The RJafroc Roc Book

Dev P. Chakraborty, PhD

2023-03-09

Contents

Preface

TBA

0.1 Rationale and Organization

- Intended as an online update to my print book (Chakraborty, 2017).
- All references in this book to `RJafroc` refer to the R package with that name (case sensitive) (Chakraborty and Zhai, 2022).
- Since its publication in 2017 `RJafroc`, on which the `R` code examples in the print book depend, has evolved considerably causing many of the examples to "break" if one uses the most current version of `RJafroc`. The code will still run if one uses `RJafroc` 0.0.1 but this is inconvenient and misses out on many of the software improvements made since the print book appeared.
- This gives me the opportunity to update the print book.
- The online book has been divided into 3 books.
  - The RJafrocQuickStartBook book.
  - **This book:** RJafrocRocBook.
  - The RJafrocFrocBook book.

0.2 TBA Acknowledgements

Dr. Xuetong Zhai

Dr. Peter Phillips

Online Latex Editor at this site

Dataset contributors

0.3 Temporary comments

This is intended to allow successful builds when a needed file is not in the build. These are indicated by, for example:

Chapter `TempComment \@ref(proper-roc-models)`

Fix these on final release.

Overview

Chapter 1

An overview of the field

1.1 How much finished

75%

1.2 Introduction

> The question addressed by this book is "how good are radiologists using medical imaging devices at diagnosing disease?" Equivalently, "how good is a computer aided detection (CAD) algorithm at detecting cancers in breast images?", or "how good is an artificial intelligence (AI) algorithm at detecting objects of interest?".

Observer performance measurements originally developed for medical imaging, and 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, signal detection theory as commonly studied by psychologists and engineers, human observer visual system modeling and statistics. Its importance in medical imaging is due to the evolution of technology and the need to objectively assess such 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 my experience, basic aspects of the subject are sometimes misunderstood and lessons from the past are sometimes forgotten.

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

This chapter provides background material and an outline 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 using simple objects in optimizing the design. Once the device is deployed medical physicists working in hospitals use phantom-based quality control measurements to maintain image quality. Lacking the complexity of clinical images, phantom measurements are not expected to 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 as yet restricted to relatively simple tasks 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. An outline 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 and 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 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 image 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 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. 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[1].

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

[1] instead of "normal" and "abnormal", or "noise" and "signal plus noise", or "target absent" and "target present", etc

"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 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. Again, 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) 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. 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. 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 developed 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 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. 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™ 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 noise. 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 an instrument called a micro-densitometer, which scans 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, which 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), the 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. Therefore, as far as image noise is concerned 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 detector introduces its own noise, due to the electronics. Furthermore, there could be engineering limitations preventing attainment of the theoretical limit. Through much iteration the designers reach 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 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 micro-densitometer the trace should show an abrupt transition as one crosses the edge of the phantom. In practice, the transition is spread out, resembling a sigmoid function. This is due to several factors. The finite size of the focal spot producing the x-rays has a penumbra effect, which blurs the edge. The spread of light, within the detector due to its finite thickness also blurs the edge. The screen absorbs photons and converts them to electrons; this process adds some blur. 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 micro-densitometer. 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 contrast of 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. 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. The spatial 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, generally involve a group of about five or six radiologists interpreting a set of patient images acquired on the prototype and on the existing standard (conventional modality). The truth (is the image of a diseased patient?) is unknown to them but must be know to the researcher. 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.

