negative and positive predictive values

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

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

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

$$NPV = P(T = 1|D = 1) \tag{2.18}$$

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

$$PPV = P(T = 2|D = 2) \tag{2.19}$$

Note that both equations are "turned around" from the definition of specificity and sensitivity, Eqn. (2.8), i.e., specificity = $P(D = 1|T = 1)$ and sensitivity = $P(D = 2|T = 2)$.

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

$$\widehat{NPV} = \frac{n(TN)}{n(TN) + n(FN)} \tag{2.20}$$

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

$$\widehat{NPV} = \frac{\widehat{P(TN)}}{\widehat{P(TN)} + \widehat{P(FN)}} \tag{2.21}$$

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

$$\widehat{P(TN)} = \widehat{P(!D)}(1 - \widehat{FPF}) \tag{2.22}$$

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

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

$$\widehat{P(FN)} = \widehat{P(D)}(1 - \widehat{TPF}) \tag{2.23}$$

Putting this all together, one has:

$$\widehat{NPV} = \frac{\widehat{P(!D)}(1 - \widehat{FPF})}{(\widehat{P(!D)}(1 - \widehat{FPF}) + (\widehat{P(D)}(1 - \widehat{TPF})} \tag{2.24}$$

For the population,

$$NPV = \frac{P(!D)(1 - FPF)}{(P(!D)(1 - FPF) + (P(D)(1 - TPF)} \tag{2.25}$$

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

$$PPV = \frac{P(D)(TPF)}{P(D)(TPF) + P(!D)FPF} \tag{2.26}$$

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

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(A)P(B|A)}{P(A)P(B|A) + P(!A)P(B|!A)} \end{aligned} \tag{2.27}$$

An easy way to remember Eqn. (2.27) is to start with the numerator on the right hand side, which is the "reversed" form of the desired probability on the left hand side, multiplied by an appropriate probability. For example, if the desired probability is $P(A|B)$, one starts with the "reversed" form, i.e., $P(B|A)$, multiplied by $P(A)$. This yields the numerator. The denominator is the sum of two probabilities: the probability of B given A, i.e., $P(B|A)$, multiplied by $P(A)$ plus the probability of B given $!A$, i.e., $P(B|!A)$, multiplied by $P(!A)$.

2.7.1 Example calculation of PPV, NPV and accuracy

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

```
# disease prevalence in
# USA screening mammography
prevalence <- 0.005 # Line 3
FPF <- 0.1 # typical operating point
TPF <- 0.8 #        do:
specificity <- 1-FPF
sensitivity <- TPF
NPV <- (1-prevalence)*(specificity)/((1-prevalence)*(specificity) + prevalence*(1-sens
PPV <- prevalence*sensitivity/(prevalence*sensitivity + (1-prevalence)*(1-specificity)
cat("NPV = ", NPV, "\nPPV = ", PPV, "\n")
#> NPV =  0.9988846
#> PPV =  0.03864734
accuracy <-(1-prevalence)*(specificity)+(prevalence)*(sensitivity)
cat("accuracy = ", accuracy, "\n")
#> accuracy =  0.8995
```

- Line 3 initializes the variable `prevalence`, the disease prevalence, to 0.005.
- Line 4 assigns `0.1` to `FPF` and line 5 assigns `0.8` to `TPF`.
- Lines 6 and 7 initialize the variables specificity and sensitivity, respectively.
- Line 8 calculates `NPV` using Eqn. (2.25).
- Line 9 calculates `PPV` using Eqn. (2.26).

2.7.2 Comments

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

2.7.3 PPV and NPV are irrelevant to laboratory tasks

According to the hierarchy of assessment methods described in (book) Chapter 01, Table 1.1, PPV and NPV are level- 3 measurements, which are calculated from "live" interpretations (recall that the higher the level the greater the clinical relevance). In the clinic, the radiologist adjusts the operating point to achieve a balance between sensitivity and specificity. The balance depends critically on the known disease prevalence. Based on geographical location and type of practice, the radiologist over time develops an idea of actual disease prevalence, or it can be found in various databases. For example, a breast-imaging clinic that specializes in imaging high-risk women will have higher disease prevalence than the general population and the radiologist is expected to err more on the side of reduced specificity because of the expected benefit of increased sensitivity. However, in the context of a laboratory study, where one uses enriched case sets, the concepts of NPV and PPV are meaningless. For example, it would be rather difficult to perform a laboratory study with 10,000 randomly sampled women, which would ensure about 50 actually diseased patients, which is large enough to get a reasonably precise estimate of sensitivity (estimating specificity is inherently more precise because most women are actually non-diseased). Rather, in a laboratory study one uses enriched data sets where the numbers of diseased-cases is much larger than in the general population, Eqn. (2.13). The radiologist cannot interpret these cases pretending that the actual prevalence is very low. Negative and positive predictive values, while they can be calculated from laboratory data, have very little, if any, clinical meanings, since they have no effect on radiologist thinking. As noted in (book) Chapter 01 the purpose of

level-3 measurements is to determine the effect on radiologist thinking. There are no diagnostic decisions riding on laboratory ROC interpretations of retrospectively acquired patient images. However, PPV and NPV do have clinical meanings when calculated from very large population based "live" studies. For example, the (Fenton et al., 2007) study sampled 684,956 women and used the results of "live" interpretations of their images. In contrast, laboratory ROC studies are typically conducted with 50-100 non-diseased and 50-100 diseased cases. A study using about 300 cases total would be considered a "large" ROC study.

2.8 Summary

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

2.9 Discussion

2.10 References

Chapter 3

Modeling the Binary Task

3.1 TBA How much finished

85%

3.2 Introduction

Chapter 3 introduced measures of performance associated with the binary decision task. Described in this chapter is a 2-parameter statistical model for the binary task, in other words it shows how one can predict quantities like sensitivity and specificity based on the values of the parameters of a statistical model. It introduces the fundamental concepts of a decision variable and a decision threshold (the latter is one of the parameters of the statistical model) that pervade this book, and shows how the decision threshold can be altered by varying experimental conditions. The receiver-operating characteristic (ROC) plot is introduced which shows how the dependence of sensitivity and specificity on the decision threshold is exploited by a measure of performance that is independent of decision threshold, namely the area AUC under the ROC curve. AUC turns out to be related to the other parameter of the model.

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

3.3 Decision variable and decision threshold

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

3.3.1 Existence of a decision variable

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

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

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

3.3.2 Existence of a decision threshold

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

- The decision threshold is a fixed value used to separate cases reported as diseased from cases reported as non-diseased.
- Unlike the random Z-sample, which varies from case to case, the decision threshold is held fixed for the duration of the study. In some of the older literature2 the decision threshold is sometimes referred to as "response bias". The author hesitates to use the term "bias" which has a negative connotation, whereas, in fact, the choice of decision threshold depends on rational assessment of costs and benefits of different outcomes.
- The choice of decision threshold depends on the conditions of the study: perceived or known disease prevalence, cost-benefit considerations, instructions regarding dataset characteristics, personal interpreting style, etc. There is a transient "learning curve" during which observer is assumed to find the optimal threshold and henceforth holds it constant for the duration of the study. The learning is expected to stabilize during a sufficiently long training interval.
- Data should only be collected in the fixed threshold state, i.e., at the end of the training session.
- If a second study is conducted under different conditions, the observer will determine, after a new training session, the optimal threshold for the new conditions and henceforth hold it constant for the duration of the second study, etc.

From assumption #2, it follows that:

$$1 - Sp = FPF = P(Z \geq \zeta | T = 1) \tag{3.1}$$

$$Se = TPF = P(Z \geq \zeta | T = 2) \tag{3.2}$$

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

There are several concepts implicit in Eqn. (3.1) and Eqn. (3.2).

- The Z-samples have an associated probability distribution; this is implicit in the notation $P(Z \geq \zeta | T = 2)$ and $P(Z \geq \zeta | T = 1)$. Diseased-cases are not homogenous; in some, disease is easy to detect, perhaps even obvious, in others the signs of disease are subtler, and in some, the disease is almost impossible to detect. Likewise, non-diseased cases are not homogenous.
- The probability distributions depend on the truth state T. The distribution of the Z-samples for non-diseased cases is in general different from that for the diseased cases. Generally, the distribution for $T = 2$ is shifted to the right of that for $T = 1$ (assuming a **positive-directed** decision variable scale). Later, specific distributional assumptions will be employed to obtain analytic expressions for the right hand sides of Eqn. (3.1) and Eqn. (3.2).
- The equations imply that via choice of the decision threshold ζ, Se and Sp are under the control of the observer. The lower the decision threshold the higher the sensitivity and the lower the specificity, and the converses are also true. Ideally both sensitivity and specificity should be large, i.e., unity (since they are probabilities they cannot exceed unity). The tradeoff between sensitivity and specificity says, essentially, that there is no "free lunch". In general, the price paid for increased sensitivity is decreased specificity and vice-versa.

3.3.3 Adequacy of the training session

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

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

3.4 Changing the decision threshold: Example I

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

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

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

3.5 Changing the decision threshold: Example II

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

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

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

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

3.6 The equal-variance binormal model

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

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

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

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

$$\left.\begin{array}{c} Z_{k_t t} \sim N(\mu_t, 1) \\ \mu_1 = 0 \\ \mu_2 = \mu \end{array}\right\} \tag{3.3}$$

In Eqn. (3.3) the subscript t denotes the truth, sometimes referred to as the "gold standard", with $t = 1$ denoting a non-diseased case and $t = 2$ denoting a diseased case. The variable $Z_{k_t t}$ denotes the random Z-sample for case $k_t t$, where k_t is the index for cases with truth state t; for example $k_1 1 = 21$ denotes the 21st non-diseased case and $k_2 2 = 3$ denotes the 3rd diseased case. To explicate $k_1 1 = 21$ further, the label k_1 indexes the case while the label 1 indicates the truth of the case. The label k_t ranges from $1, 2, ..., K_t$, where $K_t\$$ is the total number of cases with disease state t.

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

The first line in Eqn. (3.3) states that $Z_{k_t t}$ is a random sample from the $N(\mu_t, 1)$ distribution, which has unit variance regardless of the value of t (this is the reason for naming it the equal-variance binormal model). The remaining lines in Eqn. (3.3) defines μ_1 as zero and μ_2 as μ. Taken together, these equations state that non-diseased case Z-samples are distributed $N(0,1)$ and diseased case Z-samples are distributed $N(\mu,1)$. The name binormal arises from the two normal distributions underlying this model. It should not be confused with bivariate, which identifies a single distribution yielding two values per sample, where the two values could be correlated. In the binormal model, the samples from the two distributions are assumed independent of each other.

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

3.7 The normal distribution

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

By definition,

$$\phi\left(z|\mu,\sigma\right) = P(z < Z < z + dz|Z \sim N(\mu, \sigma^2)) \tag{3.4}$$

The right hand side of Eqn. (3.4) is the probability that the random variable Z, sampled from $N(\mu, \sigma^2)$, is between the fixed limits z and z + dz. For this reason $\phi(z|\mu, \sigma)$ is termed the probability density function. The special case $\phi(z|0, 1)$ is referred to as the **unit normal distribution**; it has zero mean and unit variance and the corresponding pdf is denoted $\phi(z)$. The defining equation for the pdf of this distribution is:

$$\phi\left(z\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) \tag{3.5}$$

The integral of $\phi(t)$ from $-\infty$ to z, as in Eqn. (3.6), is the probability that a sample from the unit normal distribution is less than z. Regarded as a function of z, this is termed the cumulative distribution function (CDF) and is denoted, in this book, by the symbol Φ (sometimes the term probability distribution function is used for what we are terming the CDF). The function $\Phi(z)$, specific to the unit normal distribution, is defined by:

$$\Phi\left(z\right) = \int_{-\infty}^{z} \phi(t)dt \tag{3.6}$$

Fig. 3.1 shows plots, as functions of z, of the CDF and the pdf for the unit normal distribution. Since z-samples outside ± 3 are unlikely, the plotted range, from -3 to +3 includes most of the distribution. The pdf is the familiar bell-shaped curve, centered at zero; the corresponding R function is `dnorm()`, i.e., density of the normal distribution. The CDF $\Phi(z)$ increases monotonically from 0 to unity as z increases from $-\infty$ to $+\infty$. It is the sigmoid (S-shaped) shaped curve in Fig. 3.1; the corresponding `R` function is `pnorm()`.

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

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

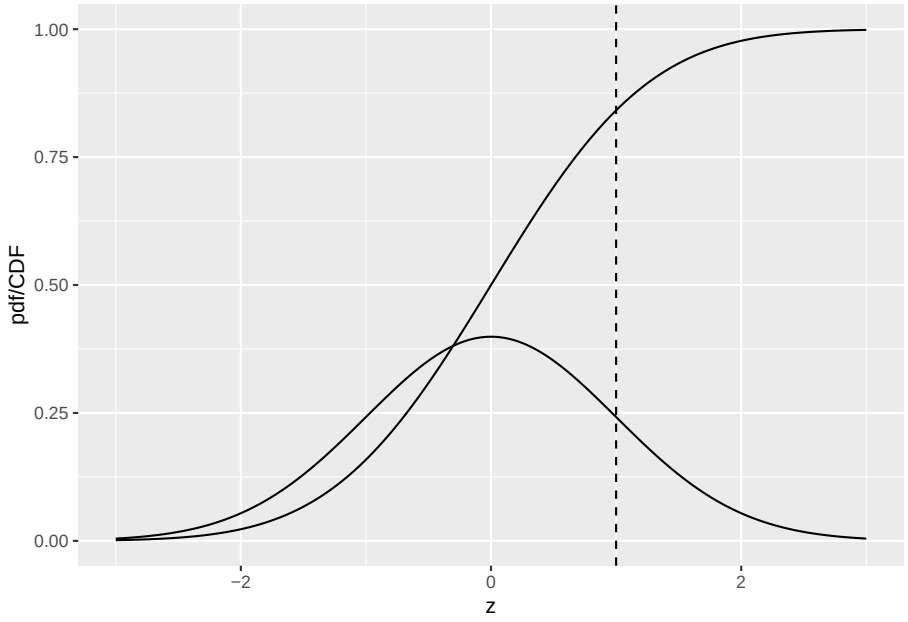


Figure 3.1: pdf-CDF plots for unit normal.

A related function is the inverse of Eqn. (3.6). Suppose the left hand side of Eqn. (3.6) is denoted p, which is a probability in the range 0 to 1.

$$p = \Phi(z) = \int_{-\infty}^{z} \phi(t)dt \tag{3.7}$$

The inverse of $\Phi(z)$ is that function which when applied to p yields the upper limit z in Eqn. (3.6), i.e.,

$$\Phi^{-1}(p) = z \tag{3.8}$$

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

$$\Phi(\Phi^{-1}(z)) = z \tag{3.9}$$

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

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

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

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

$$\Phi^{-1}(0.025) = -1.959964 \tag{3.10}$$

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

$$\Phi^{-1}(1 - 0.025) = +1.959964 \tag{3.11}$$

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

Eqn. (3.10) means that the (rounded) value -1.96 is such that the area under the pdf to the left of this value is 0.025. Similarly, Eqn. (3.11) means that the (rounded) value +1.96 is such that the area under the pdf to the left of

this value is 1-0.025 = 0.975. In other words, -1.96 captures, to its left, the 2.5th percentile of the unit-normal distribution, and 1.96 captures, to its left, the 97.5th percentile of the unit-normal distribution, Fig. 3.2. Since between them they capture 95percent of the unit-normal pdf, these two values can be used to estimate 95percent confidence intervals.

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

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

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

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

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

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

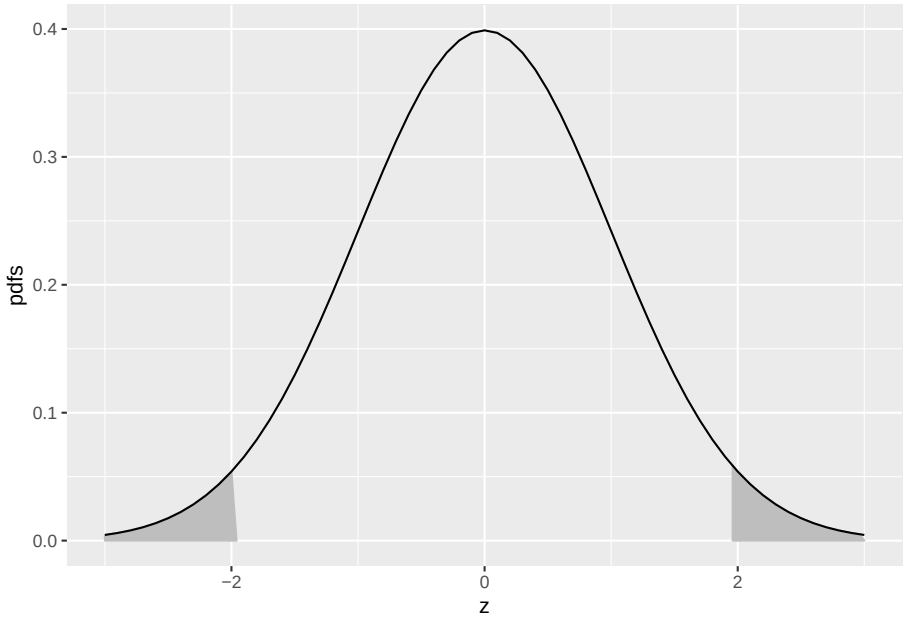


Figure 3.2: Illustrating that 95percent of the total area under the unit normal pdf is contained in the range $|Z| < 1.96$, which can be used to construct a 95percent confidence interval for an estimate of a suitably normalized statistic. The area contained in each shaded tail is 2.5percent.

3.8 Analytic expressions for specificity and sensitivity

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

$$Sp\left(\zeta\right) = P\left(Z_{k_1 1} < \zeta \mid Z_{k_1 1} \sim N\left(0,1\right)\right) = \Phi\left(\zeta\right) \tag{3.12}$$

The expression for sensitivity can be derived tediously by starting with the fact that $Z_{k_2 2}$ and then using calculus to obtain the probability that a z-sample for a disease-present case exceeds ζ. A quicker way is to consider the random variable obtaining by shifting the origin to μ. A little thought should convince the reader that $Z_{k_2 2} - \mu$ must be distributed as $N(0,1)$. Therefore, the desired probability is (the last step follows from the identity in Eqn. (3.7), with z replaced by $\zeta - \mu$:

$$\begin{aligned} Se\left(\zeta\right) & \\ = P\left(Z_{k_2 2} \geq \zeta\right) & \\ = P\left(\left(Z_{k_2 2} - \mu\right) \geq \left(\zeta - \mu\right)\right) & \\ = 1 - P\left(\left(Z_{k_2 2} - \mu\right) < \left(\zeta - \mu\right)\right) & \\ = 1 - \Phi\left(\zeta - \mu\right) & \end{aligned} \tag{3.13}$$

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

$$1 - \Phi(\zeta) = -\Phi(\zeta) 1 - \Phi(\zeta - \mu) = \Phi(\mu - \zeta) \tag{3.14}$$

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

$$Sp\left(\zeta\right) = \Phi(\zeta) Se\left(\zeta\right) = \Phi(\mu - \zeta) \tag{3.15}$$

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

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

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

The parameter μ satisfies the requirements of a scalar figure of merit (FOM). Eqn. (3.15) can be solved for μ as follows. Inverting the equations yields:

$$\zeta = \Phi^{-1}\left(Sp\left(\zeta\right)\right) \mu - \zeta = \Phi^{-1}\left(Se\left(\zeta\right)\right) \tag{3.16}$$

Eliminating ζ yields:

$$\mu = \Phi^{-1}\left(Sp\left(\zeta\right)\right) + \Phi^{-1}\left(Se\left(\zeta\right)\right) \tag{3.17}$$

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

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

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

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

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

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

Fig. 3.3 shows the equal-variance binormal model for $\mu = 3$ and $\zeta = 1$. The blue-shaded area, including the "common" portion with the vertical red lines, is the probability that a z-sample from a non-diseased case exceeds $\zeta = 1$, which is the complement of specificity, i.e., it is false positive fraction, which is 1 - `pnorm(1)` = 0.159. The red shaded area, including the "common" portion with the vertical red lines, is the probability that a z-sample from a diseased case exceeds $\zeta = 1$, which is sensitivity or true positive fraction, which is `pnorm(3-1)`= 0.977.

Demonstrated next are these concepts using R examples.

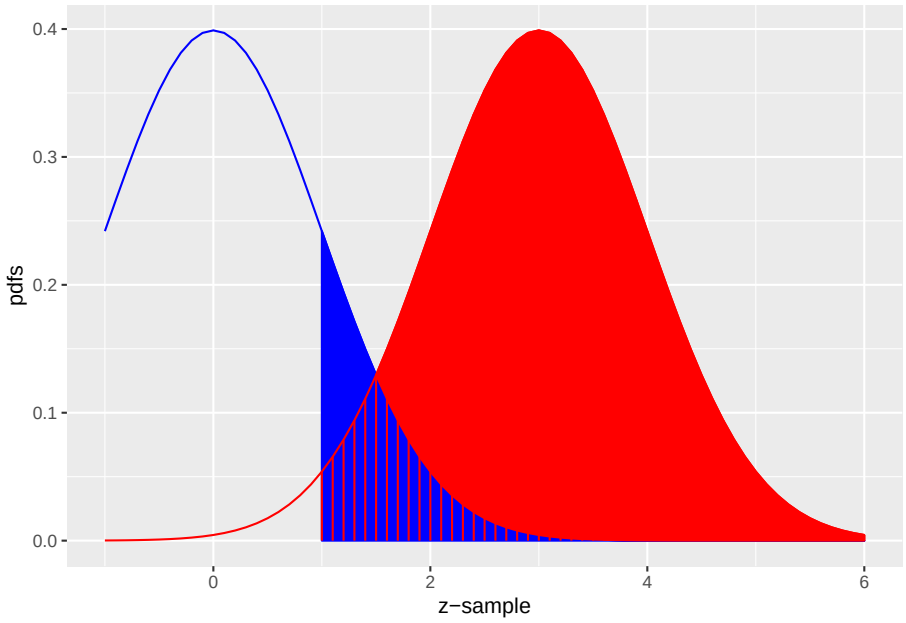


Figure 3.3: The equal-variance binormal model for mu = 3 and zeta = 1; the blue curve, centered at zero, corresponds to the pdf of non-diseased cases and the red one, centered at mu = 3, corresponds to the pdf of diseased cases. The left edge of the blue shaded region represents the threshold zeta, currently set at unity. The red shaded area, including the common portion with the vertical red lines, is sensitivity. The blue shaded area including the common portion with the vertical red lines is 1-specificity.

3.9 Demonstration of the concepts of sensitivity and specificity

3.9.1 Estimating mu from a finite sample

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

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

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

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

3.9.2 Changing the seed variable: case-sampling variability

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

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

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

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

3.9.3 Increasing the numbers of cases

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

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

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

Next we change `seed` to 101.

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

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

Notice that now the values are less sensitive to seed. Table 3.1 illustrates this trend with ever increasing sample sizes (the reader should confirm the listed values).

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

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

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

Table 3.1: Effect of sample size and seed on estimates of sensitivity, specificity and the mu-parameter.

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

Because the threshold is halfway between the two distributions, as in this example, sensitivity and specificity are identical. In words, with two unit variance distributions separated by 1.5, the area under the diseased distribution (centered at 1.5) above 0.75, namely sensitivity, equals the area under the non-diseased distribution (centered at zero) below 0.75, namely specificity, and the common value is $\Phi(0.75) = 0.773$, yielding the last row of Table 3.1, and example 1 in the above code snippet. Example 2 in the above code snippet illustrates Eqn. (3.17). The factor of two arises since in this example sensitivity and specificity are identical.

From Table 3.1, for the same numbers of cases but different seeds, comparing pairs of sensitivity and specificity values is more difficult as two pairs of numbers (i.e., four numbers) are involved. Comparing a single pair of μ values is easier as only two numbers are involved. The tendency of the pairs to become independent of case sample is discernible with fewer cases with μ, around 90/110 cases, than with sensitivity and specificity pairs. The numbers in the table might appear disheartening in terms of the implied numbers of cases needed to detect a difference in specificity. Even with 200 cases, the difference in specificity for two seed values is 0.081, which is actually a large effect considering that the scale extends from 0 to 1.0. A similar comment applies to differences in sensitivity. The situation is not quite that bad. One uses an area measure that combines sensitivity and specificity yielding less variability in the combined measure. One uses the ratings paradigm, which is more efficient than the binary one used in this chapter. Finally, one takes advantage of correlations that exist between the interpretations in matched-case matched-reader interpretations in two modalities that tend to decrease variability in the AUC-difference even further (most applications of ROC methods involved detecting differences in AUCs not absolute values).

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

The variation of sensitivity and specificity is modeled in the binormal model by the threshold parameter ζ. From Eqn. (3.12), specificity at threshold ζ is $\Phi(\zeta)$ and the corresponding expression for sensitivity is $\Phi(\mu - \zeta)$. Since the threshold ζ appears with a minus sign, the dependence of sensitivity on ζ will be the opposite of the corresponding dependence of specificity on ζ. In Fig. 3.3, the left edge of the blue shaded region represents the threshold $\zeta = 1$. As $\zeta = 1$ is moved towards the left, specificity decreases but sensitivity increases. Specificity decreases because less of the non-diseased distribution lies to the left of the new threshold, in other words fewer non-diseased cases are correctly diagnosed as non-diseased. Sensitivity increases because more of the diseased distribution lies to the right of the new threshold, in other words more diseased cases are correctly diagnosed as diseased. If an observer has higher sensitivity than another observer, but lower specificity, it is difficult to unambiguously compare them. It is not impossible (Skaane et al., 2013). The unambiguous comparison is difficult for the following reason. Assuming the second observer can be coaxed into adopting a lower threshold, thereby decreasing specificity to match that of the first observer, then it is possible that the second observer's sensitivity, formerly smaller, could now be greater than that of the first observer. A single figure of merit is desirable to the sensitivity - specificity analysis. It is possible to leverage the inverse variation of sensitivity and specificity by combing them into a single scalar measure, as was done with the μ parameter in the previous section, Eqn. (3.17). An equivalent way is by using the area under the ROC plot, discussed next.

3.11 The ROC curve

The receiver operating characteristic (ROC) is defined as the plot of sensitivity (y-axis) vs. 1-specificity (x-axis). Equivalently, it is the plot of TPF (y-axis) vs. FPF (x-axis). From Eqn. (3.15) it follows that:

$$\begin{aligned} FPF(\zeta) &= 1 - Sp(\zeta) \\ &= \Phi(-\zeta) \\ TPF(\zeta) &= Se(\zeta) \\ &= \Phi(\mu - \zeta) \end{aligned} \tag{3.18}$$

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

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

$$O(\zeta \mid \mu) = (\Phi(-\zeta), \Phi(\mu - \zeta)) \tag{3.19}$$

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

The ROC curve is the locus of the operating point for fixed μ and variable ζ. Fig. 3.4 shows examples of equal-variance binormal model ROC curves for different values of μ. Each curve is labeled with the corresponding value of μ. Each has the property that TPF is a monotonically increasing function of FPF and the slope decreases monotonically as the operating point moves up the curve. As μ increases the curves get progressively upward-left shifted, approaching the top-left corner of the ROC plot. In the limit $\mu = \infty$ the curve degenerates into two line segments, a vertical one connecting the origin to (0,1) and a horizontal one connecting (0,1) to (1,1) – the ROC plot for a perfect observer.

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

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

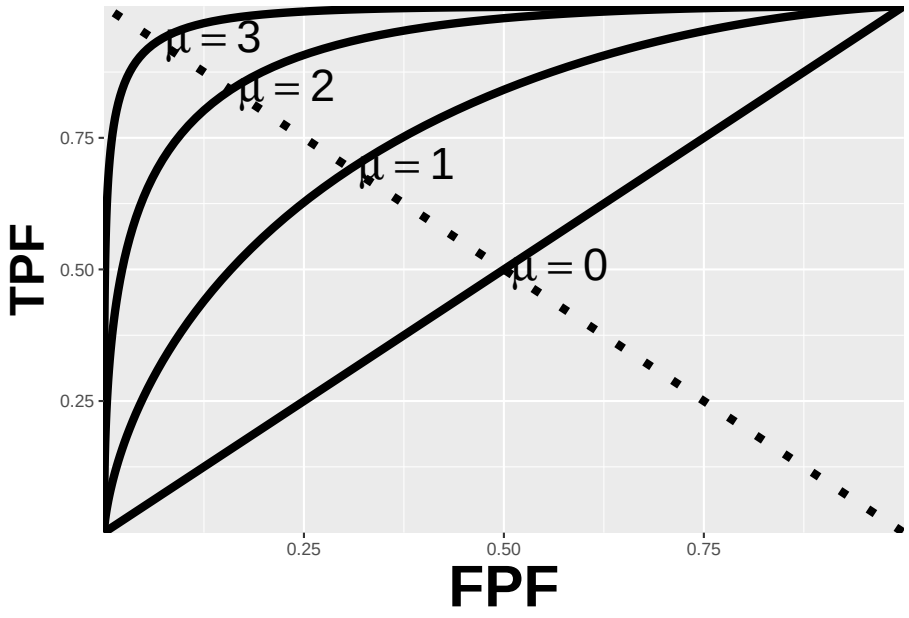


Figure 3.4: ROC plots predicted by the equal variance binormal model for different values of mu. As mu increases the intersection of the curve with the negative diagonal moves closer to the ideal operating point, (0,1) at which sensitivity and specificity are both equal to unity.

3.11.1 The chance diagonal

In Fig. 3.4 the ROC curve for $\mu = 0$ is the positive diagonal of the ROC plot, termed the chance diagonal. Along this curve $TPF = FPF$ and the observer's performance is at chance level. In the equal variance binormal model, for $\mu = 0$, the pdf of the diseased distribution is identical to that of the non-diseased distribution: both are centered at the origin. Therefore, no matter the choice of threshold ζ, $TPF = FPF$. Setting $\mu = 0$ in Eqn. (3.18) yields:

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

In this special case, the red and blue curves in Fig. 3.3 coincide. The observer is unable to find any difference between the two distributions. This can happen if the cancers are of such low visibility so that diseased cases are indistinguishable from non-diseased ones, or the observer's skill level is so poor that the observer is unable to make use of distinguishing characteristics between diseased and non-diseased cases that do exist, and which experts exploit.

3.11.2 The guessing observer

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

3.11.3 Symmetry with respect to negative diagonal

A characteristic of the ROC curves shown in Fig. 3.4 is that they are symmetric with respect to the negative diagonal, defined as the straight line joining (0,1) and (1,0) which is shown as the dotted straight line in Fig. 3.4. The symmetry property is due to the equal variance nature of the binormal model and is not true for models considered in later chapters. The intersection between the ROC curve and the negative diagonal corresponds to $\zeta = \mu/2$, in which case the operating point is:

$$\begin{aligned} FPF(\zeta) &= \Phi(-\mu/2) \\ TPF(\zeta) &= \Phi(\mu/2) \end{aligned} \tag{3.20}$$

The first equation implies:

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

Therefore,

$$TPF(\zeta) = 1 - FPF(\zeta) \tag{3.21}$$

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

3.11.4 Area under the ROC curve

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

$$\begin{aligned} A_{z;\sigma=1} &= \int_0^1 TPF(\zeta) d(FPF(\zeta)) \\ &= \int_0^1 FPF(\zeta) d(TPF(\zeta)) \end{aligned} \tag{3.22}$$

Eqn. (3.22) has the following equivalent interpretations:

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

By convention, the symbol A_z is used for the area under the binormal model predicted ROC curve. In Eqn. (3.22), the extra subscript $\sigma = 1$ is necessary to distinguish it from another one corresponding to the unequal variance binormal model to be derived later. It can be shown that:

$$A_{z;\sigma=1} = \Phi\left(\frac{\mu}{\sqrt{2}}\right) \tag{3.23}$$

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

3.11.5 Properties of the equal-variance binormal model ROC curve

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

3.11.6 Comments

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

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

$$\left.\begin{aligned} \frac{d(TPF)}{d(FPF)} &= \frac{d(\Phi(\mu-\zeta))}{d(\Phi(-\zeta))} \\ &= \frac{\phi(\mu-\zeta)}{\phi(-\zeta)} \\ &= exp(\mu(\zeta-\mu/2)) \propto exp(\mu\zeta) \end{aligned}\right\} \tag{3.24}$$

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

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

Since the slope of the ROC curve can be expressed as a power of e, it is always non-negative. Provided $\mu > 0$, then, in the limit $\zeta \to \infty$, the slope at the origin approaches ∞. Eqn. (3.24) also implies that in the limit $\zeta \to -\infty$ the slope of the ROC curve at the end-point (1,1) approaches zero, i.e., the slope is a monotone increasing function of ζ. As ζ decrease from $+\infty$ to $-\infty$, the slope decreases monotonically from $+\infty$ to 0.

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

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

3.11.7 Physical interpretation of the mu-parameter

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

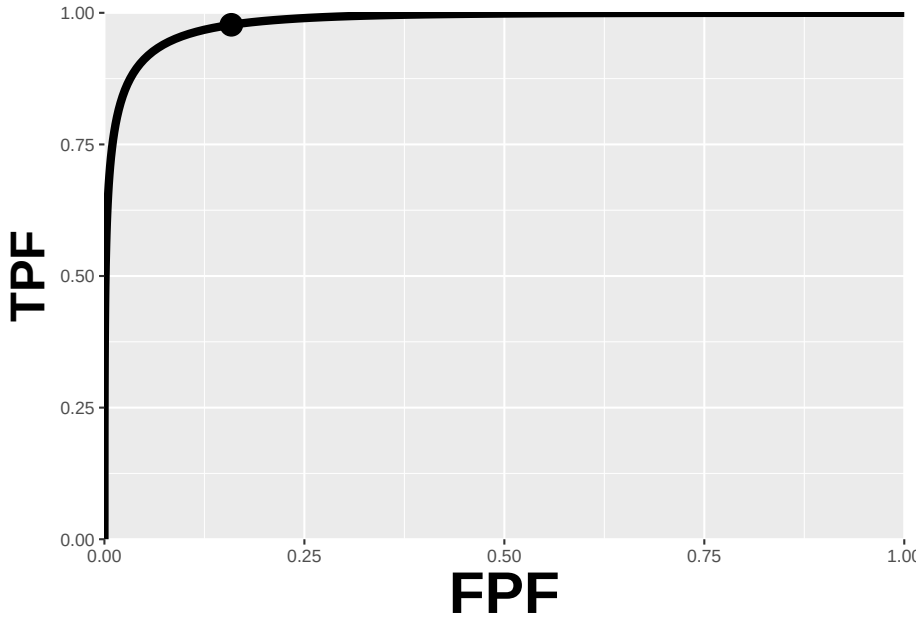


Figure 3.5: ROC curve predicted by equal variance binormal model for mu = 3. The circled operating point corresponds to zeta = 1. The operating point falls exactly on the curve, as these are analytical results. Due to sampling variability, with finite numbers of cases, this is not observed in practice.

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

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

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

3.12 Assigning confidence intervals to an operating point

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

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

$$n(FP) \sim B\left(K_1, \widehat{FPF}\right) \tag{3.25}$$

In Eqn. (3.25), $B(n,p)$ denotes the binomial distribution with success probability p and trial size n:

$$\left.\begin{array}{r} k \sim B\left(n,p\right) \\ k = 0,1,2,...,n \end{array}\right\} \tag{3.26}$$

Eqn. (3.26) states that k is a random sample from the binomial distribution $B(n,p)$. For reference, the probability mass function pmf of $B(n,p)$ is defined by (the subscript Bin denotes a binomial distribution):

$$\text{pmf}_{Bin}\left(k;n,p\right) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.27}$$

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

The binomial coefficient $\binom{n}{k}$ appearing in Eqn. (3.27), to be read as "n pick k", is defined by:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{3.28}$$

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

$$\sigma^2_{n(FP)} = K_1 \widehat{FPF}\left(1 - \widehat{FPF}\right) \tag{3.29}$$

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

$$\sigma^2_{FPF} = \frac{\widehat{FPF}\left(1 - \widehat{FPF}\right)}{K_1} \tag{3.30}$$

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

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

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

$$CI^{FPF}_{1-\alpha} = \left(\widehat{FPF} - z_{\alpha/2}\sigma_{FPF}, \widehat{FPF} + z_{\alpha/2}\sigma_{FPF}\right) \tag{3.31}$$

In Eqn. (3.31), $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the unit normal distribution, i.e., the area to the *right* under the unit normal distribution pdf from $z_{\alpha/2}$ to ∞ equals $\alpha/2$. It is the complement (i.e., plus goes to minus) of $\Phi^{-1}(\alpha/2)$ introduced earlier; the difference is that the latter uses the area to the *left*. The following code might help.

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

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

$$\left.\begin{aligned} z_{\alpha/2} &= \Phi^{-1}\left(1-\alpha/2\right) \\ \alpha/2 &= \int_{z_{\alpha/2}}^{\infty} \phi(z)dz \\ &= 1 - \Phi(z_{\alpha/2}) \end{aligned}\right\} \tag{3.32}$$

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

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

$$CI^{TPF}_{1-\alpha} = \left(\widehat{TPF} - z_{\alpha/2}\sigma_{TPF}, \widehat{TPF} + z_{\alpha/2}\sigma_{TPF}\right) \tag{3.33}$$

In Eqn. (3.33),

$$\sigma^2_{TPF} = \frac{\widehat{TPF}\left(1-\widehat{TPF}\right)}{K_2} \tag{3.34}$$

The confidence intervals are largest when the probabilities (FPF or TPF) are close to 0.5 and decrease inversely as the square root of the relevant number of cases. The symmetric binomial distribution based estimates can stray outside the allowed range (0 to 1). Exact confidence intervals9 that are asymmetric around the central value and which are guaranteed to be in the allowed range can be calculated: it is implemented in R in function `binom.test()`and used below (The approximate confidence intervals can exceed the allowed ranges, but the exact confidence intervals do not):

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

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

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

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

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

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

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

3.13 Variability in sensitivity and specificity: the Beam et al study

In this study (Beam et al., 1996) fifty accredited mammography centers were randomly sampled in the United States. "Accredited" is a legal/regulatory term implying, among other things, that the radiologists interpreting the breast cases were "board certified" by the American Board of Radiology. One hundred eight (108) certified radiologists from these centers gave blinded interpretation to a common set of 79 randomly selected enriched screening cases containing 45 cases with cancer and the rest normal or with benign lesions. Ground truth for these women had been established either by biopsy or by 2-year follow-up (establishing truth is often the most time consuming part of conducting an ROC study). The observed range of sensitivity (TPF) was 53percent and the range of FPF was 63percent; the corresponding range for AUC was 21percent, Table 3.2.

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

In Fig. 3.6, a schematic of the data, if one looks at the points labeled (B) and (C) one can mentally construct a smooth ROC curve that starts at (0,0), passes roughly through these points and ends at (1,1). In this sense, the intrinsic performances (i.e., AUCs or equivalently the parameter) of the two radiologists are similar. The only difference between them is that radiologist (B) is using

Table 3.2: The variability of 108 radiologists on a common dataset of screening mammograms. Note the reduced variability when one uses AUC, which accounts for variations in reporting thresholds (AUC variability range is 21percent compared to 53percent for sensitivity and 63percent for specificity).

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

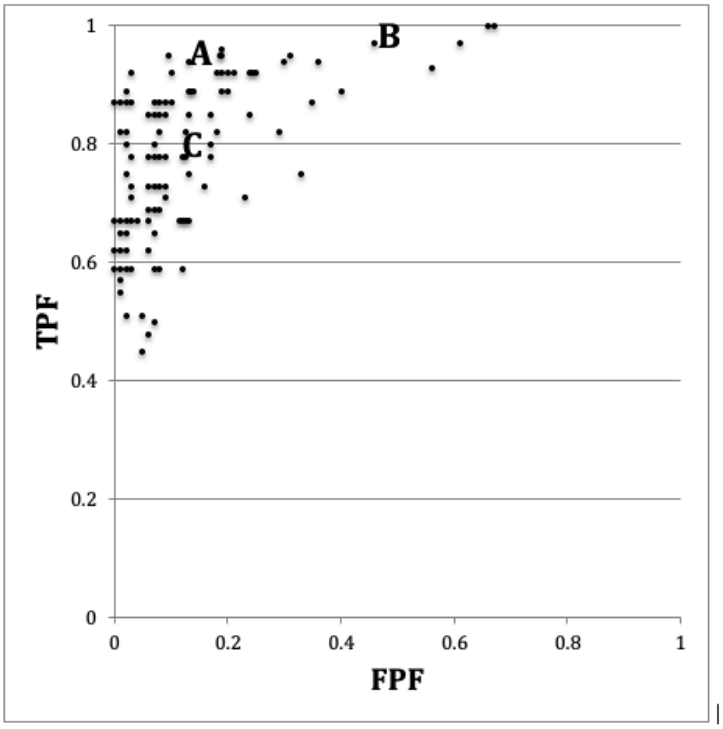


Figure 3.6: Schematic, patterned from the Beam et al study, showing the ROC operating points of 108 mammographers. Wide variability in sensitivity (40percent) and specificity (45percent) are evident. Radiologists (B) and (C) appear to be trading sensitivity for specificity and vice versa, while radiologist A's performance is intrinsically superior. See summary of important principles below.

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

Fig. 3.6 and Table 3.2 illustrate several important principles. 1. Since an operating point is characterized by two values, unless both numbers are higher (e.g., radiologist A vs. B or C), it is difficult to unambiguously compare them. 2. While sensitivity and specificity depend on the reporting threshold, the area under the ROC plot is independent of it. Using the area under the ROC curve one can unambiguously compare two readers. 3. Combining sensitivity and the complement of specificity into a single AUC measure yields the additional benefit of lower variability. In Fig. 3.6, the range for sensitivity is 53 percent while that for specificity is 63 percent. In contrast, the range for AUC is only 21 percent. This means that much of the observed variations in sensitivity and specificity are due to variations in thresholds, and using AUC eliminates this source of variability. Decreased variability of a measure is a highly desirable characteristic as it implies the measurement is more precise, making it easier to detect genuine changes between readers and / or modalities.

3.14 Summary

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

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

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

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

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

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

3.15 References

Chapter 4

Ratings Paradigm

4.1 TBA How much finished

80%

4.2 Introduction

In Chapter 2 the binary paradigm and associated concepts (e.g., sensitivity, specificity) were introduced. Chapter 2 introduced the concepts of a random scalar decision variable, or z-sample for each case, which is compared, by the observer to a fixed reporting threshold ζ, resulting in two types of decisions. It described a statistical model, characterized by two unit-variance normal distributions separated by μ, for the binary task. The concept of an underlying receiver operating characteristic (ROC) curve with the reporting threshold defining an operating point on the curve was introduced and the advisability of using the area under the curve as a measure of performance, which is independent of reporting threshold, was stressed.

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

Table 4.1: Representative counts table.

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

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

4.3 The ROC counts table

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

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

At the conclusion of the ROC study an ROC counts table is constructed. This is the generalization to rating studies of the 2 x 2 decision vs. truth table introduced in Chapter 2, Table 2.1. This type of data representation is sometimes called a frequency table, but frequency usually means a rate of number of events per some unit, so I prefer the clearer term "counts".

Table 4.1 is a representative counts table for a 5-rating study that summarizes the collected data. It is the starting point for analysis. It lists the number of counts in each ratings bin, listed separately for non-diseased and diseased cases, respectively. The data is from an actual clinical study (Barnes et al., 1989).

In this table:

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

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

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

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

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

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

The spread appears to be more pronounced for the diseased cases, e.g., five of the 50 cases appeared to be definitely non-diseased to the observer. However, one is forewarned not to jump to conclusions about the spread of the data being larger for diseased than for non-diseased cases based on observed rating alone. While it turns out to be true as will be shown later, the **ratings are merely ordered labels**, and modeling is required, see Chapter 6, that uses only the *ordering information* implicit in the labels, not the *actual values*, to reach quantitative conclusions.

4.4 Operating points from counts table

Table 4.2 illustrates how ROC operating points are calculated from the cell counts. In this table:

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

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

Table 4.2: Computation of operating points from cell counts.

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

and divides by the total number of non-diseased cases, $K_1 = 60$. This yields the abscissa of the lowest non-trivial operating point, namely $FPF_{\geq 5} = 1/60 = 0.017$. The subscript on FPF is intended to make explicit which ratings are being cumulated. The corresponding ordinate is obtained by dividing the number of diseased cases rated "5" or more and dividing by the total number of diseased cases, $K_2 = 50$, yielding $TPF_{\geq 5} = 22/50 = 0.440$. Therefore, the coordinates of the lowest operating point are (0.017, 0.44). The abscissa of the next higher operating point is obtained by dividing the number of non-diseased cases that were rated "4" or more and dividing by the total number of non-diseased cases, i.e., $TPF_{\geq 4} = 3/60 = 0.05$. Similarly the ordinate of this operating point is obtained by dividing the number of diseased cases that were rated "4" or more and dividing by the total number of diseased cases, i.e., $FPF_{\geq 4} = 34/50 = 0.680$. The procedure, which at each stage cumulates the number of cases equal to or greater (in the sense of increased confidence level for disease presence) than a specified ordered label, is repeated to yield the rest of the operating points listed in Table 4.2. Since they are computed directly from the data, without any assumption, they are called empirical or observed operating points.

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

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

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

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

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

$$r = 1, 2, ..., (R+1)$$

The operating points are defined by:

$$\left.\begin{aligned} FPF_r &= \frac{1}{K_1}\sum_{s=r}^{R+1} K_{1s} \\ TPF_r &= \frac{1}{K_2}\sum_{s=r}^{R+1} K_{2s} \end{aligned}\right\} \tag{4.1}$$

4.4.1 Labeling the points

The labeling O_n of the points follows the following convention: From Eqn. (4.1), the point corresponding to $r = 1$ would correspond to the upper right corner (1,1) of the ROC plot, a trivial operating point since it is common to all datasets, and is therefore not shown. The labeling starts with the next lower-left point, labeled O_1, which corresponds to $r = 2$; the next lower-left point is labeled O_2, corresponding to $r = 3$, etc., and the point labeled O_4 is the lowest non-trivial operating point corresponding to $r = R = 5$ and finally O_R corresponding to $r = R + 1$ is the origin (0,0) of the ROC plot, which is also a trivial operating point, because it is common to all datasets, and is therefore not shown. **To summarize, the operating points are labeled starting with the upper right corner, labeled O_1, and working down the curve, each time increasing the number by one. The total number of points is $R-1$.** The relation between n in the label and r in Eqn. (4.1) is $n = r-1$. An example of the labeling is shown in the next chapter, Fig. 5.1.

4.4.2 Examples

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

4.5 Automating all this

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

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

Notice that the values of the arrays `FPF` and `TPF` are identical to those listed in Table 4.2. Regarding the last four lines of code, it was shown in Chapter 2 that in the equal variance binormal model the operating point determines the parameters $\mu = 1.282$, Eqn. (3.17), or equivalently $A_{z;\sigma=1} = 0.818$, Eqn. (3.23). The last four lines illustrate the application of these formulae using the coordinates (0.5, 0.9) of the uppermost non-trivial operating point, i.e., one is fitting the equal variance model to the uppermost operating point.

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

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

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

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

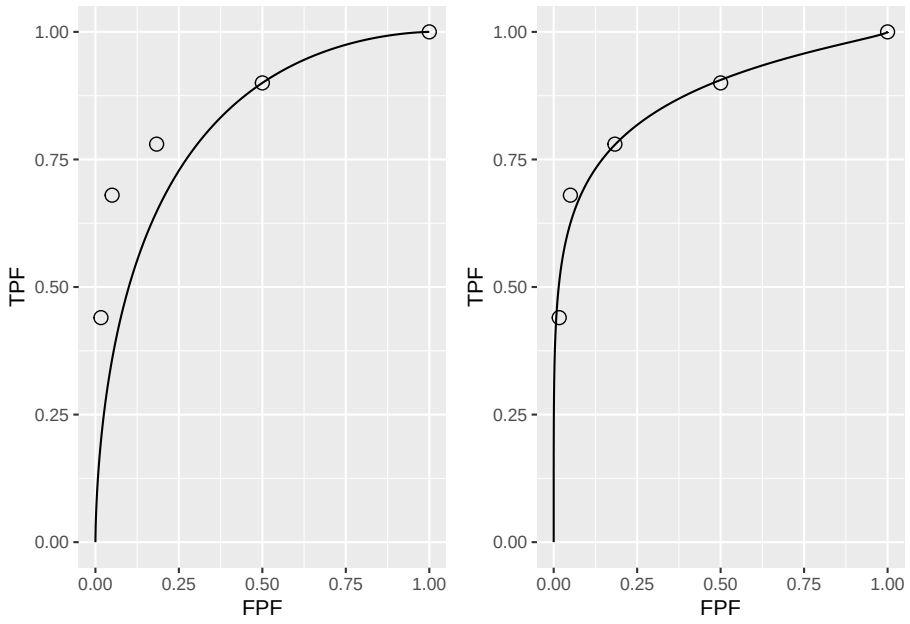


Figure 4.1: (A): The left figure is the predicted ROC curve for $\mu = 1.282$ superposed on the operating points. (B): The right figure is the same data fitted with a two-parameter model described later.

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

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

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

These are coded below:

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

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

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

- In Fig. 4.1 (A), the ROC curve, as determined by the uppermost operating point, passes exactly through this point but misses the others. If a different operating point were used to estimate μ and $A_{z;\sigma=1}$, the estimated values would have been different and the new curve would pass exactly through the *new* selected point. No single-point based choice of μ would yield a satisfactory visual fit to all the observed operating points. **This is the reason one needs a modified model, with an extra parameter, namely the unequal variance binormal model, to fit radiologist data** (the extra parameter is the ratio of the standard deviations of the two distributions).

- Fig. 4.1 (B) shows the predicted ROC curve by the unequal variance binormal model, to be introduced in Chapter 06. The corresponding parameter values are $\mu = 2.17$and $\sigma = 1.65$.

- Notice the improved visual quality of the fit. Each observed point is "not engraved in stone", rather both FPF and TPF are subject to sampling variability. Estimation of confidence intervals for FPF and TPF was addressed, see (3.31) and (3.33). [A detail: the estimated confidence interval in the preceding chapter was for a single operating point; since the multiple operating points are correlated – some of the counts used to calculate them are common to two or more operating points – the method tends to overestimate the confidence interval. A modeling approach to estimating confidence intervals accounts for these correlations and yields tighter confidence intervals.]

4.6 Relation between ratings paradigm and the binary paradigm

Table 4.1 and Table 4.2 correspond to $R = 5$. In Chapter 2 it was shown that the binary task requires a single fixed threshold parameter ζ and a decision or binning rule Eqn. (4.2): assign the case a diseased rating of 2 if $Z > \zeta$ and a rating of 1 otherwise.

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

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

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

$$\left.\begin{array}{r} if\,(\zeta_{r-1} \leq z < \zeta_r) \Rightarrow \mathrm{r}ating = r \\ r = 1, 2, ..., R \end{array}\right\} \tag{4.2}$$

For Table 4.2, the **empirical** thresholds are as follows:

$$\left.\begin{array}{l} \zeta_r = r + 1 \\ r = 1, 2, ..., R-1 \\ \zeta_0 = -\infty \\ \zeta_R = \infty \end{array}\right\} \tag{4.3}$$

The empirical thresholds are integers, as distinct from the floating point values predicted by Eqn. (4.5). **Either way one gets the same operating points.** This is a subtle and important distinction, which is related to the next section: one has enormous flexibility in the choice of the scale adopted for the decision variable axis.

In Table 4.1 the number of bins is $R = 5$. The "simultaneously conducted binary tasks" nature of the rating task can be appreciated from the following

examples. Suppose one selects the threshold for the first binary task to be $\zeta_4 = 5$. By definition, $\zeta_5 = \infty$; therefore a case rated 5 satisfies the binning rule $\zeta_4 \leq 5 < \zeta_5$, i.e., Eqn. (4.2). The operating point corresponding to $\zeta_4 = 5$, obtained by cumulating all cases rated five, yields $(0.017, 0.440)$. In the second binary-task, one selects as threshold $\zeta_3 = 4$. Therefore, a case rated four satisfies the binning rule $\zeta_3 \leq 4 < \zeta_4$. The operating point corresponding to $\zeta_3 = 4$, obtained by cumulating all cases rated four or five, yields $(0.05, 0.680)$. Similarly, for $\zeta_2 = 3$, $\zeta_1 = 2$ and $\zeta_0 = -\infty$, which yield counts in bins 3, 2 and 1, respectively. The last is a trivial operating point. The non-trivial operating points are generated by thresholds ζ_r, where $r = 1$, 2, 3 and 4. A five-rating study has four associated thresholds and a corresponding number of equivalent binary studies. In general, an R rating study has $R - 1$ associated thresholds.

4.7 Ratings are not numerical values

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

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

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

4.8 A single "clinical" operating point from ratings data

The reason for the quotes in the title to this section is that a single operating point on a laboratory ROC plot, no matter how obtained, has little relevance to how radiologists operate in the clinic. However, some consider it useful to quote an operating point from an ROC study. For a 5-rating ROC study, Table 4.1, it is not possible to unambiguously calculate the operating point of the observer in the binary task of discriminating between non-diseased and diseased cases. One possibility would be to use the "three and above" ratings to define the operating point, but one might jus have well have chosen "two and above". A second possibility is to instruct the radiologist that a "four and above" rating, for example, implies the case would be reported "clinically" as diseased. However, the radiologist can only pretend so far that this study, which has no clinical consequences, is somehow a "clinical" study.

If a single laboratory study based operating point is desired (Nishikawa, 2012), the best strategy, in my opinion, is to obtain the rating via two questions. This method is also illustrated in Table 3.1 of a book on detection theory (Macmillan and Creelman, 1991). The first question is "is the case diseased?" The binary (Yes/No) response to this question allows unambiguous calculation of the operating point, as in Chapter 2. The second question is: "what is your confidence in your previous decision?" and allow three responses, namely Low, Medium and High. The dual-question approach is equivalent to a 6-point rating scale, Fig. 4.2. The answer to the first question, is the patient diseased, allows unambiguous construction of a single "clinical" operating point for disease presence. The answer to the second question, what is your confidence level in that decision, yields multiple operating points.

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

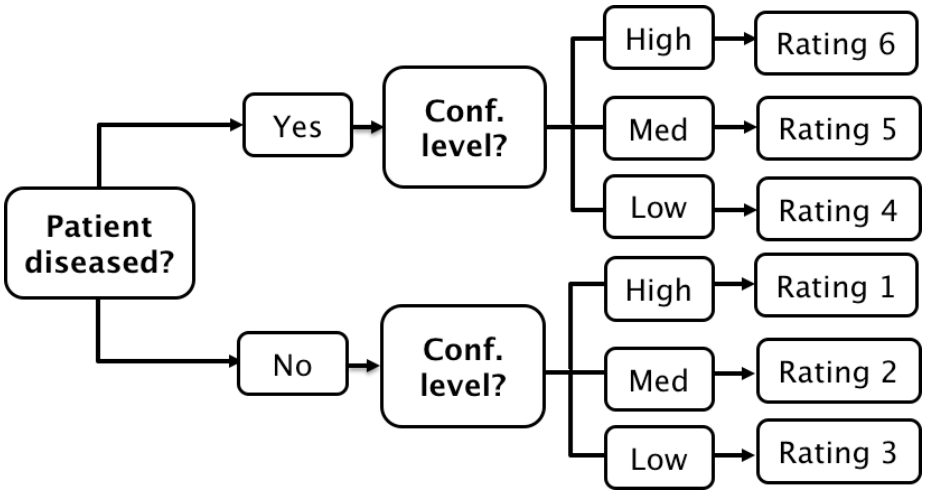


Figure 4.2: A method for acquiring ROC data on an effectively 6-point scale that also yields an unambiguous single operating point for declaring patients diseased. Note the reversal of the final ratings in the last "column" in the lower half of the figure.

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

4.9 The forced choice paradigm

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

scores is the probability of a correct choice, denoted $P(C)$. If the total number of cases presented at the same time is denoted n, then the task is termed n-alternative forced choice or nAFC (Green and Swets, 1966). If only two cases are presented, one diseased and the other non-diseased, then n = 2 and the task is 2AFC. In Fig. 4.3, in the left image a Gaussian nodule is superposed on a square region extracted from a non-diseased mammogram. The right image is a region extracted from a different non-diseased mammogram (one should not use the same background in the two images – the analysis assumes that different, i.e., independent images, are shown). If the observer clicks on the left image, a correct choice is recorded. [In some 2AFC-studies, the backgrounds are simulated non-diseased images. They resemble mammograms; the resemblance depends on the expertise of the observer: expert radiologists can tell that they are not true mammograms. They are actually created by filtering the random white noise with a 1/f3 spatial filter (Burgess, 2011).]

The 2AFC paradigm is popular, because its analysis is straightforward, and there exists a theorem4 that $P(C)$, the probability of a correct choice in the 2AFC task, equals, to within sampling variability, the *true* area under the true (not fitted, not empirical) ROC curve. Another reason for its popularity is possibly the speed at which data can be collected, sometimes only limited by the speed at which disk stored images can be displayed on the monitor. While useful for studies into human visual perception on relatively simple images, and the model observer community has performed many studies using this paradigm (Bochud et al., 1999), I cannot recommend it for clinical studies because *it does not resemble any clinical task*. In the clinic, radiologists never have to choose the diseased patient out of a pair consisting of one diseased and one non-diseased. Additionally, the forced-choice paradigm is wasteful of known-truth images, often a difficult/expensive resource to come by, because better statistics21 (tighter confidence intervals) are obtained by the ratings ROC method or by utilizing location specific extensions of the ROC paradigm. [I am not aware of the 2AFC method being actually used to assess imaging systems using radiologists to perform real clinical tasks on real images.]