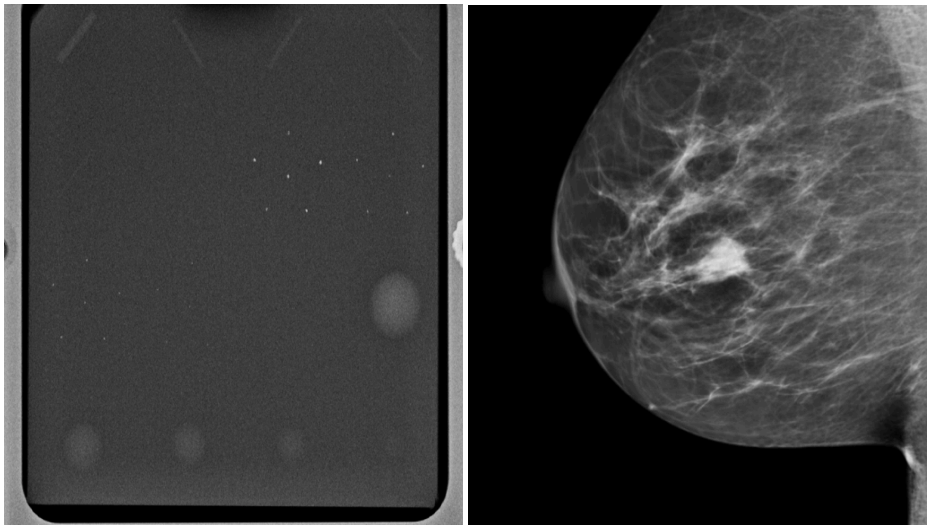


Figure 1.1: (A) Image of an ACR phantom, (B) Clinical image.

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 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 anthropomorphic 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 outline 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 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(Kundel et al., 2008, Fryback and Thornbury (1991)). 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 Thornbury 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?

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 diagnos-

tic 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.

1.10.2 How to use the book

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.

1.11 Summary

1.12 Discussion

1.13 Chapter References

ROC paradigm

Chapter 2

The Binary Task

2.1 TBA How much finished

90%

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 et al., 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 one can define 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 derive relations between the different types of measures. An `R` example of calculation of these quantities is given.

Table 2.1: Truth Table.

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

2.3 The 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 viewed.

Let D represent the radiologist's decision with D = 1 representing the decision "case is diagnosed as non-diseased" and D = 2 representing the decision "case is diagnosed as 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 number of decisions in each of the four categories defined in Table 2.1. 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 $n(N) + n(P)$, which must also equal K_1+K_2, the total number of cases denoted K. These statements are summarized in Eqn. (2.1).

Table 2.2: Individual 2x2 table cell counts and row and column sums. K_1 is the number of non-diseased cases, K_2 is the number of diseased cases and K is the total number of cases.

	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)

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

2.4 Sensitivity and specificity

The notation P(D|T) indicates the probability of diagnosis D for a given truth state T. The vertical bar is used to denote a conditional probability, i.e., the probability of what is to the left of the bar occurring when what is to the right of the bar is true:

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

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

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

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

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

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

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

Since the diagnosis must be either $D = 1$ or $D = 2$, the following must be true:

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

Equivalently, these equations can be written:

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

Comments:

- An easy way to remember Eqn. (2.7) is to start by writing down 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 yields unity, i.e., the first equation in Eqn. (2.7).

- Because there are two equations in four unknowns, only two of the four probabilities 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 $P(TP)$, 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):

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

  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 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 realizations of random variables. Intuitively, an estimate calculated over a larger number of cases is expected to be closer to the 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 denominator is the number of diseased cases K_2 and for specificity, the appropriate denominator is the number of non-diseased cases K_1. The estimation equations for sensitivity specificity are (estimates are often denoted by the "hat" or circumflex symbol $\widehat{\ }$):

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

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

2.5 Disease prevalence

Disease prevalence, often abbreviated to prevalence, is defined as the 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:

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

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

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

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.

$$\left.\widehat{P(\mathrm{D}|\mathrm{pop})}=\frac{K_2}{K_1+K_2}\right\} \tag{2.12}$$

It is important to appreciate the distinction between population prevalence and the 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:

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

2.6 Accuracy

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

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

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

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

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, for example, combine a population estimate of prevalence with a laboratory estimate of sensitivity.

Eqn. (2.15) can be understood from the following argument. $\widehat{Sp}$ is the fraction of correct decisions on non-diseased cases. Multiplying this by $\widehat{P(!D)}$ yields the

fraction of correct negative decisions on all cases. Similarly $\widehat{Se}$ is the fraction of correct decisions on diseased cases. Multiplying this by $\widehat{P(D)}$ yields the fraction of correct positive decisions on all cases. Their sum is the fraction of correct decisions on all cases.