Fig. 4.3: Example of image presentation in a 2AFC study. The left image contains, at its center, a positive contrast Gaussian shape disk superposed on a non-diseased mammogram. The right image does not contain a lesion at its center and the background is from a different non-diseased patient. If the observer clicks on the left image it is recorded as a correct choice, otherwise it is recorded as an incorrect choice. The number of correct choices divided by the number of paired presentations is an estimate of the probability of a correct choice, which can be shown to be identical, apart from sampling variability, to the true area under the ROC curve. This is an example of a signal known exactly location known exactly (SKE-LKE) task widely used by the model observer community.

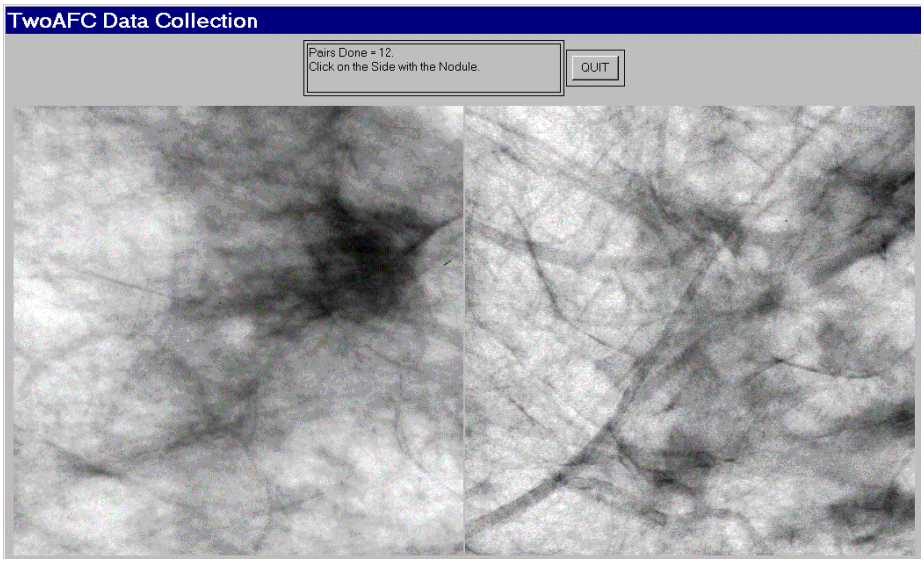


Figure 4.3: Example of image presentation in a 2AFC study.

4.10 Observer performance studies as laboratory simulations of clinical tasks

- Observer performance paradigms (ROC, FROC, LROC and ROI) should be regarded as experiments conducted in a laboratory (i.e., controlled) setting that are intended to be representative of the actual clinical task. They should not to be confused with performance in a real "live" clinical setting: there is a known "laboratory effect" (Gur et al., 2008). For example, in the just cited study radiologists performed better during live clinical interpretations than they did later, on the same cases, in a laboratory ROC study. This is to be expected because there is more at stake during live interpretations: e.g., the patient's health and the radiologist's reputation, than during laboratory ROC studies. The claimed "laboratory effect" has caused some minor controversy. A paper (Soh et al., 2013) titled "Screening mammography: test set data can reasonably describe actual clinical reporting" argues against the laboratory effect.

- Real clinical interpretations happen every day in radiology departments all over the world. On the other hand, in the laboratory, the radiologist is asked to interpret the images "as if in a clinical setting" and render a "diagnosis". The laboratory decisions have no clinical consequences, e.g., the radiologist will not be sued for mistakes and their laboratory study decisions will have no impact on the clinical management of the pa-

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

- There is no guarantee that results of the laboratory study are directly applicable to clinical practice. Indeed there is an assumption that the laboratory study correlates with clinical performance. Strict equality is not required, simply that the performance in the laboratory is related monotonically to actual clinical performance. Monotonicity assures preservation of performance orderings, e.g., a radiologist has greater performance than another does or one modality is superior to another, regardless of how they are measured, in the laboratory or in the clinic. The correlation is taken to be an axiomatic truth by researchers, when in fact it is an assumption. To the extent that the participating radiologist brings his/her full clinical expertise to bear on each laboratory image interpretation, i.e., takes the laboratory study seriously, this assumption is likely to be valid.

- This title of this section provoked a strong response from a collaborator. To paraphrase him, "*... I think it is a pity in this book chapter you argue that these studies are simulations. I mean, the reason people perform these studies is because they believe in the results*".

- I also believe in observer performance studies. Distrust of the word "simulation" seems to be peculiar to this field. Simulations are widely used in "hard" sciences, e.g., they are used in astrophysics to determine conditions dating to 10^{-31} seconds after the big bang. Simulations are not to be taken lightly. Conducting clinical studies is very difficult as there are many factors not under the researcher's control. Observer performance studies of the type described in this book are the closest that one can come to the "real thing" as they include key elements of the actual clinical task: the entire imaging system, radiologists (assuming the radiologist take these studies seriously in the sense of bringing their full expertise to bear on each image interpretation) and real clinical images. As such are expected to correlate with real "live" interpretations.

4.11 Discrete vs. continuous ratings: the Miller study

- There is controversy about the merits of discrete vs. continuous ratings (Rockette et al., 1992; Wagner et al., 2001). Since the late Prof. Charles E. Metz and the late Dr. Robert F. Wagner have both backed the latter (i.e., continuous or quasi-continuous ratings) new ROC study designs sometimes tend to follow their advice. I recommend a 6-point rating scale as outlined in Fig. 4.2. This section provides the background for the recommendation.

- A widely cited (22,909 citations at the time of writing) 1954 paper by Miller (Miller, 1956) titled "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information" is relevant. It is a readable paper, freely downloadable in several languages (www.musanim.com/miller1956/). In my judgment, this paper has not received the attention it should have in the ROC community, and for this reason portions from it are reproduced below. [George Armitage Miller, February 3, 1920 – July 22, 2012, was one of the founders of the field of cognitive psychology.]

- Miller's first objective was to comment on absolute judgments of unidimensional stimuli. Since all (univariate, i.e., single decision per case) ROC models assume a unidimensional decision variable, Miller's work is highly relevant. He comments on two papers by Pollack (Pollack, 1952, 1953). Pollack asked listeners to identify tones by assigning numerals to them, analogous to a rating task described above. The tones differed in frequency, covering the range 100 to 8000 Hz in equal logarithmic steps. A tone was sounded and the listener responded by giving a numeral (i.e., a rating, with higher values corresponding to higher frequencies). After the listener had made his response, he was told the correct identification of the tone. When only two or three tones were used, the listeners never confused them. With four different tones, confusions were quite rare, but with five or more tones, confusions were frequent. With fourteen different tones, the listeners made many mistakes. Since it is so succinct, the entire content of the first (1952) paper by Pollack is reproduced below:

- "In contrast to the extremely acute sensitivity of a human listener to discriminate small differences in the frequency or intensity between two sounds is his relative inability to identify (and name) sounds presented individually. When the frequency of a single tone is varied in equal-logarithmic steps in the range between 100 cps and 8000 cps (and when the level of the tone is randomly adjusted to reduce loudness cues), the amount of information transferred is about 2.3 bits per stimulus presentation. This is equivalent to perfect identification among only 5 tones. The information transferred, under the conditions of measurement employed, is reasonably invariant under wide variations in stimulus conditions."

- By "information" is meant (essentially) the number of levels, measured in bits (binary digits), thereby making it independent of the unit of measurement: 1 bit corresponds to a binary rating scale, 2 bits to a four-point rating scale and 2.3 bits to $2^{2.3} = 4.9$, i.e., about 5 ratings bins. Based on Pollack's' original unpublished data, Miller put an upper limit of 2.5 bits (corresponding to about 6 ratings bins) on the amount of information that is transmitted by listeners who make absolute judgments of auditory pitch. The second paper (@ Pollack, 1953) by Pollack was related to: (1) the frequency range of tones; (2) the utilization of objective reference

> tones presented with the unknown tone; and (3) the "dimensionality"—the number of independently varying stimulus aspects. Little additional gain in information transmission was associated with the first factor; a moderate gain was associated with the second; and a relatively substantial gain was associated with the third (we return to the dimensionality issue below).

- As an interesting side-note, Miller states:

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

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

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

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

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

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

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

- Another reason given for using continuous ratings is it reduces instances of data degeneracy. Data is sometimes said to be degenerate if the curve-fitting algorithm, the binormal model and the proper binormal model, cannot fit it (in simple terms, the program crashes). This occurs, for example, if there are no interior points on the ROC plot. Modifying radiologist behavior to accommodate the limitations of analytical methods seems to be inherently dubious. One could simply randomly add or subtract half an integer from the observed ratings, thereby making the rating scale more granular and reduce instances of degeneracy (this is actually

done in some ROC software to overcome degeneracy issues). Another possibility is to use the empirical (trapezoidal) area under the ROC curve, which can always be calculated; there are no degeneracy problems with it. Actually, fitting methods now exist that are robust to data degeneracy, such as discussed in TBA Chapter 18 and Chapter 20, so this reason for acquiring continuous data no longer applies.

- The rating task involves a unidimensional scale and I see no way of getting around the basic channel-limitation noted by Miller and for this reason I recommend a 6 point scale, as in Fig. 4.2.

- On the other side of the controversy (Berbaum et al., 2002), a position that I agree with, it has been argued that given a large number of allowed ratings levels the cooperating observer essentially bins the data into a much smaller number of bins (e.g., 0, 20, 40, 60, 80, 100) and then adds a zero-mean noise term to appear to be "spreading out the ratings". This ensures that the binormal model does not crash. However, if the intent is to get the observer to spread the ratings, so that the binormal model does not crash, a better approach is to use alternate models that do not crash and are, in fact, very robust with respect to degneracy of the data. More on this later (see Chapters TBA CBM and RSM).

4.12 The BI-RADS ratings scale and ROC studies

It is desirable that the rating scale be relevant to the radiologists' daily practice. This assures greater consistency – the fitting algorithms assume that the thresholds are held constant for the duration of the ROC study. Depending on the clinical task, a natural rating scale may already exist. For example, in 1992 the American College of Radiology developed the Breast Imaging Reporting and Data System (BI-RADS) to standardize mammography reporting36. There are six assessment categories: category 0 indicates need for additional imaging; category 1 is a negative (clearly non-diseased) interpretation; category 2 is a benign finding; category 3 is probably benign, with short-interval follow-up suggested; category 4 is a suspicious abnormality for which biopsy should be considered; category 5 is highly suggestive of malignancy and appropriate action should be taken. The 4th edition of the BI-RADS manual37 divides category 4 into three subcategories 4A, 4B and 4C and adds category 6 for a proven malignancy. The 3-category may be further subdivided into "probably benign with a recommendation for normal or short-term follow-up" and a 3+ category, "probably benign with a recommendation for immediate follow-up". Apart from categories 0 and 2, the categories form an ordered set with higher categories representing greater confidence in presence of cancer. How to handle the 0s and the 2s is the subject of some controversy, described next.

Table 4.3: The Barlow et al study: the ordering of the BI-RADS ratings in the first column correlates with cancer-rate in the last column.

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

4.13 The controversy

Two large clinical studies have been reported in which BI-RADS category data were acquired for $>$ 400,00 screening mammograms interpreted by many (124 in the 1st study) radiologists (Barlow et al., 2004; Fenton et al., 2007). The purpose of the first study was to relate radiologist characteristics to actual performance (e.g., does performance depend on reading volume – the number of cases interpreted per year), so it could be regarded as a more elaborate version of (Beam et al., 1996), described in Chapter 2. The purpose of the second study was to determine the effectiveness of computer-aided detection (CAD) in screening mammography.

The reported ROC analyses used the BIRADS assessments labels ordered as follows: $1 < 2 < 3 < 3+ < 0 < 4 < 5$. The last column of Table 4.3 shows that with this ordering the numbers of cancer per 1000 patients increases monotonically. The CAD study is discussed later, for now the focus is on the adopted BIRADS scale ordering that is common to both studies and which has raised controversy (the controversy appears to be limited to observer performance study analysts).

The use of the BI-RADS ratings shown in Table 4.3 has been criticized (Jiang and Metz, 2010) in an editorial titled:

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

Since BI-RADS is a clinical rating scheme widely used in mammography, the editorial, if correct, implies that ROC analysis of clinical mammography data is not possible. Since the BI-RADS scale was arrived at after considerable

deliberation, inability to perform ROC analysis with it would strike at the root of clinical utility of the ROC method. The purpose of this section is to express the reasons why I have a different take on this controversy.

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

I agree with the Barlow et al. suggested ordering of the "2s" as more likely to have cancer than the "1s". A category-2 means the radiologist found something to report, and the location of the finding is part of the clinical report. Even if the radiologist believes the finding is definitely benign, there is a finite probability that a category-2 finding is cancer, as evident in the last column of Table 4.3 ($1.43 > 0.83$). In contrast, there are no findings associated with a category-1 report. A paper (Hartmann et al., 2005) titled:

> Benign breast disease and the risk of breast cancer

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

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

Moreover, the last column of Table 4.3 suggests that any other ordering would violate an important principle, namely, optimal ordering is achieved when each case is rated according to it's likelihood ratio (defined as the probability of the case being diseased divided by the probability of the case being non-diseased). The likelihood ratio is the "betting odds" of the case being diseased, which is expected to be monotonic with the empirical probability of the case being diseased, i.e., the last column of Table 4.3. Therefore, the ordering adopted in Table 4.3 is equivalent to adopting a likelihood ratio scale and any other ordering would not be monotonic with likelihood ratio.

The likelihood ratio is described in more detail in the TBA Chapter 20, which describes ROC fitting methods that yield "proper" ROC curves, i.e., ones that have monotonically decreasing slope as the operating point moves up the curve from (0,0) to (1,1) and therefore do not (inappropriately) cross the chance diagonal. Key to these fitting methods is adoption of a likelihood ratio scale to rank-order cases, instead of the ratings assumed by the unequal variance binormal model. The proper ROC fitting algorithm implemented in PROPROC

software reorders confidence levels assumed by the binormal model, TBA Chapter 20, paragraph following Fig. 20.4. This is analogous to the reordering of the clinical ratings based on cancer rates assumed in Table 4.3. It is illogical to allow reordering of ratings in "blind" software but question the same when done in a principled way by a researcher. As expected, the modeled ROC curves in the Barlow publication, their Fig. 4, show no evidence of improper behavior. This is in contrast to a clinical study (about fifty thousands patients spread over 33 hospitals with each mammogram interpreted by two radiologists) using a non-BIRADS 7-point rating scale which yielded markedly improper ROC curves (Pisano et al., 2005) for the film modality when using ROC ratings (not BIRADS). This suggests that use of a non-clinical ratings scale for clinical studies, without independent confirmation of the ordering implied by the scale, is problematical.

The reader might be interested as to reason for the 0-ratings being more predictive of cancer than a 3+ rating, Table 4.3. In the clinic the zero rating implies, in effect, "defer decision, incomplete information, additional imaging necessary". A zero rating could be due to technical problems with the images: e.g., improper positioning (e.g., missing breast tissue close to the chest wall) or incorrect imaging technique (improper selection of kilovoltage and/or tube charge), making it impossible to properly interpret the images. Since the images are part of the permanent patient record, there are both healthcare and legal reasons why the images need to be optimal. Incorrect technical factors are expected to occur randomly and therefore not predictive of cancer. However, if there is a suspicious finding and the image quality is sub-optimal, the radiologist may be unable to commit to a decision, they may seek additional imaging, perhaps better compression or a slightly different view angle to resolve the ambiguity. Such zero ratings are expected with suspicious findings, and therefore are expected to be predictive of cancer.

As an aside, the second paper (Fenton et al., 2007) using the ordering shown in Table 4.3 questioned the utility of CAD for breast cancer screening (this was ca. 2007). This paper was met with flurry of correspondence disputing the methodology (summarized above). The finding regarding utility of CAD has been validated by more recent studies, again with very large case and reader samples, showing that usage of CAD can actually be detrimental to patient outcome (Philpotts, 2009) and a call (Fenton, 2015) for ending insurance reimbursement for CAD.

4.14 Discussion

In this chapter the widely used ratings paradigm was described and illustrated with a sample dataset. The calculation of ROC operating points from this table was detailed. A formal notation was introduced to describe the counts in this table and the construction of operating points and an R example was given. I do

not wish to leave the impression that the ratings paradigm is used only in medical imaging. In fact the historical reference (Macmillan and Creelman, 1991) to the two-question six-point scale in Fig. 4.2, namely Table 3.1 in the book by MacMillan and Creelman, was for a rating study on performance in recognizing odors. The early users of the ROC ratings paradigm were mostly experimental psychologists and psychophysicists interested in studying perception of signals, some in the auditory domain, and some in other sensory domains.

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

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

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

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

4.15 References

Chapter 5

Empirical AUC

5.1 TBA How much finished

80%

5.2 Introduction

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

One of the reasons for fitting observed counts data, such as in Table 4.1 in Chapter 04, to a parametric model, is to derive analytical expressions for the separation parameter 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, TBA Fig. 4.1 (A - B), that the equal variance binormal model did not describe a clinical dataset and that an unequal variance binormal model yielded a better visual fit. This turns out to be an almost universal finding. Before getting into the complexity of the unequal variance binormal model curve

Table 5.1: On the need for two indices to label cases in an ROC study.

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

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

5.3 The empirical ROC plot

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

5.3.1 Notation for cases

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

TBA In 5.1 the upper row shows 11 non-diseased cases, labeled N1 – N11, while the lower row shows seven diseased cases, labeled D1 – D7. To address any case one needs two indices: the row number t and the column number $k_t t$. Since in general the column number depends on the value of t, one needs two indices

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

5.3.2 An empirical operating point

Let $z_{k_t t}$ represent the 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:

$$\left.\begin{aligned} FPF\left(\zeta\right) &= \frac{1}{K_1}\sum_{k_1=1}^{K_1} I\left(z_{k_1 1} \geq \zeta\right) \\ TPF\left(\zeta\right) &= \frac{1}{K_2}\sum_{k_2=1}^{K_2} I\left(z_{k_2 2} \geq \zeta\right) \end{aligned}\right\} \tag{5.1}$$

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

In Eqn. (5.1) 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:

$$O\left(\zeta\right) = \left(FPF\left(\zeta\right), TPF\left(\zeta\right)\right) \tag{5.2}$$

The essential difference between Eqn. (5.1) and Eqn. (3.18) is that the former is non-parametric while the latter is parametric. In TBA Chapter 03 analytical (or parametric, i.e., model parameter dependent) operating points were obtained. In contrast, here one uses the observed ratings to calculate the empirical operating point.

5.4 Empirical operating points from ratings data

Consider a ratings ROC study with R bins. Describing an R-rating empirical ROC plot requires $R-1$ ordered empirical thresholds, see Eqn. (4.3).

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

$$O(\zeta_r) = (FPF(\zeta_r), TPF(\zeta_r)) \tag{5.3}$$

Its coordinates are defined by:

$$\left.\begin{aligned} FPF_r \equiv FPF(\zeta_r) &= \frac{1}{K_1}\sum_{k_1=1}^{K_1} I\left(z_{k_1 1} \geq \zeta_r\right) \\ TPF_r \equiv TPF(\zeta_r) &= \frac{1}{K_2}\sum_{k_2=1}^{K_2} I\left(z_{k_2 2} \geq \zeta_r\right) \end{aligned}\right\} \tag{5.4}$$

For example,

$$\left.\begin{aligned} FPF_4 \equiv FPF(\zeta_4) &= \frac{1}{K_1}\sum_{k_1=1}^{K_1} I\left(z_{k_1 1} \geq \zeta_4\right) \\ TPF_4 \equiv TPF(\zeta_4) &= \frac{1}{K_2}\sum_{k_2=1}^{K_2} I\left(z_{k_2 2} \geq \zeta_4\right) \\ O_4 \equiv (FPF_4, TPF_4) &= (0.017, 0.44) \end{aligned}\right\} \tag{5.5}$$

In Table 4.1 a sample clinical ratings data set was introduced. Shown below is a partial code listing of mainEmpRocPlot.R showing implementation of Eqn. (5.7). Except for the last statement, the plotting part of the code is suppressed.

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

ROCOp <- data.frame(FPF = FPF, TPF = TPF)
ROCPlot <- ggplot(
  data = ROCOp,
  mapping = aes(x = FPF, y = TPF)) +
```

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

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

The function `cumsum()` is used to calculate the cumulative sum. The `rev()`

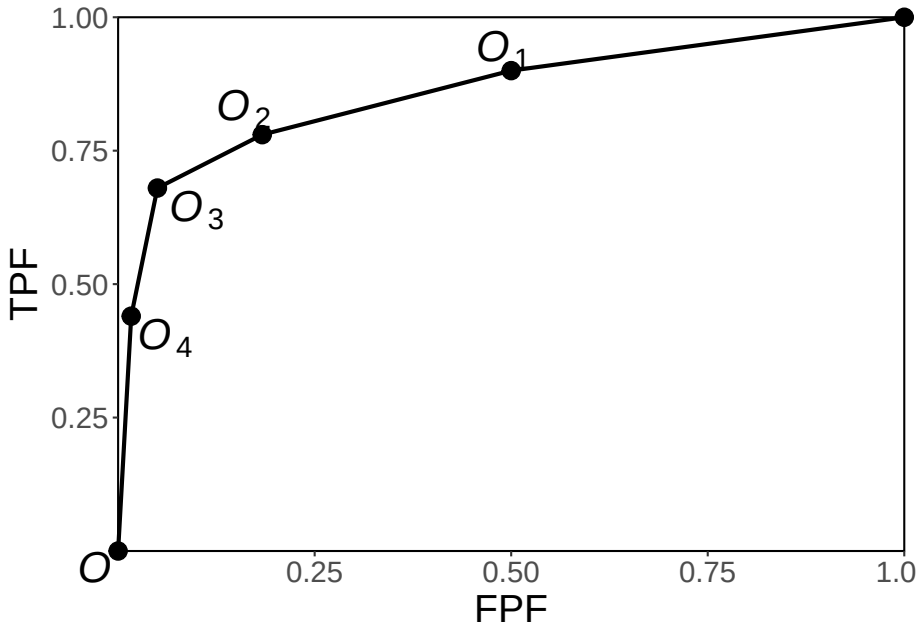


Figure 5.1: Empirical ROC plot for the data in Table 4.1. By convention the operating points are numbered starting with the uppermost non-trivial one and working down the plot and the trivial operating points (0,0) and (1,1) are not shown.

function reverses the order of the array supplied as its argument. The reader should use the debugging techniques (basically copy and paste parts of the code to the Console window and hit enter) to understand how this code implements Eqn. (5.4).

Fig. 5.1 is the empirical ROC plot. It illustrates the convention used to label the operating points introduced in TBA §4.3 is, i.e., O_1 is the uppermost non-trivial point, and the subscripts increment by unity as one moves down the plot. By convention, not shown are the trivial operating points $O_0 \equiv (FPF_0, TPF_0) = (1, 1)$ and $O_R \equiv (FPF_R, TPF_R) = (0, 0)$, where $R = 5$.

5.5 AUC under the empirical ROC plot

Fig. 5.2 shows the empirical plot for the data in Table 4.1. The area under the curve (AUC) is the shaded area. By dropping imaginary vertical lines from the non-trivial operating points onto the x-axis, the shaded area is seen to be the sum of one triangular shaped area and four trapezoids. One may be tempted to write equations to calculate the total area using elementary algebra, but that would be unproductive. There is a theorem (see below) that the empirical area is exactly equal to a particular statistic known as the Mann-Whitney-Wilcoxon statistic (Wilcoxon, 1945; Mann and Whitney, 1947), which, in this book, is abbreviated to the Wilcoxon statistic. Calculating this statistic is much simpler than calculating and summing the areas of the triangle and trapezoids, or doing planimetry.

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

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

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

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

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

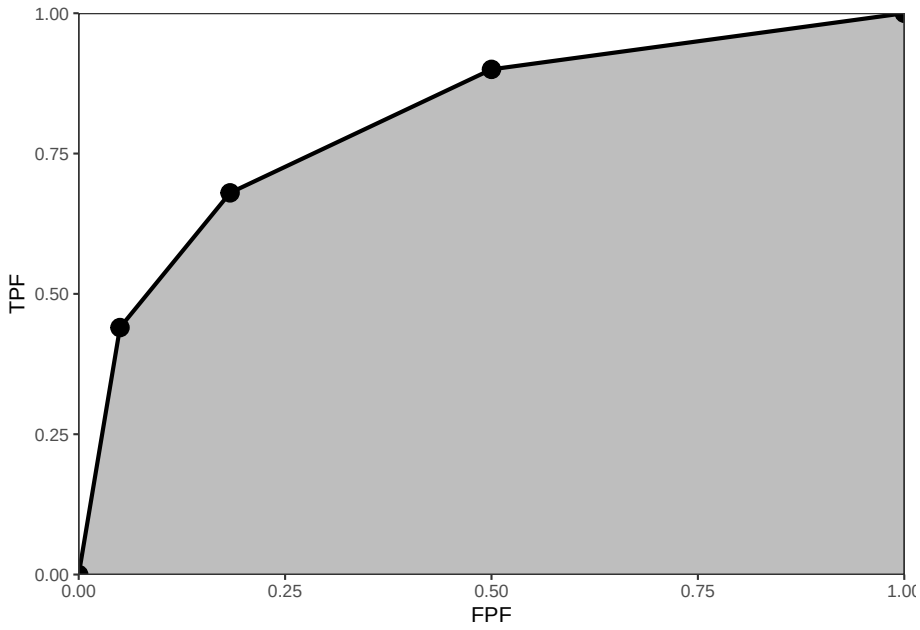


Figure 5.2: The empirical ROC plot corresponding to Table 4.1; the shaded area is the area AUC under this plot, a widely used figure of merit in non-parametric ROC analysis.

5.6 The Wilcoxon statistic

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

$$W = \frac{1}{K_1 K_2} \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \psi\left(z_{k_1 1}, z_{k_2 2}\right) \tag{5.6}$$

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

$$\left.\begin{array}{ll} \psi(x,y) = 1 & x < y \\ \psi(x,y) = 0.5 & x = y \\ \psi(x,y) = 0 & x > y \end{array}\right\} \tag{5.7}$$

The function $\psi(x, y)$ is sometimes called the kernel function. It is unity if the diseased case is rated higher, 0.5 if the two are rated the same and zero otherwise. Each evaluation of the kernel function results from a comparison of a case from the non-diseased set with one from the diseased set. In Eqn. (5.6) the two summations and division by the total number of comparisons yields the observed, i.e., empirical, probability that diseased cases are rated higher than non-diseased ones. Since it is a probability, it can range from zero to one. However, if the observer has any discrimination ability at all, one expects diseased cases to be rated equal or greater than non-diseased ones, so in practice one expects $0.5 \leq W \leq 1$. The limit 0.5 corresponds to a guessing observer, whose operating point lies on the chance diagonal of the ROC plot.

5.7 Bamber's Equivalence theorem

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

$$W = AUC \tag{5.8}$$

Numerical illustration: While hardly a proof, as an illustration of the theorem it is helpful to calculate the sum on the right hand side of Eqn. (5.6) and compare it to direct integration of the area under the empirical ROC curve (i.e., adding the area of a triangle and several trapezoids). The function is called `trapz(x,y)`, see below. It takes two array arguments, x and y, where in the current case x is FPF and y is TPF. One has to be careful to include the end-points as otherwise the area will be underestimated. The Wilcoxon W and the numerical estimate of the empirical area AUC are implemented in the following code.

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

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

RocOperatingPoints <- function( K1, K2 ) {

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

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

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

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

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

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

Note the equality of the two estimates.

The following proof is adapted from (Bamber, 1975) and while it may appear to be restricted to discrete ratings, the result is in fact quite general, i.e., it is applicable even if the ratings are acquired on a continuous scale. The reason is that in an R-rating ROC study the observed z-samples or ratings take on integer values, 1 through R. If R is large enough, ordering information present in the continuous data is not lost upon binning. In the following it is helpful to keep in mind that one is dealing with discrete distributions of the ratings, described by probability mass functions as opposed to probability density functions, e.g., $P(Z_2 = \zeta_i)$ is not zero, as would be the case for continuous ratings. The proof is illustrated with Fig. 5.3.

The abscissa of the operating point i is $P(Z_1 \geq \zeta_i)$ and the corresponding ordinate is $P(Z_2 \geq \zeta_i)$. Here Z_1 is a random sample from a non-diseased case and Z_2 is a random sample from a diseased case. The shaded trapezoid defined by drawing horizontal lines from operating points i (upper) and $i+1$ (lower) to the right edge of the ROC plot, Fig. 5.3, has height:

$$P\left(Z_2 \geq \zeta_i\right) - P\left(Z_2 \geq \zeta_{i+1}\right) = P\left(Z_2 = \zeta_i\right) \tag{5.9}$$

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

$$P\left(Z_2 \geq \zeta_i\right) = P\left(Z_2 = \zeta_i\right) + P\left(Z_2 \geq \zeta_{i+1}\right) \tag{5.10}$$

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

$$1 - P\left(Z_1 \geq \zeta_i\right) = P\left(Z_1 < \zeta_i\right) \tag{5.11}$$

and

$$1 - P\left(Z_1 \geq \zeta_{i+1}\right) = P\left(Z_1 < \zeta_{i+1}\right) \tag{5.12}$$

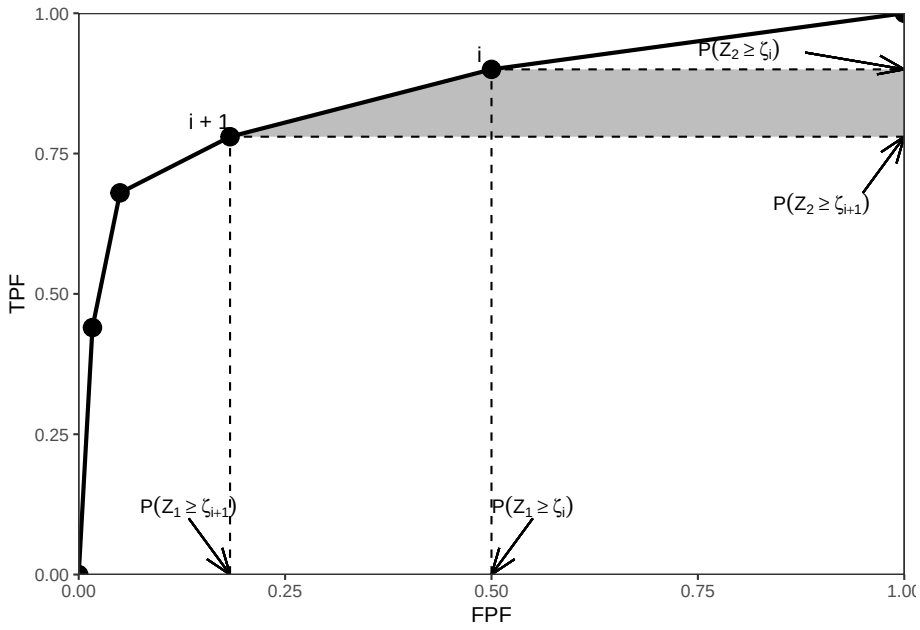


Figure 5.3: :Illustration of the derivation of Bamber's equivalence theorem. Shows an empirical ROC plot for R = 5; the shaded area is due to points labeled i and i + 1.

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

$$\left.\begin{aligned} A_i &= \frac{1}{2} P\left(Z_2 = \zeta_i\right)\left[P\left(Z_1 < \zeta_i\right) + P\left(Z_1 < \zeta_{i+1}\right)\right] \\ A_i &= P\left(Z_2 = \zeta_i\right)\left[\frac{1}{2} P\left(Z_1 < \zeta_i\right) + \frac{1}{2}\left(P\left(Z_1 = \zeta_i\right) + P\left(Z_1 < \zeta_i\right)\right)\right] \\ A_i &= P\left(Z_2 = \zeta_i\right)\left[\frac{1}{2} P\left(Z_1 = \zeta_i\right) + P\left(Z_1 < \zeta_i\right)\right] \end{aligned}\right\} \quad (5.13)$$

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

$$\left.\begin{aligned} AUC &= \sum_{i=0}^{R-1} A_i \\ &= \frac{1}{2} \sum_{i=0}^{R-1} P\left(Z_2 = \zeta_i\right) P\left(Z_1 = \zeta_i\right) + \sum_{i=0}^{R-1} P\left(Z_2 = \zeta_i\right) P\left(Z_1 < \zeta_i\right) \end{aligned}\right\} \quad (5.14)$$

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

Eqn. (5.14) can be restated as:

$$AUC = \frac{1}{2} P\left(Z_1 = Z_2\right) + P\left(Z_1 < Z_2\right) \quad (5.15)$$

The Wilcoxon statistic was defined in Eqn. (5.6). It can be seen that the comparisons implied by the summations and the weighting implied by the kernel function are estimating the two probabilities in the expression for in Eqn. (5.15). Therefore, $AUC = W$.

5.8 Importance of Bamber's theorem

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

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

5.9 Discussion / Summary

In this chapter, a simple method for estimating the area under the ROC plot has been described. The empirical AUC is a non-parametric measure of performance. Its simplicity and clear physical interpretation as the AUC under the empirical ROC (not fitted, not true) has spurred much theoretical development. These include the De Long et al method for estimating the variance of AUC of a single ROC empirical curve, and comparing pairs of ROC empirical curves5. Bamber's theorem, namely the equivalence between the empirical AUC and the Wilcoxon statistic has been derived and demonstrated.

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

5.10 Appendix 5.A: Details of Wilcoxon theorem

5.10.1 Upper triangle

For $i = 0$, Eqn. (5.13) implies (since the lowest empirical threshold is unity, the lowest allowed rating, and there are no cases rated less than one):

$$\left.\begin{aligned} A_0 =& P(Z_2 = 1)\left[\frac{1}{2}P(Z_1 = 1) + P(Z_1 < 1)\right] \\ A_0 =& \frac{1}{2}P(Z_1 = 1)\,P(Z_2 = 1) \end{aligned}\right\} \tag{5.16}$$

The base of the triangle is:

$$1 - P(Z_1 \geq 2) = P(Z_1 < 2) = P(Z_1 = 1) \tag{5.17}$$

The height of the triangle is:

$$1 - P(Z_2 \geq 2) = P(Z_2 < 2) = P(Z_2 = 1) \tag{5.18}$$

Q.E.D.

5.10.2 Lowest trapezoid

For $i = 4$, Eqn. (5.13) implies:

$$\left.\begin{aligned} A_4 =& P(Z_2 = 5)\left[\frac{1}{2}P(Z_1 = 5) + P(Z_1 < 5)\right] \\ A_4 =& \frac{1}{2}P(Z_2 = 5)\left[P(Z_1 = 5) + 2P(Z_1 < 5)\right] \\ A_4 =& \frac{1}{2}P(Z_2 = 5)\left[P(Z_1 = 5) + P(Z_1 < 5) + P(Z_1 < 5)\right] \\ A_4 =& \frac{1}{2}P(Z_2 = 5)\left[1 + P(Z_1 < 5)\right] \end{aligned}\right\} \tag{5.19}$$

The upper side of the trapezoid is

$$1 - P(Z_1 \geq 5) = P(Z_1 < 5) \tag{5.20}$$

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

$$\frac{1 + P(Z_1 < 5)}{2} \tag{5.21}$$

The height is:

$$P(Z_2 \geq 5) = P(Z_2 = 5) \tag{5.22}$$

Multiplication of the last two expressions yields A_4.

5.11 References

Chapter 6

Binormal model

6.1 TBA How much finished

70%

6.2 TBA Introduction

The equal variance binormal model was described in TBA Chapter 02. The ratings method of acquiring ROC data and calculation of operating points was discussed in TBA Chapter 04. It was shown, TBA Fig. 4.1, that for a clinical dataset the unequal-variance binormal model visually fitted the data better than the equal-variance binormal model, although how the unequal variance fit was obtained was not discussed. This chapter deals with details of the unequal-variance binormal model, often abbreviated to **binormal model**, establishes necessary notation, and derives expressions for sensitivity, specificity and the area under the predicted ROC curve).

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

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

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

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

6.3 Binormal model

6.3.1 The basic model

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

$$Z_{k_t t} \sim N\left(\mu_t, \sigma_t^2\right); t = 1, 2 \tag{6.1}$$

where

$$\left.\begin{array}{l} \mu_1 = 0 \\ \mu_2 = \mu \\ \sigma_1^2 = 1 \\ \sigma_2^2 = \sigma^2 \end{array}\right\} \tag{6.2}$$

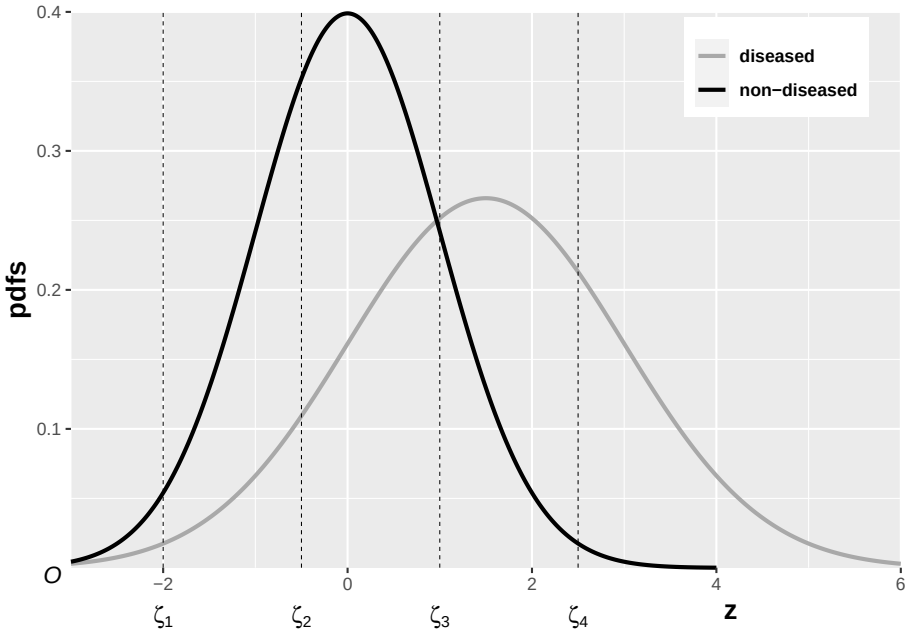
Eqn. (6.1) states that the z-samples for non-diseased cases are distributed as a $N(0, 1)$ distribution, i.e., the unit normal distribution, while the z-samples for

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

6.3.2 Additional parameters for binned data

In an R-rating ROC study the observed ratings r take on integer values, 1 through R, it being understood that higher ratings correspond to greater confidence for disease. Defining dummy cutoffs $\zeta_0 = -\infty$ and $\zeta_R = +\infty$, the binning rule for a case with realized z-sample z is (Chapter 4, Eqn. (4.2)):

$$\text{if}\,(\zeta_{r-1} \leq z \leq \zeta_r) \Rightarrow \text{rating} = r \tag{6.3}$$



[1] A more complicated version of this model allows the mean of the non-diseased distribution to be non-zero and its variance different from unity. The 4-parameter model is no more general than the 2-parameter model. The reason is that one is free to transform the decision variable, and associated thresholds, by applying arbitrary monotonic increasing function transformation, which do not change the ordering of the ratings and hence do not change the ROC curve. So if the mean of the noise distribution were non-zero, subtracting this value from all Z-samples would shift the effective mean of the non-diseased distribution to zero (the shifted Z-values are monotonically related to the original values) and the mean of the shifted diseased distribution becomes $\mu_2 - \mu_1$. Next, one scales or divides (division by a positive number is also a monotonic transformation) all the Z-samples by σ_1, resulting in the scaled non-diseased distribution having unit variance, and the scaled diseased distribution has mean $\frac{\mu_2 - \mu_1}{\sigma_1}$ and variance $(\frac{\sigma_2}{\sigma_1})^2$. Therefore, if one starts with 4 parameters then one can, by simple shifting and scaling operations, reduce the model to 2 parameters, as in Eqn. (6.1). [The author has seen a publication on Bayesian ROC estimation using the four-parameter model.]

In the unequal-variance binormal model, the variance σ^2 of the z-samples for diseased cases is allowed to be different from unity. Most ROC datasets are consistent with $\sigma > 1$. The above figure, generated with $\mu = 1.5, \sigma = 1.5, \zeta_1 = -2, \zeta_2 = -0.5, \zeta_3 = 1, \zeta_4 = 2.5$, illustrates how realized z-samples are converted to ratings, i.e., application of the binning rule (6.3). For example, a case with z-sample equal to -2.5 would be rated "1", and one with z-sample equal to -1 would be rated "2", cases with z-samples greater than 2.5 would be rated "5", etc.

6.3.3 Sensitivity and specificity

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

$$Sp(\zeta) = P(Z_1 < \zeta) = \Phi(\zeta) \tag{6.4}$$

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

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

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

$$Se(\zeta|\mu,\sigma) = P(Z_2 > \zeta) = P\left(\frac{Z_2 - \mu}{\sigma} > \frac{\zeta - \mu}{\sigma}\right) \tag{6.5}$$

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

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

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

$$Sp(\zeta) = \Phi(\zeta)\, Se(\zeta|\mu,\sigma) = \Phi\left(\frac{\mu - \zeta}{\sigma}\right) \tag{6.6}$$

The coordinates of the operating point defined by ζ are given by:

$$FPF(\zeta) = 1 - Sp(\zeta) = 1 - \Phi(\zeta) = \Phi(-\zeta) \tag{6.7}$$

$$TPF(\zeta|\mu,\sigma) = \Phi\left(\frac{\mu-\zeta}{\sigma}\right) \tag{6.8}$$

These expressions allow calculation of the operating point for any ζ. An equation for a curve is usually expressed as $y = f(x)$. An expression of this form for the ROC curve, i.e., the y-coordinate (TPF) expressed as a function of the x-coordinate (FPF), follows upon inversion of the expression for FPF, Eqn. (6.7):

$$\zeta = -\Phi^{-1}(FPF) \tag{6.9}$$

Substitution of Eqn. (6.9) in Eqn. (6.8) yields:

$$TPF = \Phi\left(\frac{\mu + \Phi^{-1}(FPF)}{\sigma}\right) \tag{6.10}$$

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

6.3.4 Binormal model in conventional notation

The following notation is widely used in the literature:

$$a = \frac{\mu}{\sigma}; b = \frac{1}{\sigma} \tag{6.11}$$

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

By dividing the z-samples by b, the variance of the distribution labeled "Noise" becomes unity, its mean stays at zero, and the variance of the distribution labeled "Signal" becomes $1/b$, and its mean becomes a/b, as shown below. It illustrates that the inverses of Eqn. (6.11) are:

$$\mu = \frac{a}{b}; \sigma = \frac{1}{b} \tag{6.12}$$

Eqns. (6.11) and (6.12) allow conversion from one notation to another.

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

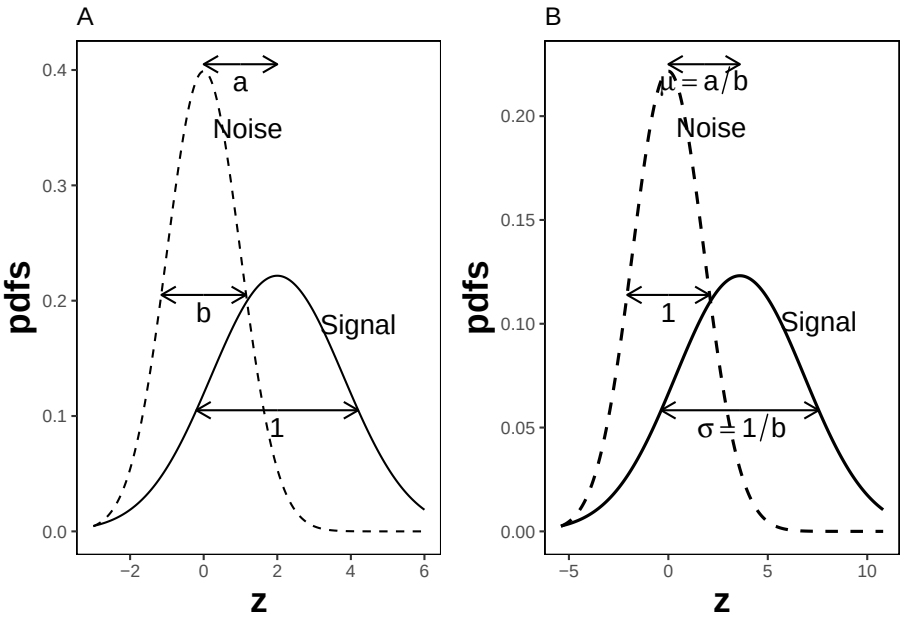


Figure 6.1: Plot A shows the definitions of the (a,b) parameters of the binormal model. In plot B the x-axis has been rescaled so that the noise distribution has unit variance, thereby illustrations between (a,b) and the (μ, σ) parameters.

6.4 Binormal ROC curve

Using the (a, b) notation, Eqn. (6.10) for the ROC curve reduces to:

$$TPF = \Phi\left(a + b\Phi^{-1}\left(FPF\right)\right) \tag{6.13}$$

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

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

From Eqn. (6.7) and Eqn. (6.8), the expressions for FPF and TPF in terms of model parameters (a, b) are:

$$\left.\begin{aligned} FPF\,(\zeta) &= \Phi\,(-\zeta) \\ TPF &= \Phi\,(a - b\zeta) \end{aligned}\right\} \tag{6.14}$$

6.5 Scalar threshold-independent measure

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

Generally accepted measures are the partial-area $A_{z;c}$ under the ROC, Eqn. (6.15), the full-area A_z under the ROC, Eqn. (6.18), and the d' index Eqn. (6.19).

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

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

6.5.1 Partial AUC

While this is a scalar measure, it does depend on choice of operating point. It is included here as it yields, as a special case, a scalar measure that does not depend on choice of operating point. The details are in Section 6.12.1, which

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

$$A_{z;c} = \int_{z_2=-\infty}^{\Phi^{-1}(c)} \int_{z_1=-\infty}^{\frac{a}{\sqrt{1+b^2}}} \phi\left(z_1, z_2; \rho\right) dz_1 dz_2 \tag{6.15}$$

The threshold ζ_1 corresponding to $FPF = c$ is given by:

$$\zeta_1 = -\Phi^{-1}\left(c\right) \tag{6.16}$$

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

$$\rho = -\frac{b}{\sqrt{1+b^2}} \tag{6.17}$$

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

```
a <- 2;b <- 1; zeta1 <- 1.5
A_z <- pnorm(a/sqrt(1+b^2))
opPtx <- pnorm(-zeta1)
opPty <- pnorm(a - b * zeta1)
rho <- -b/sqrt(1+b^2)
Lower1 <- -Inf
Upper1 <- qnorm(opPtx)
Lower2 <- -Inf
Upper2 <- a/sqrt(1+b^2)
sigma <- rbind(c(1, rho), c(rho, 1))
A_zc <- as.numeric(pmvnorm(
  c(Lower1, Lower2),
  c(Upper1, Upper2),
  sigma = sigma))
```

The partial-area measure is $A_{z;c} = 0.0352195$. The corresponding full-area measure is $A_z = 0.9213504$. $A_{z;c}$ is small because the reporting threshold is high. However, $A_{z;c}$ should not be confused with true performance of the observer, as shown in Section 6.6.

6.5.2 Full AUC

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

$$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right) \tag{6.18}$$

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

6.5.3 The d' measure

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

$$d' = \sqrt{2}\Phi^{-1}(A_z) \tag{6.19}$$

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

6.6 Partial AUC vs. true performance

- A *partial-area observer* such as in Section 6.5.1 rates cases as follows: for the sub-set of cases defined by $z \geq \zeta_1$ the observer reports *explicit* ratings exactly equal to the observed z-samples (or some monotonic transformation of the z-samples). For the remaining cases the observer assigns a *fixed value rating that is smaller than* ζ_1 (the exact value does not matter; these cases are said to be assigned *implicit* ratings).

- In contrast, the *full-area observer* reports explicit ratings *for all cases.*

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

$$A_{z;c,\text{TRUE}} = A_{z;c} + \frac{(1-FPF)\,(1+TPF)}{2} \tag{6.20}$$

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

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

$$A_{z;c,\text{TRUE}} \leq A_z \tag{6.21}$$

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

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

The following code saves 10,000 pairs of ratings in two arrays: `z[1,]` and `z[2,]`. The first array corresponds to non-diseased cases and the second to diseased cases. Note the usage, at lines 3-4, of the a, b values to define the two distributions. The array `zc`, initially a copy of `z`, is selectively binned by setting, lines 6-7, all ratings less than ζ_1 to -100. The ordering information for these z-samples is lost.

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

```
zc[1,z[1,] < zeta1] <- -100 # ratings of partial area observer
zc[2,z[2,] < zeta1] <- -100 # do:
```

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