A formal mathematical derivation follows: Eqn. (2.8) yields:

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

Therefore,

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

For the population one can dispense with the carets:

$$\left.\begin{aligned} \mathrm{Ac} &= \frac{n(TN)+n(TP)}{K} \\ &= \frac{K_1\mathrm{Sp}+K_2\mathrm{Se}}{K} \\ &= \mathrm{SpP(!D)}+\mathrm{Se}\widehat{P(D)} \end{aligned}\right\} \tag{2.18}$$

2.7 Negative and positive predictive values

Sensitivity and specificity have desirable characteristics insofar as they reward the observer for correct decisions on diseased and non-diseased cases, respectively. They are expected to be independent of disease prevalence as one is dividing by the relevant denominator. However, radiologists interpret cases in a "mixed" situation where cases could be diseased or non-diseased. Therefore 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. As an example 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 be generating false positives most of the time. The radiologist who makes too many false positives would know this 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, and 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 not interested in specificity, the normalized probability of a correct decision on non-diseased cases; rather the radiologist's interest is in the probability that a patient diagnosed as non-diseased is actually non-diseased. The reader may notice how the conditioning variable in two conditional probabilities are reversed – more on this later. Likewise, the radiologist is not interested in sensitivity, the normalized probability of a correct decision on diseased cases; rather the radiologist's 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 the probability, given a non-diseased diagnosis, that the patient is actually non-diseased:

$$\text{NPV} = P(T = 1|D = 1) \tag{2.19}$$

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

$$\text{PPV} = P(T = 2|D = 2) \tag{2.20}$$

Note that the conditioning in both equations are reversed from those in the definition of specificity and sensitivity, namely $\text{Sp} = P(D = 1|T = 1)$ and $\text{Se} = P(D = 2|T = 2)$.

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,

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

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

$$\widehat{\text{NPV}} = \frac{\widehat{P(TN)}}{\widehat{P(TN)} + \widehat{P(FN)}} \tag{2.22}$$

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

$$\widehat{P(TN)} = \widehat{P(!D)}(1 - \widehat{FPF}) \tag{2.23}$$

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 $\widehat{P(FN)}$ 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., $(\widehat{P(D)}$:

$$\widehat{P(FN)} = \widehat{P(D)}(1 - \widehat{TPF}) \tag{2.24}$$

Putting this all together, one has:

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

For the population one can dispense with the carets:

$$\text{NPV} = \frac{P(!D)(1 - FPF)}{(P(!D)(1 - FPF) + (P(D)(1 - TPF)} \tag{2.26}$$

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

$$\text{PPV} = \frac{P(D)(TPF)}{P(D)(TPF) + P(!D)FPF} \tag{2.27}$$

The equations defining NPV and PPV are actually special cases of Bayes' theorem (Larsen and Marx, 2005). The theorem is:

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(!A)P(B|!A)} \Bigg\} \tag{2.28}$$

An easy way to remember Eqn. (2.28) is to start with the numerator on the right hand side which has the "reversed" form of the desired conditioning on the left hand side. If the desired probability is $P(A|B)$ one starts with the

"reversed" form of the conditioning, i.e., $P(B|A)$, and multiplies 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 $P(!A)$.

2.8 Examples: 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?
- What is Accuracy?

```
1  # disease prevalence in
2  # USA screening mammography
3  prev <- 0.005 # Line 3
4  FPF <- 0.1 # typical operating point
5  TPF <- 0.8 #        do:
6  sp <- 1-FPF
7  se <- TPF
8  NPV <- (1-prev)*(sp)/((1-prev)*(sp)+prev*(1-se))
9  PPV <- prev*se/(prev*se+(1-prev)*(1-sp))
10 cat("NPV = ", NPV, "\nPPV = ", PPV, "\n")
11 #> NPV =  0.9988846
12 #> PPV =  0.03864734
13 ac <-(1-prev)*sp+prev*se
14 cat("accuracy = ", ac, "\n")
15 #> accuracy =  0.8995
```

- Line 3 initializes the variable `prev`, 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 sp and se.
- Line 8 calculates `NPV` using Eqn. (2.26).
- Line 9 calculates `PPV` using Eqn. (2.27).
- Line 13 calculates accuracy using Eqn. (2.18)

2.8.1 Comments

If a woman has a negative diagnosis, chances are very small that she has breast cancer: the probability that the radiologist is incorrect in the negative diagnosis is $1-\text{NPV} = 0.0011154$. Even is she has a positive diagnosis the probability that she actually has cancer is still only $\text{PPV} = 0.0386473$. That is why following

a positive screening diagnosis the woman is recalled for further imaging and if that reveals cause for reasonable suspicion 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 Ac = 0.8995. The numbers in this illustration are for expert radiologists.