```
#> A_z predicted =  0.9213504
#> A_z observed =  0.9232912
#> A_z{c;true} predicted =  0.8244498
#> A_z{c;true} observed =  0.8244189
```

Note that:

- $A_{z;c,\mathrm{TRUE}} < A_z$, because ordering information is lost for all cases with z-samples less than ζ_1.
- $A_{z;c,\mathrm{TRUE}} >> A_{z;c}$, because of the large contribution from the area under the straight line, left poanel Fig. 6.2.

6.7 Illustrative plots

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

The ROC curve for both types of observers is shown in the left panel of 6.2 for the following parameters: $a = 2$, $b = 1$ and $\zeta_1 = 1.5$; ζ_1 corresponds to $c \equiv FPF = \Phi(-\zeta_1) = 0.0668072$ and $TPF = \Phi(a - b\zeta_1) = 0.6914625$. In other words the limiting point coordinates are (0.067, 0.691), shown in the plot by the solid dot. Partial AUC $A_{z;c}$ equals 0.0352195. The full-area ROC curve, shown by the complete solid curve, extends from (0,0) to (1,1), the area under which is $A_z = 0.9213504$.

As FPF increases true-performance increases. the right panel of Fig. 6.2 shows the variation of true performance $A_{z;c,\mathrm{TRUE}}$ with FPF. The curve starts from (0, 0.5) and ends at (1.000, 0.921). For low values of FPF the curve is very steep while for FPF > 0.25 the curve levels out, approaching the maximum value defined by $A_z = 0.9213504$. True performance is maximized at $\zeta_1 = -\infty$.

Fig. 6.3, left panel, corresponding to $a = 1$, $b = 0.2$ and $\zeta_1 = 1.5$, shows an improper ROC curve. The dashed line is well above the continuous curve

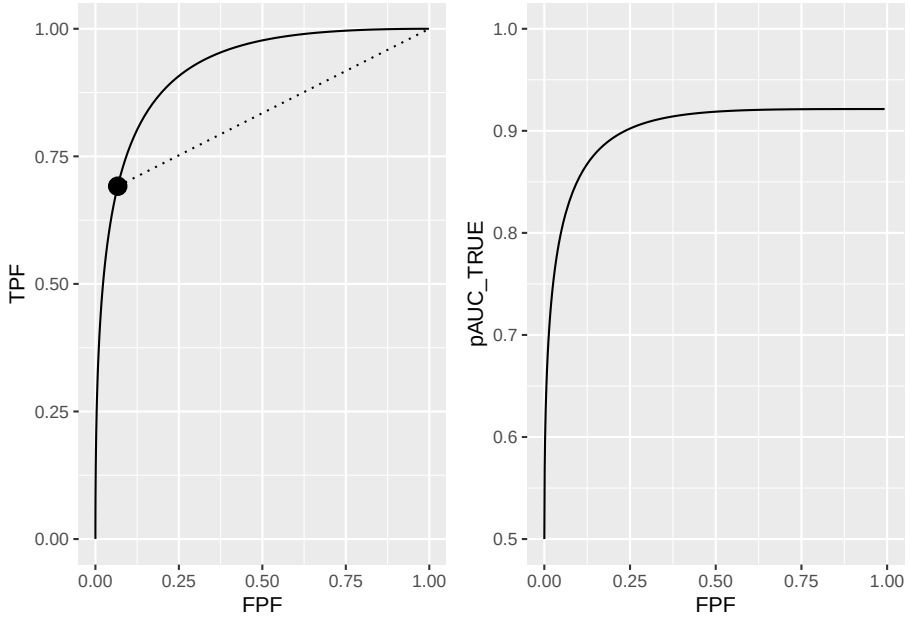


Figure 6.2: Left panel: binormal ROC curve corresponding to a = 2 and b = 1. The dot is the operating point corresponding to $\zeta_1 = 1.5$. The continuous curve extending from the origin to (1,1) represents the full ROC. Note that in the region above the dot the continuus curve is above the dotted line, meaning true performance of an observer who only rates a sub-set of cases is less than performance of an observer who rates all cases. Right panel: variation of true performance with FPF; at FPF = 0 the plot starts at ordinate equal to 0.5 and levels out at FPF = 1 at AUC = $A_z = 0.921$.

and true performance is maximized at a finite value of ζ_1, corresponding to $FPF = 0.153$, see right panel. This is an invalid conclusion since an improper ROC curve is a fitting artifact of the binormal model easily avoided by using modern curve-fitting methods (eg., PROPROC, CBM or RSM). TBA However, since the wAFROC has an operating characteristic with an improper-like feature but which is not a fitting artifact, this example serves a purpose, elaborated on in TBA Chapter (optim-op-point), where it is shown that by maximizing the area under the wAFROC one can find the optimal threshold of an algorithmic observer.

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

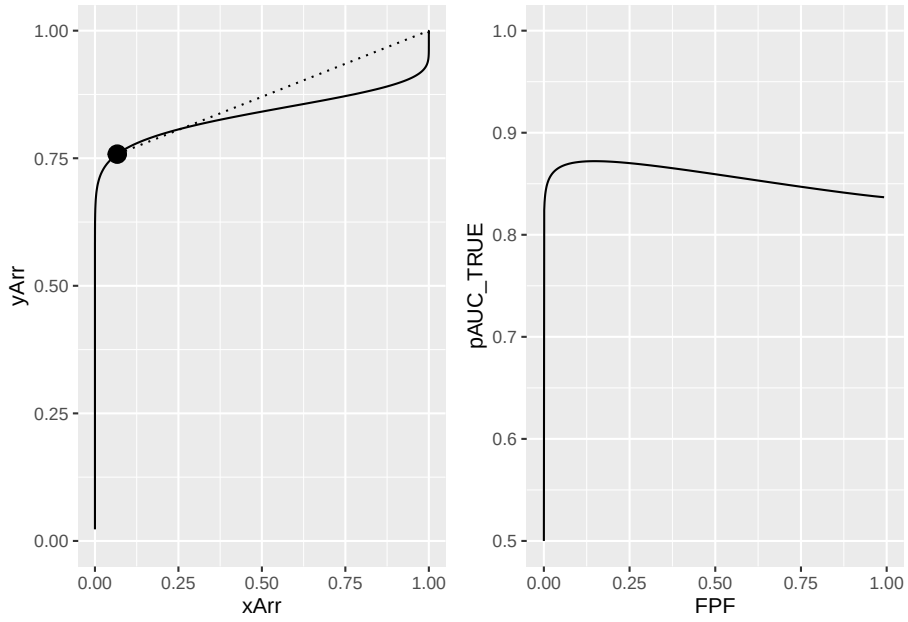


Figure 6.3: The left panel shows the visibly improper ROC curve for a = 1 and b = 0.2. The solid line is below the dotted line. The right panel shows the variation of true performance pAUC_TRUE with FPF. True performance is maximized at FPF = 0.153. Since improper ROC fits are fitting artifacts, this example does not negate the previous finding that true performance for a proper ROC curve is maximized by setting the threshold to report all cases, i.e., FPF = 1.

6.8 Geometrical argument

Defining geometrical features of a proper ROC are:

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

The geometry ensures that true performance for a proper ROC is maximized at $\zeta_1 = -\infty$, i.e., at FPF = 1, as in Fig. 6.2, right panel.

6.9 Optimal operating point on ROC

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

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

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

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

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

$$\overline{C} = C_0 + C_{\mathrm{TP}}P(\mathrm{TP}) + C_{\mathrm{TN}}P(\mathrm{TN}) + C_{\mathrm{FP}}P(\mathrm{FP}) + C_{\mathrm{FN}}P(\mathrm{FN}) \tag{6.22}$$

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

$$\left.\begin{array}{l} P(\mathrm{TP}) = P(+)\mathrm{TPF} \\ P(\mathrm{TN}) = (1-P(+))(1-\mathrm{FPF}) \\ P(\mathrm{FP}) = (1-P(+))\mathrm{FPF} \\ P(\mathrm{FN}) = P(+)(1-\mathrm{TPF}) \end{array}\right\} \tag{6.23}$$

With these substitutions one gets for the average cost:

$$\overline{C} = C_0 + C_{\mathrm{TN}}P(-) + C_{\mathrm{FN}}P(+) + (C_{\mathrm{TP}} - C_{\mathrm{FN}})P(+)\mathrm{TPF} + (C_{\mathrm{FP}} - C_{\mathrm{TN}})P(-)\mathrm{FPF} \tag{6.24}$$

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

$$\frac{d(\mathrm{TPF})}{d(\mathrm{FPF})} = \frac{C_{\mathrm{FP}} - C_{\mathrm{TN}}}{C_{\mathrm{FN}} - C_{\mathrm{TP}}} \frac{P(-)}{P(+)} \tag{6.25}$$

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

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

The costs and benefits are often difficult to quantify. If one assumes that the right hand side of Eqn. (6.25) equals unity (e.g., the four costs / benefits are equal and disease-prevalence is 50%) then the optimal operating point is defined by that point on the ROC curve where the slope is unity, which is the point of nearest approach of the curve to the upper-left corner. This corresponds to maximizing the Youden index (Youden, 1950), defined as the sum of sensitivity and specificity minus one. This is demonstrated in the following code.

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

6.10 Discussion

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

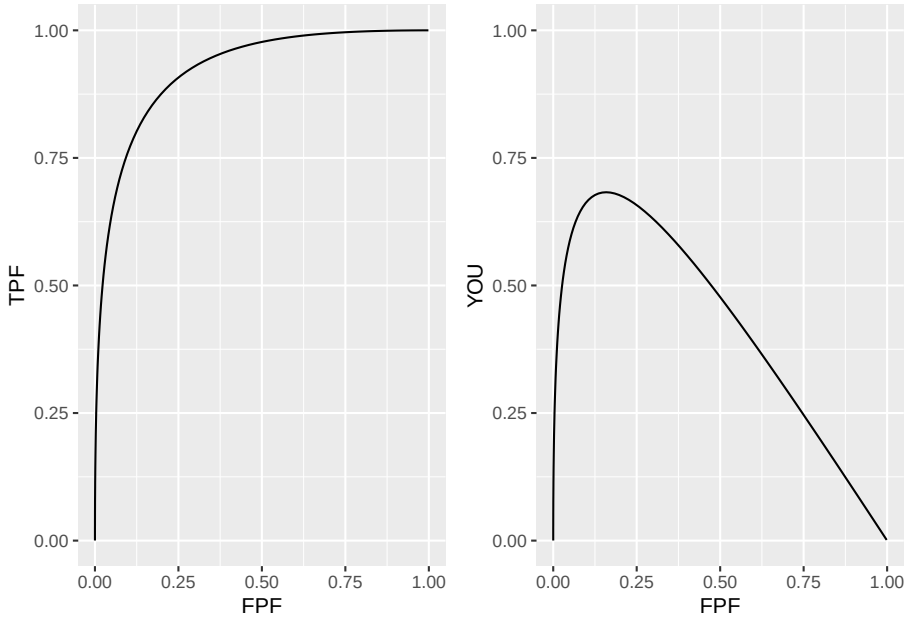


Figure 6.4: Left panel: binormal ROC curve corresponding to a = 2 and b = 1. Right panel: variation of Youden index with FPF; the plot shows a maximum at FPF = 0.1586553; this corresponds to the nearest approch of the ROC curve to the upper-left corner.

when in fact it is quite robust to apparent deviations from normality, is addressed.

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

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

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

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

The unphysical nature of the hook (predicting worse than chance-level performance for supposedly expert readers) is not the only reason for seeking alternate ROC models. The binormal model is susceptible to degeneracy problems. If the dataset does not provide any interior operating points (i.e., all observed points lie on the axes defined by FPF = 0 or TPF = 1) then the model fits these points with b = 0. The resulting straight-line segment fits do not make physical sense. These problems are addressed by the contaminated binormal model16 to be discussed in Chapter "Other proper ROC models". The first paper in the series has particularly readable accounts of data degeneracy.

To this day the binormal model is widely used to fit ROC datasets. In spite of

its limitations, the binormal model has been very useful in bringing a level of quantification to this field that did not exist prior to (Dorfman and Alf, 1969).

6.11 Appendix I: Density functions

According to Eqn. (6.1) the probability that a z-sample is smaller than a specified threshold ζ, i.e., the CDF function, is:

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

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

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

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

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

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

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

6.12 Appendix II: Area under binormal ROC

6.12.1 General case (partial-area)

This section is based on (Thompson and Zucchini, 1989). In what follows, FPF is abbreviates to x and TPF to y. Then the equation for the ROC curve is (6.13):

$$y = \Phi\left(a + b\Phi^{-1}\left(x\right)\right) \tag{6.26}$$

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

$$A_{z;c} = \int_0^c y dx = \int_0^c dx \Phi\left(a + b\Phi^{-1}\left(x\right)\right) \tag{6.27}$$

Define change of variable:

$$x = \Phi\left(x_1\right) \tag{6.28}$$

which implies:

$$\left.\begin{array}{l} x_1 = \Phi^{-1}\left(x\right) \\ dx = dx_1 \phi\left(x_1\right) \end{array}\right\} \tag{6.29}$$

This yields:

$$\left.\begin{array}{rl} A_{z;c} = & \displaystyle\int_0^c dx \Phi\left(a + b x_1\right) \\ = & \displaystyle\int_{-\infty}^{\Phi^{-1}(c)} dx_1 \phi\left(x_1\right) \Phi\left(a + b x_1\right) \end{array}\right\} \tag{6.30}$$

The right hand side of Eqn. (6.30) can be expressed as an integral over the bivariate normal distribution as follows. From the definition of the Φ function the above integral can be written as the following double integral:

$$A_{z;c} = \int_{x_1=-\infty}^{\Phi^{-1}(c)} dx_1 \phi\left(x_1\right) \int_{x_2=-\infty}^{a+bx_1} \phi\left(x_2\right) dx_2 \tag{6.31}$$

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

$$\left.\begin{array}{l} z_2 = x_1 \\ z_1 = \left(x_2 - b x_1\right) f \end{array}\right\} \tag{6.32}$$

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

$$x_2 = \frac{z_1}{f} + b x_1 = \frac{z_1}{f} + b z_2 \tag{6.33}$$

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

$$J = \begin{pmatrix} 0 & 1 \\ \frac{1}{f} & b \end{pmatrix} \tag{6.34}$$

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

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

$$A_{z;c} = \frac{1}{f} \int_{z_2=-\infty}^{\Phi^{-1}(c)} dz_2 \phi(z_2) \int_{z_1=-\infty}^{z_1^{UL}} \phi\left(\frac{z_1}{f} + bz_2\right) dz_1 \tag{6.35}$$

The upper limit of the inner integral can be calculated as follows. Using the second equation in Eqn. (6.32):

$$z_1^{UL} = (x_2^{UL} - bx_1) f = (a + bx_1 - bx_1) f = af \tag{6.36}$$

Eqn. (6.35) simplifies to:

$$A_{z;c} = \frac{1}{f} \int_{z_2=-\infty}^{\Phi^{-1}(c)} dz_2 \phi(z_2) \int_{z_1=-\infty}^{af} \phi\left(\frac{z_1}{f} + bz_2\right) dz_1 \tag{6.37}$$

Perform a change of variable from f to a correlation-like quantity ρ defined by:

$$f = \sqrt{1 - \rho^2} \tag{6.38}$$

Define ρ in terms of the b-parameter as follows:

$$b\sqrt{1 - \rho^2} = -\rho \tag{6.39}$$

This implies that ρ is given by:

$$\rho = -\frac{b}{\sqrt{1 + b^2}} \tag{6.40}$$

The argument of the right-most ϕ function in Eqn. (6.37) simplifies as follows:

$$\frac{z_1}{f} + bz_2 = \frac{z_1 + bz_2\sqrt{1 - \rho^2}}{\sqrt{1 - \rho^2}} = \frac{z_1 - \rho z_2}{\sqrt{1 - \rho^2}} \tag{6.41}$$

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

$$A_{z;c} = \frac{1}{\sqrt{1 - \rho^2}} \int_{z_2=-\infty}^{\Phi^{-1}(c)} dz_2 \phi(z_2) \int_{z_1=-\infty}^{a\sqrt{1-\rho^2}} \phi\left(\frac{z_1 - \rho z_2}{\sqrt{1 - \rho^2}}\right) dz_1 \tag{6.42}$$

Eqn. (6.39) implies:

$$1-\rho^2=\frac{1}{1+b^2} \tag{6.43}$$

Therefore,

$$A_{z;c}=\frac{1}{\sqrt{1-\rho^2}}\int_{z_2=-\infty}^{\Phi^{-1}(c)}dz_2\phi\left(z_2\right)\int_{z_1=-\infty}^{\frac{a}{\sqrt{1+b^2}}}\phi\left(\frac{z_1-\rho z_2}{\sqrt{1-\rho^2}}\right)dz_1 \tag{6.44}$$

The standard bivariate normal distribution with correlation coefficient ρ is defined by:

$$\phi\left(z_1,z_2;\rho\right)=\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{z_1^2-2\rho z_1z_2+z_2^2}{2(1-\rho^2)}\right) \tag{6.45}$$

The standard normal distribution is defined by:

$$\phi\left(z\right)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right) \tag{6.46}$$

It can be shown using these definitions that:

$$\phi\left(z_1,z_2;\rho\right)=\frac{1}{\sqrt{1-\rho^2}}\phi\left(z_2\right)\phi\left(\frac{z_1-\rho z_2}{\sqrt{1-\rho^2}}\right) \tag{6.47}$$

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

$$A_{z;c}=\int_{z_2=-\infty}^{\Phi^{-1}(c)}\int_{z_1=-\infty}^{\frac{a}{\sqrt{1+b^2}}}\phi\left(z_1,z_2;\rho\right)dz_1dz_2 \tag{6.48}$$

6.12.2 Special case (total area)

Since c is the upper limit of FPF, setting $c=1$ yields the total area under the binormal ROC curve [2]:

[2]Since the integral over z_2 is over the entire range it integrates out to unity leaving the one-dimensional density function $\phi\left(z_1\right)$ inside the integral. The last step follows from the definition of the Φ function.

$$\left.\begin{aligned} A_z &= \int_{z_2=-\infty}^{\infty} \int_{z_1=-\infty}^{\frac{a}{\sqrt{1+b^2}}} \phi\left(z_1, z_2; \rho\right) dz_1 dz_2 \\ &= \int_{z_1=-\infty}^{\frac{a}{\sqrt{1+b^2}}} \phi\left(z_1\right) dz_1 \\ &= \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \end{aligned}\right\} \tag{6.49}$$

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

$$\left.\begin{aligned} A_z &= \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \\ &= \Phi\left(\frac{\frac{a}{b}}{\sqrt{1+\frac{1}{b^2}}}\right) \\ &= \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right) \end{aligned}\right\} \tag{6.50}$$

6.13 Appendix III: Invariance property of pdfs

The binormal model is not as restrictive as might appear at first sight. Any monotone increasing transformation $Y = f(Z)$ applied to the observed z-samples, and the associated thresholds, will yield the same observed data, e.g., Table 4.1. This is because such a transformation leaves the ordering of the ratings unaltered and hence results in the same operating points. While the distributions for Y will not be binormal (i.e., two independent normal distributions), one can safely "pretend" that one is still dealing with an underlying binormal model. An alternative way of stating this is that any pair of distributions is allowed as long as they are reducible to a binormal model form by a monotonic increasing transformation of Y: e.g., $Z = f^{-1}$. [If f is a monotone increasing function of its argument, so is f^{-1}}.] For this reason, the term "pair of latent underlying normal distributions" is sometimes used to describe the binormal model. The robustness of the binormal model has been investigated (Hanley, 1988; Dorfman et al., 1997). The referenced paper by Dorfman et al has an excellent discussion of the robustness of the binormal model.

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

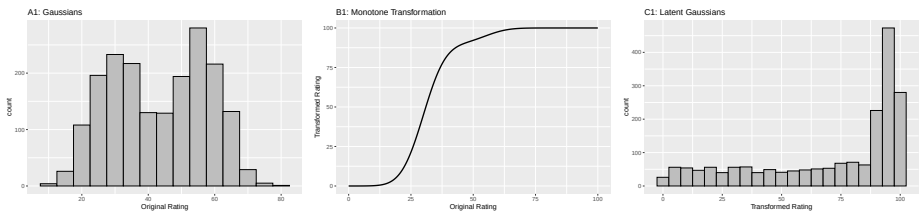
conclusions that the binormal model is inapplicable to these datasets. To quote a reviewer of one of my recent papers:

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

The reviewer is correct. The misconception is illustrated next.

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

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



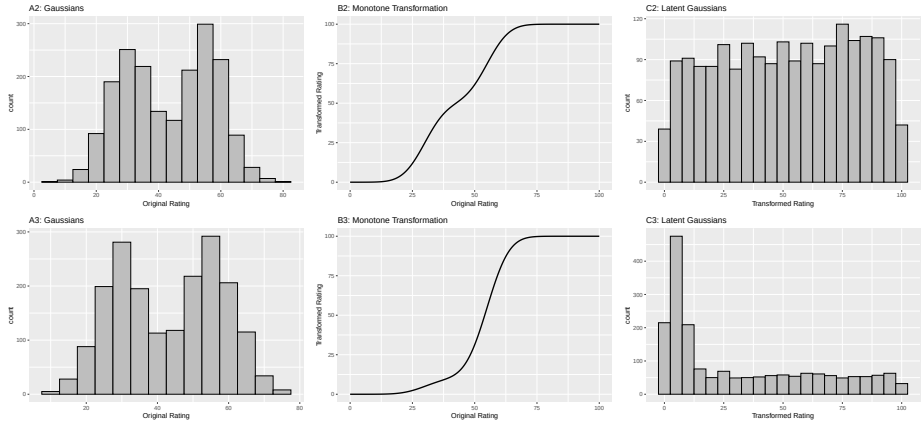


Figure captions (A1 - C3): Illustrating the invariance of ROC analysis to arbitrary monotone transformations of the ratings. Each row contains 3 plots: labeled 1, 2 and 3. Each column contains 3 plots labeled A, B and C. So, for example, plot C2 refers to the second row and third column. The for-loop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

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

$$Y(Z) = \left[(1-f)\,\Phi\left(\frac{Z-\mu_1}{\sigma_1}\right) + f\Phi\left(\frac{Z-\mu_2}{\sigma_2}\right)\right] 100 \qquad (6.51)$$

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

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

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

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

6.14 Appendix IV: Fitting an ROC curve

6.14.1 JAVA fitted ROC curve

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

6.14.2 Simplistic straight line fit to the ROC curve

To be described next is a method for fitting data such as in Table 4.1 to the binormal model, i.e., determining the parameters (a, b) and the thresholds $\zeta_r, \quad r = 1, 2, ..., R-1$, to best fit, in some to-be-defined sense, the observed cell counts. The most common method uses an algorithm called maximum likelihood. But before getting to that, I describe the least-square method, which is conceptually simpler, but not really applicable, as will be explained shortly.

6.14.2.1 Least-squares estimation

By applying the function Φ^{-1} to both sides of Eqn. (6.10), one gets (the "inverse" function cancels the "forward" function on the right hand side):

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

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

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

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

for a and b. If the reader has never done this before, one should go through these steps at least once, but it would be smarter in future to use software that does all this. In `R` the least-squares fitting function is `lm(y~x)`, which in its simplest form fits a linear model `lm(y~x)` using the method of least-squares (in case you are wondering `lm` stands for linear model, a whole branch of statistics in itself; in this example one is using its simplest capability).

```
# ML estimates of a and b (from Eng JAVA program)
# a <- 1.3204; b <- 0.6075
# # these are not used in program; just here for comparison

FPF <- c(0.017, 0.050, 0.183, 0.5)
# this is from Table 6.11, last two rows
TPF <- c(0.440, 0.680, 0.780, 0.900)
# ...do...

PhiInvFPF <- qnorm(FPF)
# apply the PHI_INV function
PhiInvTPF <- qnorm(TPF)
# ... do ...

fit <- lm(PhiInvTPF~PhiInvFPF)
print(fit)
#>
#> Call:
#> lm(formula = PhiInvTPF ~ PhiInvFPF)
#>
#> Coefficients:
#> (Intercept)    PhiInvFPF
#>    1.328844     0.630746
```

Fig. 6.5 shows operating points from Table 4.1, transformed by the Φ^{-1} function; the slope of the line is the least-squares estimate of the b parameter and the intercept is the corresponding a parameter of the binormal model.

The last line contains the least squares estimated values, $a = 1.3288$ and $b = 0.6307$. The corresponding maximum likelihood estimates of these parameters, as yielded by the Eng web code, Appendix B, are listed in line 4 of the main program: $a = 1.3204$ and $b = 0.6075$. The estimates appear to be close, particularly the estimate of a , but there are a few things wrong with the least-squares approach. First, the method of least squares assumes that the data points are independent. Because of the manner in which they are constructed, namely by cumulating points, the independence assumption is not valid for ROC operating points. Cumulating the 4 and 5 responses constrains the resulting operating point to be above and to the right of the point obtained by cumulating the 5 responses only, so the data points are definitely not independent. Similarly,

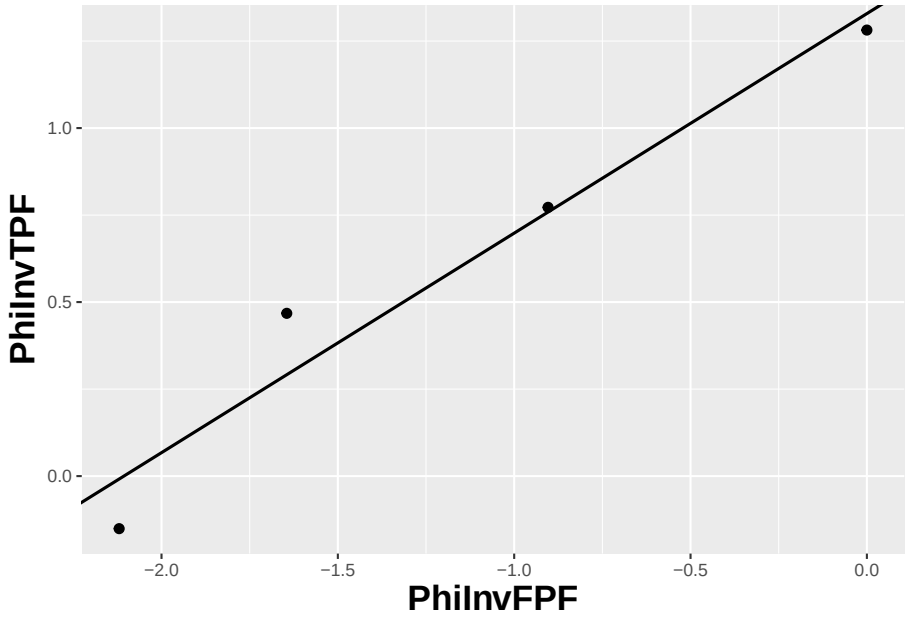


Figure 6.5: The straight line fit method of estimating parameters of the fitting model.

cumulating the 3, 4 and 5 responses constrains the resulting operating point to be above and to the right of the point obtained by cumulating the 4 and 5 responses, and so on. The second problem is the linear least-squares method assumes there is no error in measuring x; the only source of error that is accounted for is in the y-coordinate. In fact, both coordinates of an ROC operating point are subject to sampling error. Third, disregard of error in the x-direction is further implicit in the estimates of the thresholds, which according to Eqn. (6.2.19), is given by:

$$\zeta_r = -\Phi^{-1}\left(FPF_r\right)$$

These are "rigid" estimates that assume no error in the FPF values. As was shown in Chapter 2, 95% confidence intervals apply to these estimates.

A historical note: prior to computers and easy access to statistical functions the analyst had to use a special plotting paper, termed "double probability paper", that converted probabilities into x and y distances using the inverse function.

6.14.3 Maximum likelihood estimation (MLE)

The approach taken by Dorfman and Alf was to maximize the likelihood function instead of S. The likelihood function is the probability of the observed data given a set of parameter values, i.e.,

$$\mathrm{L} \equiv P\left(data \mid \text{parameters}\right)$$

Generally "data" is suppressed, so likelihood is a function of the parameters; but "data" is always implicit. With reference to Fig. 6.1, the probability of a non-diseased case yielding a count in the 2nd bin equals the area under the curve labeled "Noise" bounded by the vertical lines at ζ_1 and ζ_2. In general, the probability of a non-diseased case yielding a count in the r^{th} bin equals the area under the curve labeled "Noise" bounded by the vertical lines at ζ_{r-1} and ζ_r. Since the area to the left of a threshold is the CDF corresponding to that threshold, the required probability is $\Phi\left(\zeta_r\right) - \Phi\left(\zeta_{r-1}\right)$; we are simply subtracting two expressions for specificity, Eqn. (6.2.5).

$$\text{count in non-diseased bin } r = \Phi\left(\zeta_r\right) - \Phi\left(\zeta_{r-1}\right)$$

Similarly, the probability of a diseased case yielding a count in the rth bin equals the area under the curve labeled "Signal" bounded by the vertical lines at ζ_{r-1} and ζ_r. The area under the diseased distribution to the left of threshold ζ_r is the $1 - TPF$ at that threshold:

$$1 - \Phi\left(\frac{\mu - \zeta_r}{\sigma}\right) = \Phi\left(\frac{\zeta_r - \mu}{\sigma}\right)$$

The area between the two thresholds is:

$$\begin{aligned} P\left(\text{count in diseased bin } r\right) &= \Phi\left(\frac{\zeta_r - \mu}{\sigma}\right) - \Phi\left(\frac{\zeta_{r-1} - \mu}{\sigma}\right) \\ &= \Phi\left(b\zeta_r - a\right) - \Phi\left(b\zeta_{r-1} - a\right) \end{aligned}$$

Let K_{1r} denote the number of non-diseased cases in the rth bin, and K_{2r} denotes the number of diseased cases in the rth bin. Consider the number of counts K_{1r} in non-diseased case bin r. Since the probability of each count is $\Phi\left(\zeta_{r+1}\right) - \Phi\left(\zeta_r\right)$, the probability of the observed number of counts, assuming the counts are independent, is $\left(\Phi\left(\zeta_{r+1}\right) - \Phi\left(\zeta_r\right)\right)^{K_{1r}}$. Similarly, the probability of observing counts in diseased case bin r is $\left(\Phi\left(b\zeta_{r+1} - a\right) - \Phi\left(b\zeta_r - a\right)\right)^{K_{2r}}$, subject to the same independence assumption. The probability of simultaneously observing K_{1r} counts in non-diseased case bin r and K_{2r} counts in diseased case bin r is the product of these individual probabilities (again, an independence assumption is being used):

$$\left(\Phi\left(\zeta_{r+1}\right) - \Phi\left(\zeta_r\right)\right)^{K_{1r}} \left(\Phi\left(b\zeta_{r+1} - a\right) - \Phi\left(b\zeta_r - a\right)\right)^{K_{2r}}$$

Similar expressions apply for all integer values of r ranging from $1, 2, ..., R$. Therefore the probability of observing the entire data set is the product of expressions like Eqn. (6.4.5), over all values of r:

$$\prod_{r=1}^{R} \left[\left(\Phi\left(\zeta_{r+1}\right) - \Phi\left(\zeta_r\right)\right)^{K_{1r}} \left(\Phi\left(b\zeta_{r+1} - a\right) - \Phi\left(b\zeta_r - a\right)\right)^{K_{2r}}\right] \tag{6.52}$$

We are almost there. A specific combination of $K_{11}, K_{12}, ..., K_{1R}$ counts from K_1 non-diseased cases and counts $K_{21}, K_{22}, ..., K_{2R}$ from K_2 diseased cases can occur the following number of times (given by the multinomial factor shown below):

$$\frac{K_1!}{\prod_{r=1}^{R} K_{1r}!} \frac{K_2!}{\prod_{r=1}^{R} K_{2r}!} \tag{6.53}$$

The likelihood function is the product of Eqn. (6.52) and Eqn. (6.53):

$$\begin{aligned} L\left(a, b, \vec{\zeta}\right) = &\left(\frac{K_1!}{\prod_{r=1}^{R} K_{1r}!} \frac{K_2!}{\prod_{r=1}^{R} K_{2r}!}\right) \times \\ &\prod_{r=1}^{R} \left[\left(\Phi\left(\zeta_{r+1}\right) - \Phi\left(\zeta_r\right)\right)^{K_{1r}} \left(\Phi\left(b\zeta_{r+1} - a\right) - \Phi\left(b\zeta_r - a\right)\right)^{K_{2r}}\right] \end{aligned} \tag{6.54}$$

The left hand side of Eqn. (6.54) shows explicitly the dependence of the likelihood function on the parameters of the model, namely $a, b, \vec{\zeta}$, where the vector of thresholds $\vec{\zeta}$ is a compact notation for the set of thresholds $\zeta_1, \zeta_2, ..., \zeta_R$, (note that since $\zeta_0 = -\infty$, and $\zeta_R = +\infty$, only $R-1$ free threshold parameters are involved, and the total number of free parameters in the model is $R+1$). For example, for a 5-rating ROC study, the total number of free parameters is 6, i.e., a, b and 4 thresholds $\zeta_1, \zeta_2, \zeta_3, \zeta_4$.

Eqn. (6.54) is forbidding but here comes a simplification. The difference of probabilities such as $\Phi(\zeta_r) - \Phi(\zeta_{r-1})$ is guaranteed to be positive and less than one [the Φ function is a probability, i.e., in the range 0 to 1, and since ζ_r is greater than ζ_{r-1}, the difference is positive and less than one]. When the difference is raised to the power of K_{1r} (a non-negative integer) a very small number can result. Multiplication of all these small numbers may result in an even smaller number, which may be too small to be represented as a floating-point value, especially as the number of counts increases. To prevent this we resort to a trick. Instead of maximizing the likelihood function $L\left(a, b, \vec{\zeta}\right)$ we choose to maximize the logarithm of the likelihood function (the base of the logarithm is immaterial). The logarithm of the likelihood function is:

$$LL\left(a, b, \vec{\zeta}\right) = \log\left(L\left(a, b, \vec{\zeta}\right)\right) \tag{6.55}$$

Since the logarithm is a monotonically increasing function of its argument, maximizing the logarithm of the likelihood function is equivalent to maximizing the likelihood function. Taking the logarithm converts the product symbols in Eqn. (6.4.8) to summations, so instead of multiplying small numbers one is adding them, thereby avoiding underflow errors. Another simplification is that one can ignore the logarithm of the multinomial factor involving the factorials, because these do not depend on the parameters of the model. Putting all this together, we get the following expression for the logarithm of the likelihood function:

$$\begin{aligned} LL\left(a, b, \vec{\zeta}\right) \propto & \sum_{r=1}^{R} K_{1r} \log\left(\Phi\left(\zeta_{r+1}\right) - \Phi\left(\zeta_r\right)\right) \\ & + \sum_{r=1}^{R} K_{2r} \log\left(\Phi\left(b\zeta_{r+1} - a\right) - \Phi\left(b\zeta_r - a\right)\right) \end{aligned} \tag{6.56}$$

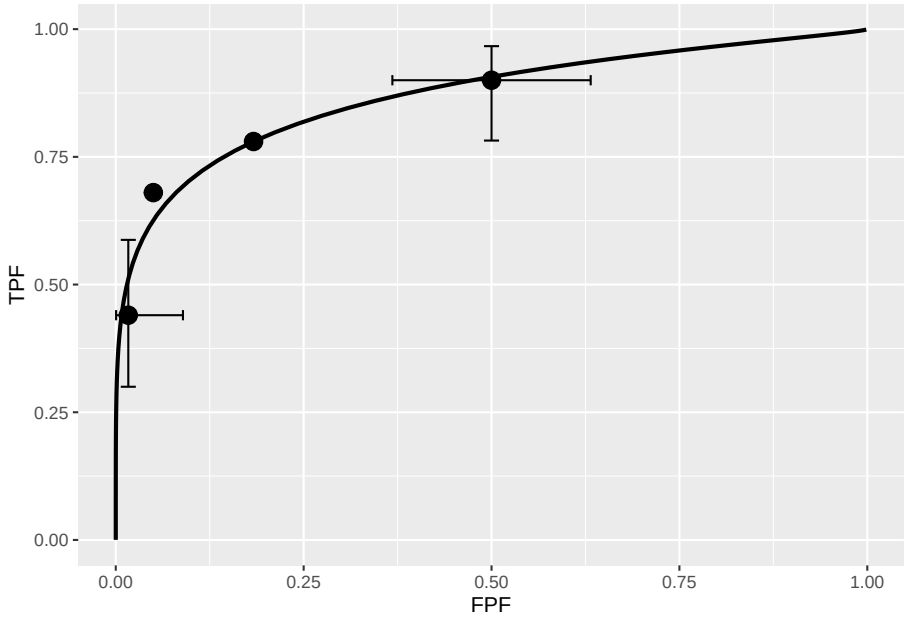
The left hand side of Eqn. (6.56) is a function of the model parameters $a, b, \vec{\zeta}$ and the observed data, the latter being the counts contained in the vectors $\overrightarrow{K_1}$ and $\overrightarrow{K_2}$, where the vector notation is used as a compact form for the counts $K_{11}, K_{12}, ..., K_{1R}$ and $K_{21}, K_{22}, ..., K_{2R}$, respectively. The right hand side of Eqn. (6.56) is monotonically related to the probability of observing the data given the model parameters $a, b, \vec{\zeta}$. If the choice of model parameters is poor, then the probability of observing the data will be small and log likelihood will

be small. With a better choice of model parameters the probability and log likelihood will increase. With optimal choice of model parameters the probability and log likelihood will be maximized, and the corresponding optimal values of the model parameters are called maximum likelihood estimates (MLEs). These are the estimates produced by the programs RSCORE and ROCFIT.

6.14.4 Code implementing MLE

```
# ML estimates of a and b (from Eng JAVA program)
# a <- 1.3204; b <- 0.6075
# these are not used in program; just there for comparison

K1t <- c(30, 19, 8, 2, 1)
K2t <- c(5,  6, 5, 12, 22)
dataset <- Df2RJafrocDataset(K1t, K2t, InputIsCountsTable = TRUE)
retFit <- FitBinormalRoc(dataset)
retFit[1:5]
#> $a
#> [1] 1.32045261
#>
#> $b
#> [1] 0.607492932
#>
#> $zetas
#>      zetaFwd1      zetaFwd2      zetaFwd3      zetaFwd4
#> 0.00768054675 0.89627306763 1.51564784976 2.39672209865
#>
#> $AUC
#> [1] 0.870452157
#>
#> $StdAUC
#>               [,1]
#> [1,] 0.0379042262
print(retFit$fittedPlot)
```



Note the usage of the `RJafroc` package (Chakraborty et al., 2020). Specifically, the function `FitBinormalRoc`. The ratings table is converted to an `RJafroc` dataset object, followed by application of the fitting function. The results, contained in `retFit` should be compared to those obtained from the website implementation of ROCFIT.

6.15 Appendix V: Validating fitting model

The above ROC curve is a good visual fit to the observed operating points. Quantification of the validity of the fitting model is accomplished by calculating the Pearson goodness-of-fit test (Pearson, 1900), also known as the chi-square test, which uses the statistic defined by (Larsen and Marx, 2001):

$$C^2 = \sum_{t=1}^{2}\sum_{r=1}^{R} \frac{\left(K_{tr} - \langle K_{tr}\rangle\right)^2}{\langle K_{tr}\rangle} K_{tr} \geq 5 \tag{6.57}$$

The expected values are given by:

$$\begin{aligned} \langle K_{1r}\rangle &= K_1\left(\Phi\left(\zeta_{r+1}\right) - \Phi\left(\zeta_r\right)\right) \\ \langle K_{2r}\rangle &= K_2\left(\Phi\left(a\zeta_{r+1} - b\right) - \Phi\left(a\zeta_r - b\right)\right) \end{aligned} \tag{6.58}$$

These expressions should make sense: the difference between the two CDF functions is the probability of a count in the specified bin, and multiplication by the total number of relevant cases should yield the expected counts (a non-integer).

It can be shown that under the null hypothesis that the assumed probability distribution functions for the counts equals the true probability distributions, i.e., the model is valid, the statistic C^2 is distributed as:

$$C^2 \sim \chi^2_{df} \tag{6.59}$$

Here $C^2 \sim \chi^2_{df}$ is the chi-square distribution with degrees of freedom df defined by:

$$df = (R-1) + (R-1) - (2 + R - 1) = (R-3) \tag{6.60}$$

The right hand side of the above equation has been written in an expansive form to illustrate the general rule: for R non-diseased cells in the ratings table, the degree of freedom is $R-1$: this is because when all but one cells are specified, the last is determined, because they must sum to K_1 . Similarly, the degree of freedom for the diseased cells is also $R-1$. Last, we need to subtract the number of free parameters in the model, which is $(2 + R - 1)$, i.e., the a, b parameters and the $R-1$ thresholds. It is evident that if $R = 3$ then $df = 0$. In this situation, there are only two non-trivial operating points and the straight-line fit shown will pass through both of them. With two basic parameters, fitting two points is trivial, and goodness of fit cannot be calculated.

Under the null hypothesis (i.e., model is valid) C^2 is distributed as χ^2_{df}. Therefore, one computes the probability that this statistic is larger than the observed value, called the *p-value*. If this probability is very small, that means that the deviations of the observed values of the cell counts from the expected values are so large that it is unlikely that the model is correct. The degree of unlikeliness is quantified by the p-value. Poor fits lead to small p values.

At the 5% significance level, one concludes that the fit is not good if $p < 0.05$. In practice one occasionally accepts smaller values of p, $p > 0.001$ before completely abandoning a model. It is known that adoption of a stricter criterion, e.g., $p > 0.05$, can occasionally lead to rejection of a retrospectively valid model (Press et al., 2007).

6.15.1 Estimating the covariance matrix

TBA See book chapter 6.4.3. This is implemented in `RJafroc`.

6.15.2 Estimating the variance of Az

TBA See book chapter 6.4.4. This is implemented in `RJafroc`.

6.16 References

Chapter 7

Sources of AUC variability

7.1 TBA How much finished

60%

7.2 Introduction

In previous chapters the area AUC under the ROC plot was introduced as the preferred way of summarizing performance in the ROC task, as compared to a pair of sensitivity and specificity values. It can be estimated either non-parametrically, as in Chapter 5, or parametrically, as in Chapter 6, and even better ways of estimating it are described in TBA Chapter 18 and Chapter 20.

Irrespective of how it is estimated AUC is a realization of a random variable, and as such, it is subject to sampling variability. Any measurement based on a finite number of samples from a parent population is subject to sampling variability. This is because no finite sample is unique: someone else conducting a similar study would, in general, obtain a different sample. [Case-sampling variability is estimated by the binormal model in the previous chapter. It is related to the sharpness of the peak of the likelihood function, §6.4.4. The sharper that the peak, the smaller the case sampling variability. This chapter focuses on general sources of variability affecting AUC, regardless of how it is estimated, and other (i.e., not binormal model based) ways of estimating it.]

Here is an outline of this chapter. The starting point is the identification of different sources of variability affecting AUC estimates. Considered next is dependence of AUC on the case-set index $\{c\}$, $c = 1, 2, ..., C$. Considered next is estimating case-sampling variability of the empirical estimate of AUC by an analytic method. This is followed by descriptions of two resampling-based

methods, namely the bootstrap and the jackknife, both of which have wide applicability (i.e., they are not restricted to ROC analysis). The methods are demonstrated using `R` code and the implementation of a calibrated simulator is shown and used to demonstrate their validity, i.e., showing that the different methods of estimating variability agree. The dependence of AUC on reader expertise and modality is considered. An important source of variability, namely the radiologist's choice of internal sensory thresholds, is described. A cautionary comment is made regarding indiscriminate usage of empirical AUC as a measure of performance.

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

7.3 Three sources of variability

Statistics deals with variability. Understanding sources of variability affecting AUC is critical to an appreciation of ROC analysis. Three sources of variability are identified in (Swets and Pickett, 1982): case sampling, between-reader and within-reader variability.

1. Consider a single reader interpreting different case samples. Case-sampling variability arises from the finite number of cases comprising the dataset, compared to the potentially very large population of cases. [If one could sample every case there exists and have them interpreted by the same reader, there would be no case-sampling variability and the poor reader's AUC values (from repeated interpretations of the entire population) would reflect only within reader variability, see #3 below.] Each case-set $\{c\}$, consisting of K_1 non-diseased and K_2 diseased cases interpreted by the reader, yields an AUC value. The notation $\{c\}$ means different *case sets*. Thus $\{c\} = \{1\}, \{2\}$, etc., denote different case sets, each consisting of K_1 non-diseased and K_2 diseased cases.

There is much "data compression" in going from individual case ratings to AUC. For a single reader and given case-set $\{c\}$, the ratings can be converted to an $A_{z\{c\}}$ estimate, TBA Eqn. (6.49). The notation shows explicitly the dependence of the measure on the case-set $\{c\}$. One can conceptualize the distribution of $A_{z\{c\}}$'s over different case-sets, each of the same size $K_1 + K_2$, as a normal distribution, i.e.,

$$A_{z\{c\}} \sim N(A_{z\{\bullet\}}, \sigma^2_{\text{cs+wr}}) \tag{7.1}$$

The dot notation $\{\bullet\}$ denotes an average over all case sets. Thus, $A_{z\{\bullet\}}$ is an estimate of the case-sampling mean of A_z for a single fixed reader and $\sigma^2_{\text{cs+wr}}$ 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 specified reader interpreting different case-sets effectively samples different parts of the population of cases, resulting in variability in measured A_z. Sometimes easier cases are sampled, and sometimes more difficult ones. This source of variability is expected to decrease with increasing case-set size, i.e., increasing $K_1 + K_2$, which is the reason for seeking large numbers of cases in clinical trials. Case-sampling and within-reader variability also decreases as the cases become more homogenous. An example of a more homogenous case sample would be cases originating from a small geographical region with, for example, limited ethnic variability. This is the reason for seeking multi-institutional clinical trials, because they tend to sample more of the population than patients seen at a single institution.

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

$$A_{z;j} \sim N(A_{z;\bullet}, \sigma^2_{\text{br+wr}}) \tag{7.2}$$

where $A_{z;\bullet}$ is an estimate of the reader population AUC mean (the bullet symbol replacing the reader index averages over a set of readers) for the fixed case-set $\{1\}$ and $\sigma^2_{\text{br+wr}}$ 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 groups of J readers interpret the same case set $\{1\}$, thereby sampling different parts of the reader distribution, causing fluctuations in the measured $A_{z;j}$ of the readers. Sometimes better readers are sampled and sometimes not so good ones are sampled. This time there is no "data compression" – each reader in the sample has an associated $A_{z;j}$. However, variability of the average $A_{z;\bullet}$ over the J readers is expected to decrease with increasing J. This is the reason for seeking large reader-samples.

3. Consider a fixed reader, e.g., $j = 1$, interpreting a fixed case-sample $\{1\}$. Within-reader variability is due to variability of the ratings for the same case: the same reader interpreting the same case on different occasions will give different ratings to it, causing fluctuations in the measured AUC. This assumes that memory effects are minimized, for example, by sufficient time between successive interpretations as otherwise, if a case is

shown twice in succession, the reader would give it the same rating each time. Since this is an intrinsic source of variability (analogous to the internal noise of a voltmeter) affecting each reader's interpretations, it cannot be separated from case sampling variability, i.e., it cannot be "turned off". The last sentence needs further explanation. A measurement of case-sampling variability requires a reader, and the reader comes with an intrinsic source of variability that gets added to the case-sampling variance, so what is measured is the sum of case sampling and within-reader variances, denoted $\sigma^2_{\text{cs+wr}}$. 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, yielding $\sigma^2_{\text{br+wr}}$. To emphasize this point, an estimate of case-sampling variability *always* includes within reader variability. Likewise, an estimate of between-reader variability *always* includes within-reader variability.

With this background, the purpose of this chapter is to delve into variability in some detail and in particular describe computational methods for estimating them. This chapter introduces the concept of resampling a dataset to estimate variability and the widely used bootstrap and jackknife methods of estimating variance are described. In a later chapter, these are extended to estimating covariance (essentially a scaled version of the correlation) between two random variables.

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

7.4 Dependence of AUC on the case sample

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

$$AUC \rightarrow AUC_{\{c\}} \tag{7.3}$$

Note that $\{c\}$ is not an individual *case* index, rather it is a *case-set* index, i.e., different integer values of c denote different sets, or samples, or groups, or

collections of cases. [The dependence of AUC on the case sample index is not explicitly shown in the literature.]

What does the dependence of AUC on the c index mean? Different case samples differ in their *difficulty* levels. A difficult case set contains a greater fraction of difficult cases than is usual. A difficult diseased case is one where disease is difficult to detect. For example, the lesions could be partly obscured by overlapping normal structures in the patient anatomy; i.e., the lesion does not "stick out". Alternatively, variants of normal anatomy could mimic a lesion, like a blood vessel viewed end on in a chest radiograph, causing the radiologist to miss the real lesion(s) and mistake these blood vessels for lesions. An easy diseased case is one where the disease is easy to detect. For example, the lesion is projected over smooth background tissue, because of which it "sticks out", or is more conspicuous2. 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.

7.4.1 Case sampling variability of AUC

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

Even if one could get a radiologist to interpret different case-sets, it is even more impractical to actually acquire the different case samples of truth-proven cases. Patients do not come conveniently labeled as non-diseased or diseased. Rather, one needs to follow-up on the patients, perhaps do other imaging tests, in order to establish true disease status, or ground-truth. In screening mammography, a woman who continues to be diagnosed as non-diseased on successive yearly screening tests in the US, and has no other symptoms of breast disease, is probably disease-free. Likewise, a woman diagnosed as diseased and the diagnosis is confirmed by biopsy (i.e., the biopsy comes back showing a malignancy in the sampled tissues) is known to be diseased. However, not all patients who are diseased are actually diagnosed as diseased: a typical false negative fraction is

20% in screening mammography3. 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 6. 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 6 with increased numbers of cases. The variability is seen to decrease.

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

7.5 DeLong method

If the figure-of-merit is the empirical AUC, then a procedure developed by DeLong et al4 (henceforth abbreviated to DeLong) is applicable that is based on earlier work by (Noether, 1967) and (Bamber, 1975). The author will not go into details of this procedure but limit to showing that it "works". However, before one can show that it "works", one needs to know the true value of the variance of empirical AUC. Even if data were simulated using the binormal model, one cannot use the binormal model based estimate of variance as it is an estimate, not to be confused with a true value. Estimates are realizations of random numbers and are themselves subject to variability, which decreases with increasing case-set size. Instead, a "brute-force" (i.e., simulated population sampling) approach is adopted to determine the true value of the variance of AUC. The simulator provides a means of repeatedly generating case-sets interpreted by the same radiologist, and by sampling it enough time, e.g., $C = 10{,}000$ times, each time calculating AUC, one determines the population mean and standard deviation.

The standard deviation determined this way is compared to that yielded by the DeLong method to check if the latter actually works.

```
bruteForceEstimation <-
  function(seed, mu, sigma, K1, K2) {
    # brute force method to
    # find the population
    # meanempAuc and stdDevempAuc
    empAuc <- array(dim = 10000)
    for (i in 1:length(empAuc)) {
      zk1 <- rnorm(K1)
      zk2 <- rnorm(K2, mean = mu, sd = sigma)
      empAuc[i] <- Wilcoxon(zk1, zk2)
    }
    stdDevempAuc  <-  sqrt(var(empAuc))
    meanempAuc   <-  mean(empAuc)
    return(list(
      meanempAuc = meanempAuc,
      stdDevempAuc = stdDevempAuc
    ))
  }
```

```
seed <- 1;set.seed(seed)
mu <- 1.5;sigma <- 1.3;K1 <- 50;K2 <- 52
ret <- bruteForceEstimation(seed, mu, sigma, K1, K2)
# one more trial
zk1 <- rnorm(K1)
zk2 <- rnorm(K2, mean = mu, sd = sigma)
empAuc <- Wilcoxon(zk1, zk2)
ret1  <- DeLongVar(zk1,zk2)
stdDevDeLong <- sqrt(ret1)
cat("brute force estimates:",
    "\nempAuc = ",
    ret$meanempAuc,
    "\npopulation standard deviation =",
    ret$stdDevempAuc, "\n")
#> brute force estimates:
#> empAuc =  0.819178
#> population standard deviation = 0.04176683

cat("single sample estimates = ",
    "\nempirical AUC",
    empAuc,
    "\nstandard deviation DeLong = ",
    stdDevDeLong, "\n")
#> single sample estimates =
```

```
#> empirical AUC 0.8626923
#> standard deviation DeLong =  0.03804135
```

Two functions needed for this code to work are not shown: `Wilcoxon()` calculates the Wilcoxon statistic and the `DeLongVar()` implements the DeLong variance computation method (the DeLong method also calculates co-variances, but these are not needed in the current context). Line 1 sets the `seed` of the random number generator to 1. The `seed` variable is completely analogous to the case-set index `c`. Keeping `seed` fixed realizes the same random numbers each time the program is run. Different values of `seed` result in different, i.e., statistically independent, random samples. Line 2 initialize the values (μ, σ, K_1, K_2) needed by the data simulator: the normal distributions are separated by $\mu = 1.5$, the standard deviation of the diseased distribution is $\sigma = 1.3$, and there are $K_1 = 50$ non-diseased and $K_2 = 52$ diseased cases. Line 3 calls `bruteForceEstimation`, the "brute force" method for estimating mean and standard deviation of the population distribution of AUC, returned by this function, which are the "correct" value to which the DeLong standard deviation estimate will be compared. Lines 4-9 generates a fresh ROC dataset to which the DeLong method is applied.

Two runs of this code were made, one with the smaller sample size, and the other with 10 times the sample size (the second run takes much longer). A third run was made with the larger sample size but with a different seed value. The results follow:

```
seed <- 2;set.seed(seed)
mu <- 1.5;sigma <- 1.3;K1 <- 500;K2 <- 520
ret <- bruteForceEstimation(seed, mu, sigma, K1, K2)
# one more trial
zk1 <- rnorm(K1)
zk2 <- rnorm(K2, mean = mu, sd = sigma)
empAuc <- Wilcoxon(zk1, zk2)
ret1  <- DeLongVar(zk1,zk2)
stdDevDeLong <- sqrt(ret1)
cat("brute force estimates:",
    "\nempAuc = ",
    ret$meanempAuc,
    "\npopulation standard deviation =",
    ret$stdDevempAuc, "\n")
#> brute force estimates:
#> empAuc =  0.8194988
#> population standard deviation = 0.01300203

cat("single sample estimates = ",
    "\nempirical AUC",
    empAuc,
```

```
    "\nstandard deviation DeLong = ",
    stdDevDeLong, "\n")
#> single sample estimates =
#> empirical AUC 0.8047269
#> standard deviation DeLong =  0.01356696
```

1. An important observation is that as sample-size increases, case-sampling variability decreases: 0.0417 for the smaller sample size vs. 0.01309 for the larger sample size, and the dependence is as the inverse square root of the numbers of cases, as expected from the central limit theorem.
2. With the smaller sample size (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 (0.01300203 vs. 0.01356696 – the latter value is for seed = 2).
3. Notice also that the one sample empirical AUC for the smaller case-size is 0.863, which is less than two standard deviations from the population mean 0.819. The "two standard deviations" comes from rounding up 1.96: as in Eqn. (3.32), where $z_{\alpha/2}$ was defined as the upper $1-\alpha/2$ quantile of the unit normal distribution and $z_{0.025} = 1.96$.
4. To reiterate, with clinical data the DeLong procedure estimates case sampling plus within reader variability. With simulated data as in this example, there is no within-reader variability as the simulator yields identical values for fixed seed.

This demonstration should convince the reader that one does have recourse other than the "brute force" method, at least when the figure of merit is the empirical area under the ROC. That should come as a relief, as population sampling is impractical in the clinical context. It should also impress the reader, as the DeLong method is able to use information present in a *single dataset* to tease out its variability. [This is not magic: the MLE estimate is also able to tease out variability based on a parametric fit to a single dataset and examination of the sharpness of the peak of the log-likelihood function, Chapter 6, as are the resampling methods described next.]

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

Table 7.1: Representative counts table.

	$r = 5$	$r = 4$	$r = 3$	$r = 2$	$r = 1$
non-diseased	0	0	9	16	35
diseased	19	8	7	9	7

7.6 Bootstrap method

The simplest resampling method, at least at the conceptual level, is the bootstrap. *The bootstrap method is based on the assumption that one can regard the observed sample as defining the population from which it was sampled.* Since by definition a population cannot be exhausted, the idea is to resample, *with replacement*, from the observed sample. Each resampling step realizes a particular bootstrap sample set denoted $\{b\}$, where $b = 1, 2, ..., B$. The curly brackets emphasize that different integer values of b denote different *sets of cases*, not individual cases. [In contrast, the notation (k) will be used to denote *removing* a specific case, k, as in the jackknife procedure to be described shortly. The index b should not be confused with the index c, the case sampling index; the latter denotes repeated sampling from the population, which is impractical in real life; the bootstrap index denotes repeated sampling from the dataset, which is quite feasible.] The procedure is repeated B times, typically B can be as small as 200, but to be safe I generally use about 1000 - 2000 bootstraps. The following example uses Table 4.1 from Chapter 4.

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 that it is free to be possibly picked again. This is repeated 60 times for the non-diseased hat resulting in 60 ratings from non-diseased cases. A similar procedure is performed using the diseased hat, resulting in 50 ratings from diseased cases. The author has just described, in painful detail (one might say) the realization of the 1st bootstrap sample, denoted $\{b = 1\}$. This is used to construct the 1st bootstrap counts table, Table 7.1.

So what happened? Consider the 35 non-diseased cases with a 1 rating. If each non-diseased case rated 1 in Table 4.1 were picked one time, the total would have been 30, but it is 35. Therefore, some of the original non-diseased cases rated

1 must have been picked multiple times, but one must also make allowance as there is no guarantee that a specific case was picked at all. Still focusing on the 35 non-diseased cases with a 1 rating in the first bootstrap sample, the picked labels, reordered after the fact, with respect to the first index, might be:

$$k_{2,1}, k_{2,1}, k_{4,1}, k_{4,1}, k_{4,1}, k_{6,1}, k_{7,1}, k_{7,1}, k_{9,1}, ..., k_{28,1}, k_{28,1}, k_{30,1}, k_{30,1} \qquad (7.4)$$

In this example, case $k_{1,1}$ was not picked, case $k_{2,1}$ was picked twice, case $k_{3,1}$ was not picked, case $k_{4,1}$ was picked three times, case $k_{5,1}$ was not picked, case $k_{6,1}$ was picked once, etc. The total number of cases in Eqn. (7.4) is 35, and similarly for the other cells in Table 7.1. Next, one estimates AUC for this table. Using the Eng website referred to earlier, one gets AUC = 0.843. [It is OK to use a parametric FOM since the bootstrap is a general procedure applicable, in principle, to any FOM, not just the empirical AUC, unlike the DeLong method, which is restricted to empirical AUC.] The corresponding value for the original data, Table 4.1, was AUC = 0.870. The first bootstrapped dataset yielded a smaller value than the original dataset because one happened to have picked an unusually difficult bootstrap sample.

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

To complete the description of the bootstrap method, one repeats the procedure described in the preceding paragraphs $B = 200$ times, each time running the website calculator and the final result is B values of AUC, denoted:

$$AUC_{\{1\}}, AUC_{\{2\}}, ..., AUC_{\{B\}}$$

where $AUC_{\{1\}} = 0.843$, etc. The bootstrap estimate of the variance of AUC is defined by (Efron and Tibshirani, 1993):

$$\text{Var}\left(AUC\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(AUC_{\{b\}} - AUC_{\{\bullet\}}\right)^2 \tag{7.5}$$

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 empirical AUC, described in Chapter 5 and the binormal model estimate of AUC, described in Chapter 6.

7.6.1 Demonstration of the bootstrap method

To minimize clutter, several `R` functions are not shown, but they are compiled. To display them `clone` or 'fork' the book repository and look at the `Rmd` file corresponding to this output and the sourced `R` files listed below:

```
source(here("R/CH07-Variability/Transforms.R"))
source(here("R/CH07-Variability/LL.R"))
source(here("R/CH07-Variability/RocfitR.R"))
source(here("R/CH07-Variability/RocOperatingPoints.R"))
source(here("R/CH07-Variability/FixRocCountsTable.R"))
source(here("R/CH07-Variability/WilcoxonCountsTable.R"))
```

```
doBootstrap <- function(parametricFOM, B, seed, RocTable) {
  # this is the K vector
  K <- c(sum(RocTable[1,]), sum(RocTable[2,]))

  if (parametricFOM) {
    ret <- RocfitR(RocTable)
  } else {
    ret <- WilcoxonCountsTable(RocTable)
  }
  OrigAUC  <- ret$AUC

  # ready to bootstrap
  # first put the counts data into a linear array
  # convert counts table to array
  z1  <- rep(1:length(RocTable[1,]),
             RocTable[1,])
  z2  <- rep(1:length(RocTable[2,]),
             RocTable[2,])#do:
  AUC <- array(dim = B)#to save the bs AUC values
  for ( b in 1 : B){
```

```
    while (1) {
      RocTable_bs  <-
        array(dim = c(2,length(RocTable[1,])))
      # bs indices for non-diseased
      k1_b <- ceiling( runif( K[ 1 ] ) * K[ 1 ] )
      # bs indices for diseased
      k2_b <- ceiling( runif( K[ 2 ] ) * K[ 2 ] )
      bsTable <- table(z1[k1_b])
      #convert array to frequency table
      RocTable_bs[1,as.numeric(names(bsTable))] <-
        bsTable
      bsTable <- table(z2[k2_b])
      #do:
      RocTable_bs[2,as.numeric(names(bsTable))] <-
        bsTable
      #replace NAs with zeroes
      RocTable_bs[is.na(RocTable_bs )] <- 0
      if (parametricFOM) {
        temp <- RocfitR(RocTable_bs)
      } else {
        temp <- WilcoxonCountsTable(RocTable_bs)
      }
      AUC[b]  <- temp$AUC
      # a return of -1 means AUC did not converge
      if (AUC[b] != -1) break
    }
  }
  meanAUCboot <- mean(AUC)
  Var <- var(AUC)
  stdAUCboot  <- sqrt(Var)
  return(list(
    OrigAUC = OrigAUC,
    meanAUCboot = meanAUCboot,
    stdAUCboot = stdAUCboot
  ))
}
```

Since the bootstrap method is applicable to any scalar figure of merit, two options are provided in the code. If `parametricFOM` is set to `TRUE`, then the binormal model estimate is used, and if set to `FALSE`, the empirical AUC is used. The first set of results are obtained with `parametricFOM` set to `TRUE`.

```
parametricFOM <- TRUE
B <- 200;seed <- 1;set.seed(seed)
RocTable = array(dim = c(2,5))
```

```
RocTable[1,]  <- c(30,19,8,2,1)
RocTable[2,]  <- c(5,6,5,12,22)

ret <- doBootstrap(parametricFOM, B, seed, RocTable)
OrigAUC <- ret$OrigAUC
meanAUCboot <- ret$meanAUCboot
stdAUCboot <- ret$stdAUCboot

cat("Bootstrap variance estimation:",
    "\nparametricFOM = ", parametricFOM,
    "\nseed = ", seed,
    "\nB = ", B,
    "\nOrigAUC = ", OrigAUC,
    "\nmeanAUCboot = ", meanAUCboot,
    "\nstdAUCboot = ", stdAUCboot, "\n")
#> Bootstrap variance estimation:
#> parametricFOM =  TRUE
#> seed =  1
#> B =  200
#> OrigAUC =  0.8704519
#> meanAUCboot =  0.8671713
#> stdAUCboot =  0.04380523
```

This shows that the AUC of the original data (i.e., before performing any bootstrapping) is 0.870, the mean AUC of the B = 200 bootstrapped datasets is 0.867, and the standard deviation of the 200 bootstraps is 0.0438. If one runs the website calculator referenced in the previous chapter on the dataset shown in Table 4.1, one finds that the MLE of the standard deviation of the AUC of the fitted ROC curve is 0.0378. The standard deviation is itself a statistic and there is sampling variability associated with it, i.e., there exists such a beast as a standard deviation of a standard deviation; the bootstrap estimate is not too far from the MLE estimate. By setting `seed` to different values, one gets an idea of the variability of the estimate of the standard deviation of AUC. For example, with `seed` = 2, one gets:

```
#> Bootstrap variance estimation:
#> parametricFOM =  TRUE
#> seed =  2
#> B =  200
#> OrigAUC =  0.8704519
#> meanAUCboot =  0.8673155
#> stdAUCboot =  0.03815402
```

Note that both the mean of the bootstrap samples and the standard deviation have changed, but both are close to the MLE values. Examined next is the

dependence of the estimates on `B`, the number of bootstraps. With `seed` = 1 and `B` = 2000 one gets:

```
#> Bootstrap variance estimation:
#> parametricFOM =  TRUE
#> seed =  1
#> B =  2000
#> OrigAUC =  0.8704519
#> meanAUCboot =  0.8674622
#> stdAUCboot =  0.03833508
```

The estimates are evidently rather insensitive to B, but the computation time was longer, ~13 seconds (running MLE 2000 times in 13 seconds is not bad!). It is always a good idea to test the stability of the results to different B and `seed` values. Unlike the DeLong et al method, which is restricted to the Wilcoxon statistic (which equals empirical AUC as per the Bamber theorem), the bootstrap is broadly applicable to other figures of merit, including non-ROC paradigm figures of merit. However, beware that it depends on the assumption that the sample itself is representative of the population. With limited numbers of cases, this could be a bad assumption. [With small numbers of cases it is relatively easy to enumerate the different outcomes of the sampling process and, more importantly, their respective probabilities, leading to what is termed the *exact bootstrap*. It is "exact" in the sense that there is no seed variable or number of bootstrap dependence.]

Finally, here is the output when using non-parametric AUC, with `seed` = 1.