2.8.2 PPV and NPV are irrelevant to laboratory tasks

According to the hierarchy of assessment methods described in (book) Chapter 01, TBA Table 1.1, PPV and NPV are level- 3 measurements. These are calculated from "live" interpretations as opposed to retrospectively calculated quantities like sensitivity and specificity. The radiologist adjusts the reporting threshold to achieve a balance between sensitivity and specificity. The balance depends critically on the expected disease prevalence. Based on geographical location and type of practice, the radiologist over time develops a feel for 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 from increased sensitivity.

In a laboratory study where one uses enriched case sets, the concepts of NPV and PPV are meaningless. For example, it would be impossible 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 enriched 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 meaning. No diagnostic decisions are riding on laboratory interpretations of retrospectively acquired patient images. In contrast PPV and NPV do have clinical meanings when calculated from large population based "live" studies. For example, the (Fenton et al., 2007) study sampled 684,956 women and used the results of "live" interpretations. 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.

2.9 Discussion

This chapter introduced the terms sensitivity, specificity, disease prevalence, positive and negative predictive values and accuracy. Due to its strong dependence on disease prevalence, accuracy is a relatively poor measure of performance. Radiologists generally have a good understanding of positive and negative predictive values, as these terms are relevant in the clinical context, being in effect their "batting averages". They do not care as much for sensitivity and specificity. A caveat on the use of PPV and NPV calculated from laboratory studies is noted.

2.10 Chapter References

Chapter 3

Modeling the binary Task

3.1 How much finished 95%

3.2 Introduction

Chapter 3 introduced measures of performance associated with the binary task. Described in this chapter is a 2-parameter statistical model for the binary task. It shows how one can predict quantities like sensitivity and specificity based on the values of the parameters of a statistical model of the binary task. Introduced are the fundamental concepts of a **decision variable** and a **reporting threshold** or simple **threshold** that occur frequently in this book. It is shown that the reporting threshold can be altered by varying the experimental conditions. The receiver-operating characteristic (ROC) plot is introduced. It is shown how the dependence of sensitivity and specificity on the reporting threshold can be exploited by a measure of performance that is independent of reporting threshold.

The dependence of variability of the operating point on the numbers of cases is examined. The concept of random sampling is introduced and it is shown that the results become more stable with larger numbers of cases, i.e., larger sample sizes. These are perhaps intuitively obvious concepts but it is important to see them demonstrated, Online Appendix 3.A. Formulae for confidence intervals for estimates of sensitivity and specificity are derived and the calculations are shown explicitly using `R` code.

3.3 Decision variable and reporting threshold

The model for the binary task involves three assumptions:

1. Each case has an associated decision variable sample.

2. The observer (e.g., radiologist) adopts a case-independent reporting threshold.

3. An adequate number of training session(s) are used to get the observer to a steady state.

4. The observer is "blinded" to the truth of each case 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 value of a random scalar *sensory variable* which is a unidirectional measure of evidence of disease.

- The sensory variable is sensed internally by the observer and as such cannot be directly measured. Alternative terminology includes "psychophysical variable", "perceived variable", "perceptual variable" or "confidence level". The last term is the most common. It is a subjective variable since it depends 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 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 is used and specific realized values are termed z-samples. It is a random variable in the sense that it varies randomly from case to case.

- The decision variable rank-orders cases with respect to evidence for presence of disease. Unlike a traditional rank-ordering scheme, where "1" is the highest rank, the scale is inverted with larger values corresponding to greater evidence of disease. Without loss of generality, one assumes that the decision variable ranges from $-\infty$ to $+\infty$ with large positive values indicative of strong evidence for presence of disease and large negative values indicative of strong evidence for absence of disease. The zero value indicates no evidence for presence or absence of disease. [1] In this book such a decision scale, with increasing values corresponding to increasing evidence of disease, is termed **positive-directed**.

[1] The $-\infty$ to $+\infty$ scale is not an assumption. The decision variable scale could just as well range from a to b, where a $<$ b; with appropriate re-scaling of the decision variable, there will be no changes in the rank-orderings, and the scale can be made extend from $-\infty$ to $+\infty$.

3.3.2 Existence of a reporting threshold

Assumption 2: The radiologist adopts a single case-independent reporting threshold ζ and states: "case is diseased" if $z \geq \zeta$ or "case is non-diseased" if $z < \zeta$.