```
#> Bootstrap variance estimation:
#> parametricFOM =  FALSE
#> seed =  1
#> B =  200
#> OrigAUC =  0.8606667
#> meanAUCboot =  0.8604575
#> stdAUCboot =  0.04125475
```

7.7 Jackknife method

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

In this method, the first case is removed, or jackknifed, from the set of cases and the MLE (or empirical estimation) is conducted on the resulting dataset, which

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

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

The corresponding jackknife pseudovalues Y_k are defined by:

$$Y_k = K \times AUC - (K-1) \times AUC_{(k)} \tag{7.7}$$

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 TBA Chapter 09. The *jackknife AUC values*, defined by Eqn. (7.6), should not be confused with jackknife derived psuedovalues, defined by Eqn. (7.7).

The jackknife estimate of the variance is defined by (Efron and Tibshirani, 1993):

$$\mathrm{Var}_{\mathrm{jack}} = \frac{(K-1)^2}{K} \frac{1}{K-1} \sum_{k=1}^{K} \left(AUC_{(k)} - AUC_{(\bullet)} \right)^2 \tag{7.8}$$

Since variance of K scalars is defined by:

$$\mathrm{Var}\,(x) = \frac{1}{K-1} \sum_{k=1}^{K} (x_k - x_\bullet)^2 \tag{7.9}$$

It follows that:

$$\mathrm{Var}_{\mathrm{jack}}\,(\mathrm{AUC}) = \frac{(K-1)^2}{K} \mathrm{Var}\,(\mathrm{AUC}) \tag{7.10}$$

In Eqn. (7.8) I have deliberately not simplified the right hand side by canceling out $K-1$. The purpose is to show, Eqn. (7.10), that the usual expression for the variance (of the jackknife FOM values) needs to be multiplied by a **variance inflation factor** $\frac{(K-1)^2}{K}$, which is approximately equal to K, in order to obtain the correct jackknife estimate of variance of AUC. This factor was not necessary when one used the bootstrap method. That is because the bootstrap samples are more representative of the actual spread in the data. The jackknife samples are more restricted than the bootstrap samples, so the spread of the data is smaller; hence the need for the variance inflation factor (Efron and Tibshirani, 1993).

```
doJackknife <- function(parametricFOM, RocTable) {
  # this is the K vector
  K <- c(sum(RocTable[1,]), sum(RocTable[2,]))

  if (parametricFOM) {
    ret <- RocfitR(RocTable)
  } else {
    ret <- WilcoxonCountsTable(RocTable)
  }
  OrigAUC  <- ret$AUC

  # first put the counts data into a linear array
  z1  <- rep(1:length(RocTable[1,]),
             RocTable[1,])
  z2  <- rep(1:length(RocTable[1,]),
             RocTable[2,])

  AUC_jack  <- array(dim = sum(K))
  Y_k <- array(dim = sum(K))
  z_jk  <- array(dim = sum(K))
  # ready to jackknife
  for ( k in 1 : sum(K)){
    RocTable_jk  <- array(dim = c(2,length(RocTable[1,])))
    if ( k <= K[ 1 ]){
      z1_jk <- z1[ -k ]
      z2_jk <- z2
    }else{
      z1_jk <- z1
      z2_jk <- z2[ -(k - K[ 1 ]) ]
    }
    #convert array to frequency table
    RocTable_jk[1,1:length(table(z1_jk))]  <-
      table(z1_jk)
    RocTable_jk[2,1:length(table(z2_jk))]  <-
      table(z2_jk)
    #replace NAs with zeroes
    RocTable_jk[is.na(RocTable_jk)] <- 0
    # AUC_jack for observed data
    if (parametricFOM) {
      temp <- RocfitR(RocTable_jk)
    } else {
      temp <- WilcoxonCountsTable(RocTable_jk)
    }
    AUC_jack[k]  <- temp$AUC
    Y_k[k] <- sum(K)*OrigAUC - (sum(K)-1)*AUC_jack[k]
```

```
    if (AUC_jack[k] == -1)
      stop("RocfitR did not converge in jackknife loop")
  }
  meanAUCjack <- mean(AUC_jack)
  #Efron and Stein's paper, include jackknife inflation factor
  Var_jack <- var(AUC_jack) * ( sum(K) - 1)^2 / sum(K)
  stdAUCjack  <- sqrt(Var_jack)
  return(list(
    OrigAUC = OrigAUC,
    meanAUCjack = meanAUCjack,
    stdAUCjack = stdAUCjack
  ))
}
```

Since the jackknife method is applicable to any scalar figure of merit, two options are provided in the code. If `parametricFOM` is set to `TRUE`, then the binormal model estimate is used, and if set to `FALSE`, the empirical AUC is used. The first set of results are obtained with `parametricFOM` set to `TRUE`. Notice that the code does not use a `set.seed()` statement, as no random number generator is needed in the jackknife method. Systematically removing and replacing each case in sequence, one at a time, is not random sampling, which should further explain the need for the variance inflation factor in Eqn. (7.10).

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

ret <- doJackknife(parametricFOM, RocTable)
OrigAUC <- ret$OrigAUC
meanAUCjack <- ret$meanAUCjack
stdAUCjack <- ret$stdAUCjack

cat("Jackknife variance estimation:",
    "\nparametricFOM = ", parametricFOM,
    "\nOrigAUC = ", OrigAUC,
    "\nmeanAUCjack = ", meanAUCjack,
    "\nstdAUCjack = ", stdAUCjack, "\n")
#> Jackknife variance estimation:
#> parametricFOM =  TRUE
#> OrigAUC =  0.8704519
#> meanAUCjack =  0.8704304
#> stdAUCjack =  0.03861591
```

The next output is with the non-parametric figure of merit:

```
#> Jackknife variance estimation:
#> parametricFOM =  FALSE
#> OrigAUC =  0.8606667
#> meanAUCjack =  0.8606667
#> stdAUCjack =  0.03689264
```

It may be noticed that the mean of the jackknife figure of merit values, i.e., 0.8606667, exactly equals the original figure of merit 0.8606667 (i.e., that calculated including all cases). This can be shown analytically to be true so long as the figure of merit is the empirical AUC. A similar relation is not true for the bootstrap.

7.8 Calibrated simulator

7.8.1 The need for a calibrated simulator

The population sampling method used previously, 7.5, to compare the DeLong method to a known standard used arbitrarily set simulator values, i.e., $\mu = 1.5$ and $\sigma = 1.3$. One does not know if these values actually represent real clinical data. In this section a simple method of implementing population sampling using a *calibrated simulator* is described. A calibrated simulator is one whose parameters are chosen to match those of an actual clinical dataset. This way one has some assurance that the simulator is realistic and therefore its verdict on a proposed method or analysis (in our case method of estimating AUC variability) is likely to be correct.

7.8.2 Implementation of a simple calibrated simulator

The simple simulator described here is limited to a single reader single modality dataset. A more complex simulator describing multiple readers in multiple modalities is described in a later chapter (TBA). Consider a clinical dataset, such as in Table 4.1. Analyzed by MLE, this yields binormal model parameters, `a`, `b` and the thresholds $\zeta_1, \zeta_2, \zeta_3, \zeta_4$. After conversion to $\mu = a/b$ and $\sigma = 1/b$ and new zetas $\zeta = \zeta/b$, the values are (in the same order): 2.173597, 1.646099, 0.01263423, 1.475351, 2.494901, 3.945221 (see code output below):

```
# mu_sigma is the mu-sigma notation
mu_sigma <- c(2.173597, 1.646099, 0.01263423, 1.475351, 2.494901, 3.945221)
# ab is the a-b notation
ab <- c(1.320453, 0.607497, 0.007675259, 0.8962713, 1.515645, 2.39671)
ab[1]/ab[2] # this is mu
#> [1] 2.173596
1/ab[2]    # this is sigma
```

```
#> [1] 1.646099
ab[3:6]/ab[2] # this is zeta in mu-sigma notation
#> [1] 0.01263423 1.47535099 2.49490121 3.94522113
```

[The reason for dividing ζ by b is that when re-scaling the decision variable axis by b one must also re-scale the cutoffs.] The values μ, σ, ζ define the calibrated simulator, in the sense that the parameter values are calibrated to match the dataset in Table 4.1.

Here is the function `doCalSimulator()` that will be used to perform the initial calibration followed by population sampling from the calibrated simulator:

```
doCalSimulator <- function(P, parametricFOM, RocCountsTable) {
  K <- c(sum(RocCountsTable[1,]),
         sum(RocCountsTable[2,]))
  # perform the initial calibration
  ret <- RocfitR(RocCountsTable) # AUC for observed data
  a <- ret$a
  b  <- ret$b
  zetas <- ret$zeta
  mu  <- a/b
  sigma <- 1/b
  zetas <- zetas/b # need to also scale zetas
   # AUC for observed data
  if (parametricFOM) {
    OrigAUC <- RocfitR(RocCountsTable)$AUC
  } else {
    OrigAUC <- WilcoxonCountsTable(RocCountsTable)$AUC
  }
  # perform the population sampling
  AUC <- array(dim = P)
  for ( p in 1 : P){
    while (1) {

      RocCountsTableSimPop <-
        SimulateRocCountsTable(K, mu, sigma, zetas)
      if (parametricFOM) {

        # AUC for fitted curve
        temp <- RocfitR(RocCountsTableSimPop)
        # a return of -1 means RocFitR did not converge
        if (temp[1] != -1) {
          AUC[p]  <- temp$AUC
          break
        }
```

```
34        } else {
35          AUC[p] <- (WilcoxonCountsTable(RocCountsTableSimPop))$AUC
36          break
37        }
38      }
39    }
40    AUC <- AUC[!is.na(AUC)]
41    meanAUC  <- mean(AUC)
42    stdAUC  <- sqrt(var(AUC))
43    return(list(
44      mu = mu, # these define the calibration simulator
45      sigma = sigma, #do:
46      zetas = zetas, #do:
47      OrigAUC = OrigAUC,
48      meanAUC = meanAUC,
49      stdAUC = stdAUC
50    ))
51  }
```

In the function `doCalSimulator(P, parametricFOM, RocCountsTable)`, `P` is the desired number of population samples, `parametricFOM` is a logical, if set to `TRUE` the binormal model is used to calculate *fitted* AUC and otherwise the Wilcoxon statistic is used to calculate *empirical* AUC, and `RocCountsTable` contains the ROC data, such as Table 4.1, to which the simulator is to be calibrated to. Lines 2-3 construct the K-vector, containing K_1, K_2. Line 5 performs the maximum likelihood fit, using function `RocfitR(RocCountsTable)`. The returned variable contains a, b, ζ as a `list`, which are extracted at lines 6-8. Lines 9-11 converts these to the mu-sigma notation. In essence, lines 5 - 11 calibrates the simulator and the calibrated values of the simulator are contained in μ, σ, ζ. Lines 13-17 calculates `OrigAUC`, the AUC of the original data, using parametric `RocfitR` or the Wilcoxon statistic, as appropriate, depending on the value of `parametricFOM`. After defining a length `P` array, at line 19, to hold the sampled `AUC` values, lines 20-39 begins and ends a `for` loop to conduct the `P` population samples. Each pass through the `for` loop yields K_1 samples from the non-diseased distribution and K_2 samples from the diseased distribution, returned in the variable `RocCountsTableSimPop`, which is similar in structure to a counts table like Table 4.1. Within the `for` loop there is an endless `while` loop, needed because `RocfitR` can sometimes fail to converge, signaled by the first member of the returned `list` being minus 1, in which case another iteration of the `while` loop is performed (see line 30) and otherwise the `break` statement (line 32) causes program execution to proceed to the next iteration of the `for` loop. After entering the `while` loop, lines 22-23, a new ROC counts table is generated. The returned `list` is saved to `temp` at line 28, and if `temp[1] != -1` (i.e., `RocfitR` did converge) the AUC value is saved to `AUC[p]`, line 31. Upon exiting the code one has `P` values of AUC in the array `AUC`.

7.8.2.1 Parametric AUC results

The following code uses the function just described and prints out the results.

```
parametricFOM <- TRUE
seed <- 1
set.seed(seed)
P <- 2000
RocCountsTable = array(dim = c(2,5))
RocCountsTable[1,]  <- c(30,19,8,2,1)
RocCountsTable[2,]  <- c(5,6,5,12,22)
ret <- doCalSimulator(P, parametricFOM, RocCountsTable)
mu <- ret$mu
sigma <- ret$sigma
zetas <- ret$zetas
meanAUC_1_2000 <- ret$meanAUC
stdAUC_1_2000 <- ret$stdAUC
```

After setting `parametricFOM` to `TRUE` (for a parametric fit), `seed` to 1 and `P` to 2000, the ROC counts table is defined and the function `doCalSimulator()` is called. The returned `list` contains the parameter values for the calibrated simulator: $\mu = 2.1735969$, $\sigma = 1.6460988$ and $\zeta = 0.0126342$, 1.4753512, 2.4949012, 3.9452209. It also contains `OrigAUC`, the AUC of the original data, calculated by `RocfitR()`, in this case `OrigAUC` = 0.8704519, and the mean and standard deviation of the 2000 AUC values, equal to 0.8676727 and 0.0403331, respectively.

The simulations were repeated with `seed` = 2. This time the mean and standard deviation of the 2000 AUC values, are equal to 0.8681855 and 0.0405516, respectively. The respective values corresponding to the two `seed` values are quite close to each other (to within a percent).

More variability is observed, as expected, when the above two simulations are repeated with `P` = 200:

For `seed` = 1 and `P` = 200 the mean and standard deviation of the 200 AUC values, are 0.8727151 and 0.0355281, respectively.

For `seed` = 2 and `P` = 200 the mean and standard deviation of the 200 AUC values, are 0.8649385 and 0.0450947, respectively. Note the greater variability induced by the change in `seed`, as compared to `P` = 2000.

7.8.2.2 Non-parametric AUC results

The next simulation is with `seed` = 1 and `P` = 2000, but this time `parametricFOM` is set to `FALSE`. The calibration proceeds as before, using `RocfitR` to determine the parameters of the simulation model, calibrating the

simulator requires a parametric fit, but this empirical AUC is used to obtained the 2000 AUC samples. The mean and standard deviation of the AUC values, are 0.8497634 and 0.0367476, respectively. Note that these are smaller than the corresponding parametric estimates. The empirical AUC is expected to be smaller than the corresponding parametric AUC as joining adjacent points with straight lines will underestimate the area under the smooth ROC curve. Repeating with `seed` = 2, the mean and standard deviation of the AUC values, are 0.8503732 and 0.0369091, respectively, which are close to the `seed` = 1 values.

7.9 Discussion

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

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

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

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

TBA: what to do with removed sections?

7.10 References

Significance Testing

Chapter 8

Hypothesis Testing

8.1 TBA How much finished

60%

8.2 Introduction

The problem addressed here is how to decide whether an estimate of AUC is consistent with a pre-specified value. One example of this is when a single-reader rates a set of cases in a single-modality, from which one estimates AUC, and the question is whether the estimate is statistically consistent with a pre-specified value. From a clinical point of view, this is generally not a useful exercise, but its simplicity is conducive to illustrating the broader concepts involved in this and later chapters. The clinically more useful analysis is when multiple readers interpret the same cases in two or more modalities. [With two modalities, for example, one obtains an estimate AUC for each reader in each modality, averages the AUC values over all readers within each modality, and computes the inter-modality difference in reader-averaged AUC values. The question forming the main subject of this book is whether the observed difference is consistent with zero.]

Each situation outlined above admits a binary (yes/no) answer, which is different from the estimation problem that was dealt with in connection with the maximum likelihood method in (book) Chapter 06, where one computed numerical estimates (and confidence intervals) of the parameters of the fitting model.

Hypothesis testing is the process of dichotomizing the possible outcomes of a statistical study and then using probabilistic arguments to choose one option over the other.

The two options are termed the *null hypothesis* (NH) and the *alternative hypothesis* (AH). The hypothesis testing procedure is analogous to the jury trial system in the US, with 20 instead of 12 jurors, with the NH being the presumption of innocence and the AH being the defendant is guilty. The decision rule is to assume the defendant is innocent unless all 20 jurors agree the defendant is guilty. If even one juror disagrees, the defendant is deemed innocent (equivalent to choosing an α – defined below - of 0.05, or 1/20).

8.3 Single-modality single-reader ROC study

The binormal model described in Chapter 06 can be used to generate sets of ratings to illustrate the methods being described in this chapter. To recapitulate, the model is described by:

$$Z_{k_1 1} \sim N(0,1)$$
$$Z_{k_2 2} \sim N(\mu, \sigma^2)$$

The following code chunk encodes the `Wilcoxon` function:

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

In the next code chunk we set $\mu = 1.5$ and $\sigma = 1.3$ and simulate $K_1 = 50$ non-diseased cases and $K_2 = 52$ diseased cases. The `for`-loop draws 50 samples from the $N(0,1)$ distribution and 52 samples from the $N(\mu, \sigma^2)$ distribution, calculates the empirical AUC using the Wilcoxon, and the process is repeated 10,000 times, the AUC values are saved to a huge array `AUC_c` (the c-subscript is for case sample, where each case sample represents 102 cases). After exit from the `for`-loop we calculate the mean and standard deviation of the `AUC` values.

```
seed <- 1;set.seed(seed)
mu <- 1.5;sigma <- 1.3;K1 <- 50;K2 <- 52

# cheat to find the population mean and std. dev.
AUC_c <- array(dim = 10000)
for (c in 1:length(AUC_c)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu, sd = sigma)
  AUC_c[c] <- Wilcoxon(zk1, zk2)
}
meanAUC   <-  mean(AUC_c);sigmaAUC  <-  sd(AUC_c)
cat("pop mean AUC_c = ", meanAUC,
    ", pop sigma AUC_c = ", sigmaAUC, "\n")
#> pop mean AUC_c =  0.819178 , pop sigma AUC_c =  0.04176683
```

By the simple (if unimaginative) approach of sampling 10,000 times, one has estimates of the *population* mean and standard deviation of empirical AUC, denoted below by AUC_{pop} and σ_{AUC}, respectively.

The next code-chunk simulates one more independent ROC study with the same numbers of cases, and the resulting area under the empirical curve is denoted AUC in the code.

```
# one more trial, this is the one we want
# to compare to meanAUC
zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu, sd = sigma)
AUC <- Wilcoxon(zk1, zk2)
cat("New AUC = ", AUC, "\n")
#> New AUC =  0.8626923

z <- (AUC - meanAUC)/sigmaAUC
cat("z-statistic = ", z, "\n")
#> z-statistic =  1.04184
```

Is the new value, 0.8626923, sufficiently different from the population mean, 0.819178, to reject the null hypothesis NH : $\text{AUC} = \text{AUC}_{pop}$? Note that the answer to this question can be either yes or no: equivocation is not allowed!

The new value is "somewhat close" to the population mean, but how does one decide if "somewhat close" is close enough? Needed is the statistical distribution of the random variable AUC under the hypothesis that the true mean is AUC_{pop}. In the limit of a large number of cases, the pdf of AUC under the null hypothesis is a normal distribution $N\left(\text{AUC}_{pop}, \sigma^2_{\text{AUC}}\right)$:

$$\text{pdf}_{\text{AUC}}\left(\text{AUC} \mid \text{AUC}_{pop}, \sigma_{\text{AUC}}\right) = \frac{1}{\sigma_{\text{AUC}}\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{\text{AUC} - \text{AUC}_{pop}}{\sigma_{\text{AUC}}}\right)^2\right)$$

The translated and scaled value is distributed as a unit normal distribution, i.e.,

$$Z \equiv \frac{\text{AUC} - \text{AUC}_{pop}}{\sigma_{\text{AUC}}} \sim N(0,1)$$

[The Z notation here should not be confused with z-sample, decision variable or rating of a case in an ROC study; the latter, when sampled over a set of non-diseased and diseased cases, yield a realization of AUC. The author trusts the distinction will be clear from the context.] The observed magnitude of z is 1.0418397. [Upper-case for random variable, lower-case for realized or observed value.]

The ubiquitous p-value is the probability that the observed magnitude of z, or larger, occurs under the null hypothesis (NH) that the true mean of Z is zero. Stated somewhat differently, but equivalently, it is the probability that a random sample from $N(0,1)$ exceeds z.

The p-value corresponding to an observed z of 1.0418397 is given by:

$$\begin{aligned}\Pr(|Z| \geq |z| \mid Z \sim N(0,1)) &= \Pr(|Z| \geq 1.042 \mid Z \sim N(0,1)) \\ &= 2\Phi(-1.042) \\ &= 0.2975\end{aligned}$$

To recapitulate statistical notation, $\Pr(|Z| \geq |z| \mid Z \sim N(0,1))$ is parsed as $\Pr(A \mid B)$, that is, the probability $|Z| \geq |z|$ given that $Z \sim N(0,1)$. The second line in the preceding equation follows from the symmetry of the unit normal distribution, i.e., the area above 1.042 must equal the area below -1.042.

Since z is a continuous variable, there is zero probability that a sampled value will exactly equal the observed value. Therefore, one must pose the statement as above, namely the probability that Z is at least as extreme as the observed value (by "extreme" I mean further from zero, in either positive or negative directions). If the observed was $z = 2.5$ then the corresponding p-value would be $2\Phi(-2.5)$=0.01242, which is smaller than 0.2975. Under the zero-mean null hypothesis, the larger the magnitude of the observed value z, the smaller the p-value, and the more unlikely that the data supports the NH. **The p-value can be interpreted as the degree of unlikelihood that the data is consistent with the NH.**

By convention one adopts a fixed value of the probability, denoted and usually $\alpha = 0.05$, which is termed *the significance level* of the test, and the decision rule is to reject the null hypothesis if the observed p-value $< \alpha$. α is also referred to as the *size* of the test.

$$p < \alpha \Rightarrow \text{Reject NH}$$

If the p-value is exactly 0.05 (unlikely with ROC analysis, but one needs to account for it), then one does not reject the NH. In the 20-juror analogy, of one juror insists the defendant is not guilty, the observed p-value is 0.05, and one does not reject the NH that the defendant is innocent (the double negatives, very common in statistics, can be confusing; in plain English, the defendant goes home).

According to the previous discussion, the critical magnitude of z that determines whether to reject the null hypothesis is given by:

$$z_{\alpha/2} = -\Phi^{-1}(\alpha/2)$$

For $\alpha = 0.05$ this evaluates to 1.95996 (which is sometimes rounded up to two, good enough for "government work" as the saying goes) and the decision rule is to reject the null hypothesis only if the observed magnitude of z is larger than $z_{\alpha/2}$.

The decision rule based on comparing the observed z to a critical value is equivalent to a decision rule based on comparing the observed p-value to α. It is also equivalent, as will be shown later, to a decision rule based on a $(1-\alpha)$ confidence interval for the observed statistic. One rejects the NH if the closed confidence interval does not include zero.

8.4 Type-I errors

Just because one rejects the null hypothesis does not mean that the null hypothesis is false. Following the decision rule puts an upper limit on, or "caps", the probability of incorrectly rejecting the null hypothesis at α. In other words, by agreeing to reject the NH only if $p \leq \alpha$, one has set an upper limit, namely α, on errors of this type, termed *Type-I* errors. These could be termed false positives in the hypothesis testing sense, not to be confused with false positive occurring on individual case-level decisions. According to the definition of α:

$$\Pr(\text{Type I error} \mid \text{NH}) = \alpha$$

To demonstrate the ideas one needs to have a very cooperative reader interpreting new sets of independent cases not just one more time, but 2000 more times (the reason for the 2000 trials will be explained below). The simulation code follows:

```
seed <- 1;set.seed(seed)
mu <- 1.5;sigma <- 1.3;K1 <- 50;K2 <- 52
```

```
nTrials <- 2000
alpha <- 0.05 # size of test
reject = array(0, dim = nTrials)
for (trial in 1:length(reject)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu, sd = sigma)
  AUC <- Wilcoxon(zk1, zk2)
  z <- (AUC - meanAUC)/sigmaAUC
  p <- 2*pnorm(-abs(z)) # p value for individual trial
  if (p < alpha) reject[trial] = 1
}

CI <- c(0,0); width <- -qnorm(alpha/2)
ObsvdTypeIErrRate <- sum(reject)/length(reject)
CI[1] <- ObsvdTypeIErrRate -
  width*sqrt(ObsvdTypeIErrRate*(1-ObsvdTypeIErrRate)/nTrials)
CI[2] <- ObsvdTypeIErrRate +
  width*sqrt(ObsvdTypeIErrRate*(1-ObsvdTypeIErrRate)/nTrials)
cat("alpha = ", alpha, "\n")
#> alpha =  0.05
cat("ObsvdTypeIErrRate = ", ObsvdTypeIErrRate, "\n")
#> ObsvdTypeIErrRate =  0.0535
cat("95% confidence interval = ", CI, "\n")
#> 95% confidence interval =  0.04363788 0.06336212
exact <- binom.test(sum(reject),n = 2000,p = alpha)
cat("exact 95% CI = ", as.numeric(exact$conf.int), "\n")
#> exact 95% CI =  0.04404871 0.06428544
```

The population means were calculated in an earlier code chunk. One initializes `NTrials` to 2000 and α to 0.05. The `for-loop` describes our captive reader interpreting independent sets of cases 2000 times. *Each completed interpretation of 102 cases is termed a trial.* For each trial one calculates the observed value of `AUC`, the observed `z` statistic and the the observed p-value. The observed p-value is compared against the fixed value α and one sets the corresponding `reject[trial]` flag to unity if $p < \alpha$. In other words, if the trial-specific p-value is less than α one counts an instance of rejection of the null hypothesis. The process is repeated 2000 times.

Upon exit from the for-loop, one calculates the observed Type-I error rate, denoted `ObsvdTypeIErrRate` by summing the reject array and dividing by 2000. One calculates a 95% confidence interval for `ObsvdTypeIErrRate` based on the binomial distribution, as in (book) Chapter 03.

The observed Type-I error rate is a realization of a random variable, as is the estimated 95% confidence interval. The fact that the confidence interval includes $\alpha = 0.05$ is no coincidence - it shows that the hypothesis testing procedure is working as expected. To distinguish between the selected α (a fixed value) and

that observed in a simulation study (a realization of a random variable), the term *empirical* α is sometimes used to denote the observed rejection rate.

It is a mistake to state that one wishes to minimize the Type-I error probability. The minimum value of α (a probability) is zero. Run the software with this value of α: one finds that the NH is never rejected. The downside of minimizing the expected Type-I error rate is that the NH will never be rejected, even when the NH is patently false. The aim of a valid method of analyzing the data is not minimizing the Type-I error rate, rather, the observed Type-I error rate should equal the specified value of α (0.05 in our example), allowance being made for the inherent variability in it's estimate. This is the reason 2000 trials were chosen for testing the validity of the NH testing procedure. With this choice, the 95% confidence interval, assuming that observed value is close to 0.05, is roughly ± 0.01 as explained next.

Following analogous reasoning to (book) Chapter 03, Eqn. (3.10.10), and defining f as the observed rejection fraction over T trials, and as usual, F is a random variable and f a realized value,

$$\sigma_f = \sqrt{f(1-f)/TF} \sim N\left(f, \sigma_f^2\right)$$

An approximate $(1-\alpha)100$ percent CI for f is:

$$CI_f = \left[f - z_{\alpha/2}\sigma_f, f + z_{\alpha/2}\sigma_f\right]$$

If f is close to 0.05, then for 2000 trials, the 95% CI for f is $f \pm 0.01$, i.e., `qnorm(alpha/2) * sqrt(.05*(.95)/2000)` $= 0.009551683 \sim 0.01$.

The only way to reduce the width of the CI, and thereby run a more stringent test of the validity of the analysis, is to increase the number of trials T. Since the width of the CI depends on the inverse square root of the number of trials, one soon reaches a point of diminishing returns. Usually $T = 2000$ trials are enough for most statisticians and me, but studies using more simulations have been published.

8.5 One vs. two sided tests

The test described above is termed 2-tailed. Here, briefly, is the distinction between 2-tailed vs. 1-tailed p-values:

```
alpha  <- 0.05
# Example 1
# p value for two-sided AH
p2tailed <- pnorm(-abs(z)) + (1-pnorm(abs(z)))
```

```
cat("pvalue 2-tailed, AH: z ne 0 = ", p2tailed, "\n")
#> pvalue 2-tailed, AH: z ne 0 =  0.2943993

# Example 2
# p value for one-sided AH gt 0
p1tailedGT <- 1-pnorm(z)
cat("pvalue 1-tailed, AH: z gt 0 = ", p1tailedGT, "\n")
#> pvalue 1-tailed, AH: z gt 0 =  0.8528004

# Example 2
# p value for one-sided AH lt 0
p1tailedLT <- pnorm(z)
cat("pvalue 1-tailed, AH: z lt 0 = ", p1tailedGT, "\n")
#> pvalue 1-tailed, AH: z lt 0 =  0.8528004

df <- data.frame(p2tailed = p2tailed,
                 p1tailedGT = p1tailedGT,
                 p1tailedGT = p1tailedGT)
print(df)
#>    p2tailed p1tailedGT p1tailedGT.1
#> 1 0.2943993  0.8528004    0.8528004
```

The only difference between these tests is in how the alternative hypotheses is stated.

- For a two-tailed test the alternative hypothesis is AUC $\neq$ AUC$_{pop}$. Large deviations, in either direction, cause rejection of the NH.
- For the first one-tailed test the alternative hypothesis is AUC $>$ AUC$_{pop}$. Large positive observed values of z result in rejection of the NH. Large negative values do not.
- For the second one-tailed test the alternative hypothesis is AUC $<$ AUC$_{pop}$. Large negative observed values of z result in rejection of the NH. Large positive values do not.
- The last two statements are illustrated below with the following code-fragments:

```
# p1tailedGT
1-pnorm(1) # do not reject
#> [1] 0.1586553
1-pnorm(2) # reject
#> [1] 0.02275013
1-pnorm(-2) # do not reject
#> [1] 0.9772499
```

```
# p1tailedGT
pnorm(-1) # do not reject
#> [1] 0.1586553
pnorm(-2) # reject
#> [1] 0.02275013
pnorm(2) # do not reject
#> [1] 0.9772499
```

Note that the p-value of the 1-tailed tests are half that of the 2-tailed test. Further discussion of the difference between 2-tailed and 1-tailed tests, and when the latter might be appropriate, is given below.