- Unlike the random Z-sample, which varies from case to case, the reporting threshold is held fixed for the duration of the study. In some of the older literature the reporting threshold is sometimes referred to as "response bias". The term "bias" which has a negative connotation, whereas, in fact, the choice of reporting threshold depends on a rational assessment of costs and benefits of different outcomes.
- The choice of reporting 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 after 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 is assumed to determine after a new training session the optimal threshold for the new conditions and henceforth holds it constant for the duration of the second study.

From Assumption 2, it follows that:

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

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

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, therefore 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. The diseased cases are not homogenous: in some the 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 in this chapter, 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 reporting threshold ζ, both Se and Sp are under the control of the observer. The lower the reporting threshold the higher the sensitivity and the lower the specificity and the converse is also true. Ideally both sensitivity and specificity should be large, i.e., unity. The tradeoff between sensitivity and specificity says that there is no "free lunch": 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 distributions is entirely consistent with not knowing which distribution a specific z-sample is coming from.

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

3.4 Changing the reporting threshold: I

Suppose that in the first study a radiologist interprets a set of cases subject to the instructions that it is important to identify actually diseased cases and to worry less 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 and that the case set contains equal numbers of non-diseased and diseased cases and the radiologist is informed of these experimental conditions. It is also assumed that the radiologist is allowed to reach a steady state and responds rationally to the payoff agreement. Under these circumstances, the radiologist is expected to set the reporting 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 reporting 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 90 percent and specificity is 10 percent. 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 pay-off is \$1 for each TP decision and \$10 for each TN decision. Suppose, further, that sufficient time has elapsed from the previous study so 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 reporting 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 10 percent and specificity is 90 percent.

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 reporting threshold: 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 in the experimental study, i.e., the experimenter is not deceiving the radiologist. [2] In other words, only five out of every 1000 cases are actually diseased. This information will cause the radiologist to adopt a high threshold thereby becoming more reluctant to state: "case is diseased".

[2]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 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.

By simply diagnosing all cases as non-diseased, i.e., without using any image information, the radiologist will be correct on every disease absent case and earn $995, which is 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 reporting 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 disease prevalence, the reputation of the radiologist, malpractice, etc.

3.6 The equal-variance binormal model

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

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

The constraint $\mu \geq 0$ 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. [3] A more general model termed the unequal-variance binormal model is generally used for modeling human observer data, discussed later, but for the moment,

[3] 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.

one does not need that complication. The equal-variance binormal model is defined by:

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

In Eqn. (3.3) the subscript t denotes the truth 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 state of the case. The label k_t ranges from $1, 2, ..., K_t$ where K_t is the number of cases with disease state t.

As you can see I am departing from the usual convention in this field which labels the cases with a single index k ranging 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 the other 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 calling 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.

A few facts concerning the normal distribution are summarized next.

3.7 The normal distribution

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 $\phi(z|\mu, \sigma)$. The special case $N(0, 1)$ is referred to as the **unit normal distribution**; it has zero mean and unit variance and the corresponding pdf is denoted $\phi(z)$.

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 $N(\mu,\sigma^2)$ falls between the fixed limits z and z + dz. For this reason $\phi(z|\mu,\sigma)$ is termed the probability density function. 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 by Φ (sometimes the term probability distribution function is used for what we are terming the CDF). The function $\Phi(z)$ 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()` the density of the unit 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 dashed line corresponds to the reporting threshold $\zeta = 1$. The area under the pdf to the left of ζ equals the value of CDF at the selected ζ (`pnorm(1)` = 0.841).

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.

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

The inverse of $\Phi(z)$ is that function which when applied to p yields the upper limit z in Eqn. (3.7), i.e.,