If the null hypothesis is rejected anytime the magnitude of the observed value of z exceeded the critical value $-\Phi^{-1}(\alpha/2)$. This is a statement of the alternative hypothesis (AH) AUC $\neq$ AUC$_{pop}$, in other words too high or too low values of z *both* result in rejection of the null hypothesis. This is referred to as a two-sided AH and the resulting p-value is termed a *two-sided* p-value. This is the most common one used in the literature.

Suppose the additional trial performed by the radiologist was performed after an intervention following which the radiologist's performance is expected to increase. To make matters clearer, assume the interpretations in the 10,000 trials used to estimate AUC$_{pop}$ were performed with the radiologist wearing an old pair of eye-glasses, possibly out of proper strength, and the additional trial is performed after the radiologist gets a new set of prescription eye-glasses. Because the radiologist's eyesight has improved, the expectation is that performance should increase. In this situation, it is appropriate to use the one-sided alternative hypothesis AUC $>$ AUC$_{pop}$. Now, large positive values of z result in rejection of the null hypothesis, but large negative values do not. The critical value of z is defined by $z_\alpha = \Phi(1-\alpha)$, which for $\alpha = 0.05$ is 1.645 (i.e., `qnorm(1-alpha) = 1.644854`). Compare 1.64 to the critical value $-\Phi^{-1}(\alpha/2)$ = 1.96 for a two-sided test. If the change is in the expected direction, it is more likely that one will reject the NH with a one-sided than with a two-sided test. The p-value for a one-sided test is given by:

$$\Pr(Z \geq 1.042 \mid \text{NH}) = \Phi(-1.042) = 0.1487$$

Notice that this is half the corresponding two-sided test p-value; this is because one is only interested in the area under the unit normal that is above the observed value of z. If the intent is to obtain a significant finding, it is tempting to use one-sided tests. The down side of a one-sided test is that even with a large excursion of the observed z in the other direction one cannot reject the null hypothesis. So if the new eye-glasses are so bad as to render the radiologist practically blind (think of a botched cataract surgery) the observed z would be large and negative, but one cannot reject the null hypothesis AUC = AUC$_{pop}$.

The one-sided test could be run the other way, with the alternative hypothesis being stated as AUC $<$ AUC$_{pop}$. Now, large negative excursions of the observed value of AUC cause rejection of the null hypothesis, but large positive excursions do not. The critical value is defined by $z_\alpha = \Phi^{-1}(\alpha)$, which for $\alpha = 0.05$ is -1.645. The p-value is given by (note the reversed sign compared to the previous one-sided test:

$$\Pr(Z \leq 1.042 \mid \text{NH}) = \Phi(1.042) = 1 - 0.1487 = 0.8513$$

This is the complement of the value for a one-sided test with the alternative hypothesis going the other way: obviously the probability that Z is smaller than the observed value (1.042) plus the probability that Z is larger than the same value must equal one.

8.6 Statistical power

So far, focus has been on the null hypothesis. The Type-I error probability was introduced, defined as the probability of incorrectly rejecting the null hypothesis, the control, or "cap" on which is α, usually set to 0.05. What if the null hypothesis is actually false and the study fails to reject it? This is termed a Type-II error, the control on which is denoted β, the probability of a Type-II error. **The complement of β is called statistical power.**

The following table summarizes the two types of errors and the two correct decisions that can occur in hypothesis testing. In the context of hypothesis testing, a Type-II error could be termed a false negative, not to be confused with false negatives occurring on individual case-level decisions.

Truth	Fail to reject NH	Reject NH
NH is True	1 - α	α (FPF)
NH is False	β (FNF)	Power = 1 - β

This resembles the 2 x 2 table encountered in (book) Chapter 02, which led to the concepts of FPF, TPF and the ROC curve. Indeed, it is possible think of an analogous plot of empirical (i.e., observed) power vs. empirical α, which looks like an ROC plot, with empirical α playing the role of FPF and empirical power playing the role of TPF, see below. If $\alpha = 0$, then power = 0; i.e., if Type-I errors are minimized all the way to zero, then power is zero and one makes Type-II errors all the time. On the other hand, if $\alpha = 1$ then Power = 1, and one makes Type-I errors all the time.

A little history is due at this point. The author's first FROC study, which led to his entry into this field (Chakraborty et al., 1986), was published in Radiology

in 1986 after a lot of help from a reviewer, who we (correctly) guessed was the late Prof. Charles E. Metz. Prof. Gary T. Barnes (my mentor at that time at the University of Alabama at Birmingham) and I visited Prof. Charles Metz in Chicago for a day ca. 1986, to figuratively "pick Charlie's brain". Prof. Metz referred to the concept outlined in the previous paragraph, as an *ROC within an ROC*.

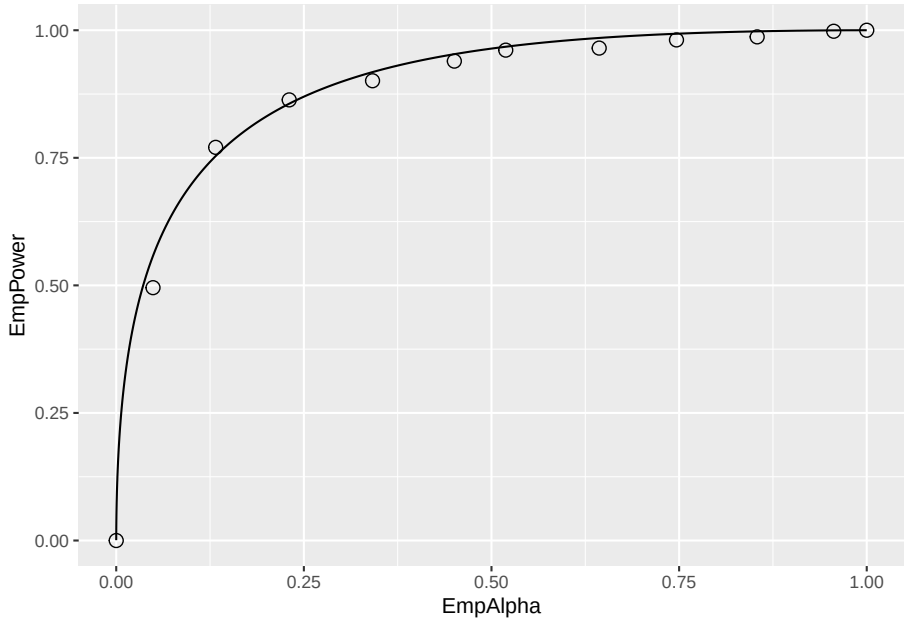
This curve does not summarize the result of a single ROC study. Rather it summarizes the probabilistic behavior of the two types of errors that occur when one conducts thousands of such studies, under both NH and AH conditions, each time with different values of α, with each trial ending in a decision to reject or not reject the null hypothesis. The long sentence is best explained with an example.

```
seed <- 1;set.seed(seed)
muNH <- 1.5;muAH <- 2.1;sigma <- 1.3;K1 <- 50;K2 <- 52# Line 6

# cheat to find the population mean and std. dev.
AUC <- array(dim = 10000) # line 8
for (i in 1:length(AUC)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = muNH, sd = sigma)
  AUC[i] <- Wilcoxon(zk1, zk2)
}
sigmaAUC <- sqrt(var(AUC));meanAUC <- mean(AUC) # Line 14

T <- 2000  # Line 16
mu <- c(muNH,muAH) # Line 17
alphaArr <- seq(0.05, 0.95, length.out = 10)
EmpAlpha <- array(dim = length(alphaArr))
EmpPower <- array(dim = length(alphaArr))
for (a in 1:length(alphaArr)) { # Line 20
  alpha <- alphaArr[a]
  reject <- array(0, dim = c(2, T))
  for (h in 1:2) {
    for (t in 1:length(reject[h,])) {
      zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = mu[h], sd = sigma)
      AUC <- Wilcoxon(zk1, zk2)
      obsvdZ <- (AUC - meanAUC)/sigmaAUC
      p <- 2*pnorm(-abs(obsvdZ)) # p value for individual t
      if (p < alpha) reject[h,t] = 1
    }
  }
  EmpAlpha[a] <- sum(reject[1,])/length(reject[1,])
  EmpPower[a] <- sum(reject[2,])/length(reject[2,])
}
EmpAlpha <- c(0,EmpAlpha,1); EmpPower <- c(0,EmpPower,1) # Line 19
```

```
pointData <- data.frame(EmpAlpha = EmpAlpha, EmpPower = EmpPower)
zetas <- seq(-5, 5, by = 0.01)
muRoc <- 1.8
curveData <- data.frame(EmpAlpha = pnorm(-zetas),
  EmpPower = pnorm(muRoc - zetas))
alphaPowerPlot <- ggplot(mapping = aes(x = EmpAlpha, y = EmpPower)) +
  geom_point(data = pointData, shape = 1, size = 3) +
  geom_line(data = curveData)
print(alphaPowerPlot)
```



Relevant line numbers are shown above as comments. Line 6 creates two variables, `muNH` = 1.5 (the binormal model separation parameter under the NH) and `muAH` = 2.1 (the separation parameter under the AH). Under either hypotheses, the same diseased case standard deviation `sigma` = 1.3 and 50 non-diseased and 52 diseased cases are assumed. As before, lines 8 – 14 use the "brute force" technique to determine population AUC and standard deviation of AUC under the NH condition. Line 16 defines the number of trials `T` = 2000. Line 17 creates a vector `mu` containing the NH and AH values defined at line 6. Line 18 creates `alphaArr`, a sequence of 10 equally spaced values in the range 0.05 to 0.95, which represent 10 values for α. Line 19 creates two arrays of length 10 each, named `EmpAlpha` and `EmpPower`, to hold the values of the observed Type-I error rate, i.e., empirical α, and the empirical power, respectively. The program will run `T` = 2000 NH and `T` = 2000 AH trials using as α each successive value in `alphaArr` and save the observed Type-I error rates and observed powers to

the arrays `EmpAlpha` and `EmpPower`, respectively.

Line 20 begins a for-loop in `a`, an index into `alphaArr.` Line 21 selects the appropriate value for `alpha` (0.05 on the first pass, 0.15 on the next pass, etc.). Line 22 initializes `reject[2,2000]` with zeroes, to hold the result of each trial; the first index corresponds to hypothesis `h` and the second to trial `t`. Line 23 begins a for-loop in `h`, with `h` = 1 corresponding to the NH and `h` = 2 to the AH. Line 24 begins a for-loop in `t`, the trial index. The code within this block is similar to previous examples. It simulates ratings, computes AUC, calculates the p-value, and saves a rejection of the NH as a one at the appropriate array location `reject[h,t]`. Lines 32 – 33 calculate the empirical α and empirical power for each value of α in `alphaArr`. After padding the ends with zero and ones (the trivial points), the remaining lines plot the "ROC within an ROC".

Each of the circles in the figure corresponds to a specific value of α. For example the lowest non-trivial corresponds to $\alpha = 0.05$, for which the empirical α is 0.049 and the corresponding empirical Power is 0.4955. True α increases as the operating point moves up the plot, with empirical α and empirical power increasing correspondingly. The AUC under this curve is determined by the effect size, defined as the difference between the AH and NH values of the separation parameter. If the effect size is zero, then the circles will scatter around the chance diagonal; the scatter will be consistent with the 2000 trials used to generate each coordinate of a point. As the effect size increases, the plot approaches the perfect "ROC", i.e., approaching the top-left corner. One could use AUC under this "ROC" as a measure of the incremental performance, the advantage being that it would be totally independent of α, but this would not be practical as it requires replication of the study under NH and AH conditions about 2000 times each and the entire process has to be repeated for several values of α. The purpose of this demonstration was to illustrate the concept behind Metz's profound remark.

It is time to move on to factors affecting statistical power in a single study.

8.6.1 Factors affecting statistical power

- Effect size: effect size is defined as the difference in AUC_{pop} values between the alternative hypothesis condition and the null hypothesis condition. Recall that AUC_{pop} is defined as the true or population value of the empirical ROC-AUC for the relevant hypothesis. One can use the "cheat method" to estimate it under the alternative hypothesis. The formalism is easier if one assumes it is equal to the asymptotic binormal model predicted value. The binormal model yields an estimate of the parameters, which only approach the population values in the asymptotic limit of a large number of cases. In the following, it is assumed that the parameters on the right hand side are the population values) It follows that effect size (ES) is given by (all quantities on the right hand side of Eqn. (8.13) are population values):

$$\text{AUC} = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right)$$

It follows that effect size (ES) is given by (all quantities on the right hand side of above equation are population values):

$$ES = \Phi\left(\frac{\mu_{AH}}{\sqrt{1+\sigma^2}}\right) - \Phi\left(\frac{\mu_{NH}}{\sqrt{1+\sigma^2}}\right)$$

```
EffectSize <- function (muNH, sigmaNH, muAH, sigmaAH)
{
  ES <- pnorm(muAH/sqrt(1+sigmaAH^2)) - pnorm(muNH/sqrt(1+sigmaNH^2))
  return (ES)
}

seed <- 1;set.seed(seed)
muAH <- 2.1 # NH value, defined previously, was mu = 1.5

T <- 2000
alpha <- 0.05 # size of test
reject = array(0, dim = T)
for (t in 1:length(reject)) {
  zk1 <- rnorm(K1);zk2 <- rnorm(K2, mean = muAH, sd = sigma)
  AUC <- Wilcoxon(zk1, zk2)
  obsvdZ <- (AUC - meanAUC)/sigmaAUC
  p <- 2*pnorm(-abs(obsvdZ)) # p value for individual t
  if (p < alpha) reject[t] = 1
}

ObsvdTypeIErrRate <- sum(reject)/length(reject)
CI <- c(0,0);width <- -qnorm(alpha/2)
CI[1] <- ObsvdTypeIErrRate -
  width*sqrt(ObsvdTypeIErrRate*(1-ObsvdTypeIErrRate)/T)
CI[2] <- ObsvdTypeIErrRate +
  width*sqrt(ObsvdTypeIErrRate*(1-ObsvdTypeIErrRate)/T)
cat("obsvdPower = ", ObsvdTypeIErrRate, "\n")
#> obsvdPower =  0.489
cat("95% confidence interval = ", CI, "\n")
#> 95% confidence interval =  0.4670922 0.5109078
cat("Effect Size = ", EffectSize(mu, sigma, muAH, sigma), "\n")
#> Effect Size =  0.08000617 0
```

The ES for the code above is 0.08 (in AUC units). It should be obvious that if effect size is zero, then power equals α. This is because then there is no

distinction between the null and alternative hypotheses conditions. Conversely, as effect size increases, statistical power increases, the limiting value being unity, when every trial results in rejection of the null hypothesis. The reader should experiment with different values of `muAH` to be convinced of the truth of these statements.

- Sample size: increase the number of cases by a factor of two, and run the above code chunk.

```
#> pop NH mean AUC =  0.8594882 , pop NH sigma AUC =  0.02568252
#> num. non-diseased images =  100 num. diseased images =  104
#> obsvdPower =  0.313
#> 95% confidence interval =  0.2926772 0.3333228
#> Effect Size =  0.08000617 0
```

So doubling the numbers of cases (both non-diseased and diseased) results in statistical power increasing from 0.509 to 0.844. Increasing the numbers of cases decreases σ_{AUC}, the standard deviation of the empirical AUC. The new value of σ_{AUC} is 0.02947, which should be compared to the value 0.04177 for K1 = 50, K2 = 52. Recall that σ_{AUC} enters the denominator of the Z-statistic, so decreasing it will increase the probability of rejecting the null hypothesis.

- Alpha: Statistical power depends on *alpha*. The results below are for two runs of the code, the first with the original value $\alpha = 0.05$, the second with $\alpha = 0.01$:

```
#> alpha =  0.05 obsvdPower =  0.1545
#> alpha =  0.01 obsvdPower =  0.0265
```

Decreasing α results in decreased statistical power.

8.7 Comments

The Wilcoxon statistic was used to estimate the area under the ROC curve. One could have used the binormal model, introduced in Chapter 06, to obtain maximum likelihood estimates of the area under the binormal model fitted ROC curve. The reasons for choosing the simpler empirical area are as follows. (1) With continuous ratings and 102 operating points, the area under the empirical ROC curve is expected to be a close approximation to the fitted area. (2) With maximum likelihood estimation, the code would be more complex – in addition to the fitting routine one would require a binning routine and that would introduce yet another variable in the analysis, namely the number of

bins and how the bin boundaries were chosen. (3) The maximum likelihood fitting code can sometimes fail to converge, while the Wilcoxon method is always guaranteed to yield a result. The non-convergence issue is overcome by modern methods of curve fitting described in later chapters. (4) The aim was to provide an understanding of null hypothesis testing and statistical power without being bogged down in the details of curve fitting.

8.8 Why alpha is chosen as 5%

One might ask why α is traditionally chosen to be 5%. It is not a magical number, rather the result of a cost benefit tradeoff. Choosing too small a value of α would result in greater probability $(1-\alpha)$ of the NH not being rejected, even when it is false. Sometimes it is important to detect a true difference between the measured AUC and the postulated value. For example, a new eye-laser surgery procedure is invented and the number of patients is necessarily small as one does not wish to subject a large number of patients to an untried procedure. One seeks some leeway on the Type-I error probability, possibly increasing it to $\alpha = 0.1$, in order to have a reasonable chance of success in detecting an improvement in performance due to better eyesight after the surgery. If the NH is rejected and the change is in the right direction, then that is good news for the researcher. One might then consider a larger clinical trial and set α at the traditional 0.05, making up the lost statistical power by increasing the number of patients on which the surgery is tried.

If a whole branch of science hinges on the results of a study, such as discovering the Higg's Boson in particle physics, statistical significance is often expressed in multiples of the standard deviation (σ) of the normal distribution, with the significance threshold set at a much stricter level (e.g. 5σ). This corresponds to $\alpha \sim 1$ in 3.5 million (`1/pnorm(-5)` = 3.5 x 10^6, a one-sided test of significance). There is an article in Scientific American (https://blogs.scientificamerican.com/observations/five-sigmawhats-that/) on the use of $n\sigma$, where `n` is an integer, e.g. 5, to denote the significance level of a study, and some interesting anecdotes on why such high significance levels (ie., small α) are used in some fields of research.

Similar concerns apply to manufacturing where the cost of a mistake could be the very expensive recall of an entire product line. For background on Six Sigma Performance, see http://www.six-sigma-material.com/Six-Sigma.html. An article downloaded 3/30/17 from https://en.wikipedia.org/wiki/Six_Sigma is included as supplemental material to this chapter (Six Sigma.pdf). It has an explanation of why 6σ translates to one defect per 3.4 million opportunities (it has to do with short-term and long-term drifts in a process). In my opinion, looking at other fields offers a deeper understanding of this material than simply stating that by tradition one adopts alpha = 5%.

Most observer performance studies, while important in the search for better

imaging methods, are not of such "earth-shattering" importance, and it is somewhat important to detect true differences at a reasonable alpha, so alpha = 5% and beta = 20% represent a good compromise. If one adopted a 5σ criterion, the NH would never be rejected, and progress in image quality optimization would come to a grinding halt. That is not to say that a 5σ criterion cannot be used; rather if used, the number of patients needed to detect a reasonable difference (effect size) with 80% probability would be astronomically large. Truth-proven cases are a precious commodity in observer performance studies. Particle physicists working on discovering the Higg's Boson can get away with 5σ criterion because the number of independent observations and/or effect size is much larger than corresponding numbers in observer performance research.

8.9 Discussion

In most statistics books, the subject of hypothesis testing is demonstrated in different (i.e., non-ROC) contexts. That is to be expected since the ROC-analysis field is a small sub-specialty of statistics (Prof. Howard E. Rockette, private communication, ca. 2002). Since this book is about ROC analysis, I decided to use a demonstration using ROC analysis. Using a data simulator, one can "cheat" by conducting a very large number of simulations to estimate the population AUC under the null hypothesis. This permitted us to explore the related concepts of Type-I and Type-II errors within the context of ROC analysis. Ideally, both errors should be zero, but the nature of statistics leads one to two compromises. Usually one accepts a Type-I error capped at 5% and a Type-II error capped at 20%. These translate to $\alpha = 0.05$ and desired statistical power = 80%. The dependence of statistical power on α, the numbers of cases and the effect size was explored.

In TBA Chapter 11 sample-size calculations are described that allow one to estimate the numbers of readers and cases needed to detect a specified difference in inter-modality AUCs with expected statistical power = $1 - \beta$. The word "detect" in the preceding sentence is shorthand for "reject the NH with incorrect rejection probability capped at α".

This chapter also gives the first example of validation of a hypothesis testing method. Statisticians sometimes refer to this as showing a proposed test is a "5% test". What is meant is that one needs to be assured that when the NH is true the probability of NH rejection is consistent with the expected value. Since the observed NH rejection rate over 2000 simulations is a random variable, one does not expect the NH rejection rate to exactly equal 5%, rather the constructed 95% confidence interval (also a random interval variable) should include the NH value with probability 1 - α.

Comparing a single reader's performance to a specified value is not a clinically interesting problem. The next few chapters describe methods for significance testing of multiple-reader multiple-case (MRMC) ROC datasets, consisting of

interpretations by a group of readers of a common set of cases in typically two modalities. It turns out that the analyses yield variability estimates that permit sample size calculation. After all, sample size calculation is all about estimation of variability, the denominator of the z-statistic. The formulae will look more complex, as interest is not in determining the standard deviation of AUC, but in the standard deviation of the inter-modality reader-averaged AUC difference. However, the basic concepts remain the same.

8.10 References

Bibliography

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415.

Barlow, W. E., Chi, C., Carney, P. A., Taplin, S. H., D'Orsi, C., Cutter, G., Hendrick, R. E., and Elmore, J. G. (2004). Accuracy of screening mammography interpretation by characteristics of radiologists. *Journal of the National Cancer Institute*, 96(24):1840–1850.

Barnes, G., Sabbagh, E., Chakraborty, D., Nath, P., Luna, R., Sanders, C., and Fraser, R. (1989). A comparison of dual-energy digital radiography and screen-film imaging in the detection of subtle interstitial pulmonary disease. *Investigative Radiology*, 24(8):585–591.

Beam, C. A., Layde, P. M., and Sullivan, D. C. (1996). Variability in the interpretation of screening mammograms by us radiologists. findings from a national sample. *Archives of Internal Medicine*, 156(2):209–13.

Berbaum, K. S., Dorfman, D. D., Franken, E. A., and Caldwell, R. T. (2002). An empirical comparison of discrete ratings and subjective probability ratings. *Academic Radiology*, 9(7):756–763.

Bochud, F., Abbey, C., and Eckstein, M. (1999). Visual signal detection in structured backgrounds iv, calculation of figures of merit for model observers in non-stationary backgrounds. *Journal of the Optical Society of America, A, Optics, Image Science, and Vision*, 17(2):206–17.

Burgess, A. E. (2011). Visual perception studies and observer models in medical imaging. In *Seminars in nuclear medicine*, volume 41, pages 419–436. Elsevier.

Chakraborty, D., Breatnach, E., Yester, M., Soto, B., Barnes, G., and Fraser, R. (1986). Digital and conventional chest imaging: A modified ROC study of observer performance using simulated nodules. *Radiology*, 158:35–39.

Chakraborty, D., Phillips, P., and Zhai, X. (2020). *RJafroc: Artificial Intelligence Systems and Observer Performance*. R package version 2.0.1.9000.

Chakraborty, D. P. (2017). *Observer Performance Methods for Diagnostic Imaging: Foundations, Modeling, and Applications with R-Based Examples*. CRC Press, Boca Raton, FL.

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44:837–845.

Dorfman, D. and Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals - rating-method data. *Journal of Mathematical Psychology*, 6:487–496.

Dorfman, D., Berbaum, K., Metz, C., Lenth, R., Hanley, J., and Abu Dagga, H. (1997). Proper receiving operating characteristic analysis: The bigamma model. *Acad. Radiol.*, 4(2):138–149.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, Boca Raton.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. Academic Press Series in Cognition and Perception. Academic Press, Inc., New York, first edition.

Fenton, J. (2015). Is it time to stop paying for computer-aided mammography? *JAMA Intern Med*.

Fenton, J. J., Taplin, S. H., Carney, P. A., Abraham, L., Sickles, E. A., D'Orsi, C., Berns, E. A., Cutter, G., Hendrick, R. E., Barlow, W. E., and Elmore, J. G. (2007). Influence of computer-aided detection on performance of screening mammography. *N Engl J Med*, 356(14):1399–1409.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley and Sons, New York.

Gur, D., Bandos, A. I., Cohen, C. S., Hakim, C. M., Hardesty, L. A., Ganott, M. A., Perrin, R. L., Poller, W. R., Shah, R., Sumkin, J. H., Wallace, L. P., and Rockette, H. E. (2008). The "laboratory" effect: Comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. *Radiology*, 249(1):47–53.

Hanley, J. A. (1988). The robustness of the "binormal" assumptions used in fitting ROC curves. *Med. Decis. Making*, 8(3):197–203.

Hanley, J. A. and Hajian-Tilaki, K. O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: An update. *Academic Radiology*, 4(1):49–58.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Hartmann, L. C., Sellers, T. A., Frost, M. H., Lingle, W. L., Degnim, A. C., Ghosh, K., Vierkant, R. A., Maloney, S. D., Pankratz, V. S., Hillman, D. W., et al. (2005). Benign breast disease and the risk of breast cancer. *New England Journal of Medicine*, 353(3):229–237.

Jiang, Y. and Metz, C. E. (2010). BI-RADS data should not be used to estimate ROC curves. *Radiology*, 256(1):29–31.

Kundel, H., Berbaum, K., Dorfman, D., Gur, D., Metz, C. E., and Swensson, R. G. (2008). Receiver operating characteristic analysis in medical imaging (icru report 79). Report, International Commission on Radiation Units and Measurments.

Larsen, R. J. and Marx, M. L. (2001). *An Introduction to Mathematical Statistics and Its Applications*. Prentice-Hall Inc, Upper Saddle River, NJ, 3rd edition.

Lusted, L. B. (1971). Signal detectability and medical decision making. *Science*, 171:1217–1219.

Macmillan, N. and Creelman, C. (1991). *Detection Theory: A User's Guide*. Cambridge University Press, New York.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50—60.

Metz, C. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.

Metz, C. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigative Radiology*, 24:234–245.

Metz, C. E. (1986). ROC methodology in radiologic imaging. *Investigative Radiology*, 21(9):720–733.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97.

Nishikawa, R. (2012). Estimating sensitivity and specificity in an ROC experiment. *Breast Imaging*, pages 690–696.

Noether, G. E. (1967). Elements of nonparametric statistics. Report, Wiley and Sons.

Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Philpotts, L. E. (2009). Can computer-aided detection be detrimental to mammographic interpretation? *Radiology*, 253(1):17–22.

Pisano, E., Gatsonis, C., Hendrick, E., Yaffe, M., Baum, J., Acharyya, S., Conant, E., Fajardo, L., Bassett, L., D'Orsi, C., Jong, R., and Rebner, M. (2005). Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med*, 353(17):1773–1783.

Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America*, 24(6):745–749.

Pollack, I. (1953). The information of elementary auditory displays. ii. *The Journal of the Acoustical Society of America*, 25(4):765–769.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (2007). *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge, 3 edition.

Rockette, H., Gur, D., and Metz, C. (1992). The use of continous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. *Investigative Radiology*, 27:169–172.

Skaane, P., Bandos, A. I., Gullien, R., Eben, E. B., Ekseth, U., Haakenaasen, U., Izadi, M., Jebsen, I. N., Jahr, G., and Krager, M. (2013). Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology*, 267(1):47–56.

Soh, B. P., Lee, W., McEntee, M. F., Kench, P. L., Reed, W. M., Heard, R., Chakraborty, D. P., and Brennan, P. C. (2013). Screening mammography: test set data can reasonably describe actual clinical reporting. *Radiology*, 268(1):46–53.

Stein, S. K. and Barcellos, A. (1992). *Calculus and analytic geometry*. McGraw-Hill Companies, 5 edition.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Series in Cognition and Perception. Academic Press, New York, first edition.

Thompson, M. L. and Zucchini, W. (1989). On the statistical analysis of roc curves. *Statistics in medicine*, 8(10):1277–1290.

USAirForce, R. (1947). A statistical theory of target detection by pulsed radar.

Wagner, R. F., Beiden, S. V., and Metz, C. E. (2001). Continuous versus categorical data for ROC analysis: Some quantitative considerations. *Academic Radiology*, 8(4):328–334.

Wilcoxon, F. (1945). Individual comparison by ranking methods. *Biometrics*, 1:80–83.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, 3(1):32–35.

Zhou, X.-H., Obuchowski, N. A., and McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine*. John Wiley and Sons, New York.