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

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

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

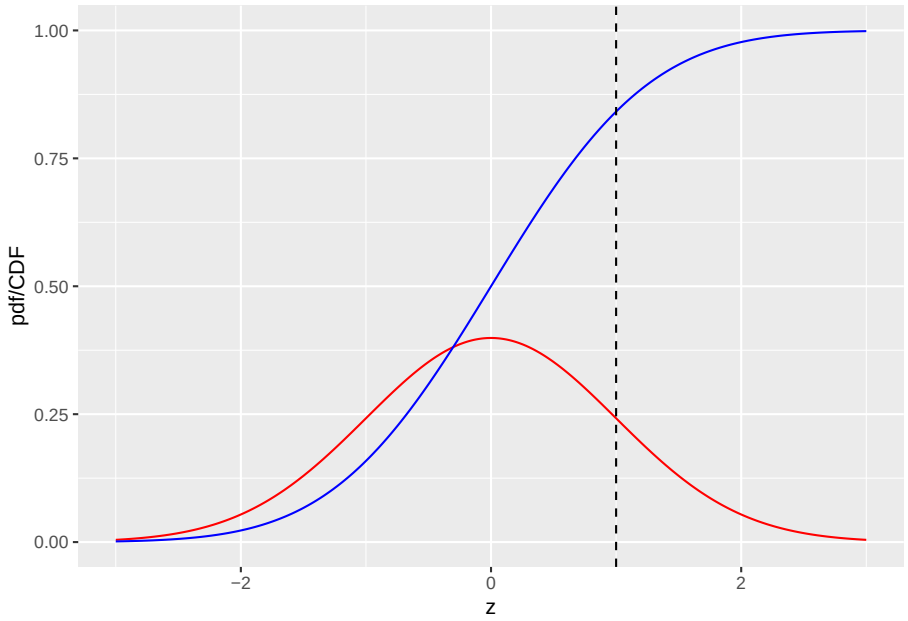


Figure 3.1: pdf-CDF plots for the unit normal distribution. The red curve is the pdf and the blue line is the CDF. The dashed line is the reporting threshold $\zeta = 1$.

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

To summarize the convention used in `R`, `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.

```
1 qnorm(0.025)
2 #> [1] -1.959964
3 qnorm(1-0.025)
4 #> [1] 1.959964
5 pnorm(qnorm(0.025))
6 #> [1] 0.025
7 qnorm(pnorm(-1.96))
8 #> [1] -1.96
```

Line 1 demonstrates the identity:

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

Line 3 demonstrates the identity:

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

Lines 5 and 7 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 95 percent of the unit-normal pdf, these two values can be used to estimate 95 percent confidence intervals.

> 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 95 percent confidence interval, and the observed value plus 1.96 defines the upper limit of its 95 percent confidence interval.

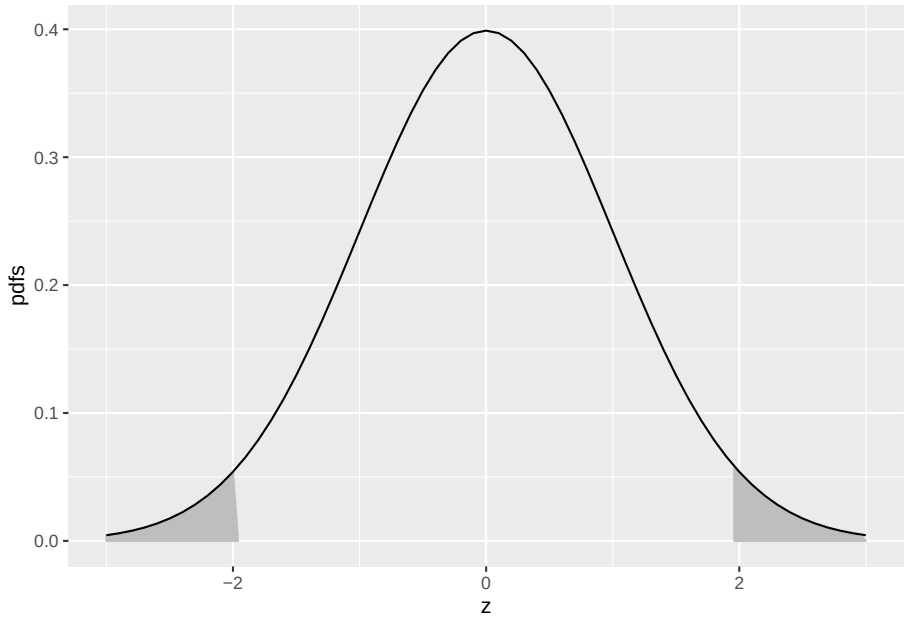


Figure 3.2: Illustrating that 95 percent of the total area under the unit normal pdf is contained in the range $|Z| < 1.96$, which can be used to construct a 95 percent confidence interval for an estimate of a suitably normalized statistic. The area in each shaded tail is 0.025.

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:

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

The expression for sensitivity can be derived. Consider that the random variable obtaining by shifting the origin to μ. A little thought should convince the reader that $Z_{k_2 2} - \mu$ is distributed as $N(0,1)$. Therefore, the desired probability is:

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

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:

$$\left.\begin{aligned} 1 - \Phi(\zeta) = & -\Phi(\zeta) \\ 1 - \Phi(\zeta - \mu) = & \Phi(\mu - \zeta) \end{aligned}\right\} \tag{3.14}$$

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 Eqn. (3.15) the threshold ζ appears with different signs because specificity is the area under a pdf to the left of the threshold while sensitivity is the area to the right.

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 first system is unambiguously better than the second. But if sensitivity is greater for the

first but specificity is greater for the second the comparison is ambiguous. 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 performancer measure, termed a figure of merit (FOM). Eqn. (3.15) can be solved for μ as follows. Inverting the equations yields:

$$\left.\begin{aligned} \zeta &= \Phi^{-1}\left(\mathrm{Sp}\left(\zeta\right)\right) \\ \mu - \zeta &= \Phi^{-1}\left(\mathrm{Se}\left(\zeta\right)\right) \end{aligned}\right\} \tag{3.16}$$

Eliminating ζ yields:

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

This is a useful relation, as it converts a *pair* of numbers into a *scalar* performance measure. Now it is almost trivial to compare two modalities: the one with the higher μ is better. In reality, the comparison is not trivial since like sensitivity and specificity μ has to be estimated from a finite dataset and one must account for sampling variability.

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., 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.

See Appendix 3.14 for a demonstration of the concepts of sensitivity and specificity using `R`.

3.9 Inverse variation of sensitivity and specificity

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 $\mathrm{Sp} = \Phi(\zeta)$ and sensitivity is $\mathrm{Se} = \Phi(\mu - \zeta)$. Since the threshold ζ appears with different signs the dependence of sensitivity on ζ will be the opposite of that of specificity. 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 lowered threshold, in other words fewer non-diseased cases are correctly diagnosed as non-diseased. Sensitivity increases because more of the

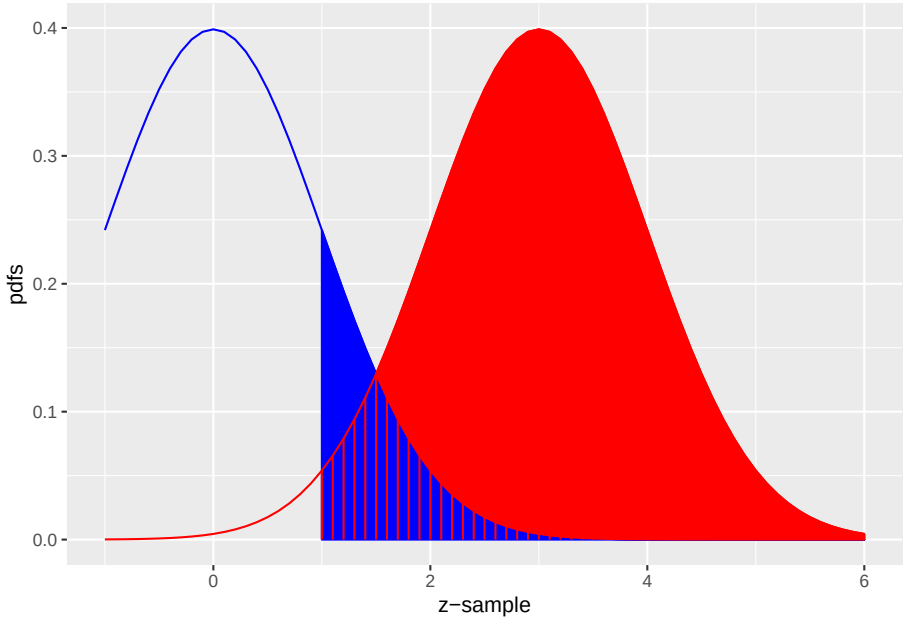


Figure 3.3: The equal-variance binormal model for $\mu = 3$ and $\zeta = 1$; the blue curve, centered at zero, is the pdf of non-diseased cases and the red one, centered at $\mu = 3$, is the pdf of diseased cases. The left edge of the blue shaded region represents the threshold $\zeta = 1$. The red shaded 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.

diseased distribution lies to the right of the lowered threshold, in other words more diseased cases are correctly diagnosed as diseased.

If Observer 1 has higher sensitivity than Observer 2 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 Observer 2 can be coaxed into adopting a lower threshold, thereby decreasing specificity to match that of Observer 1 then it is possible that the Observer 2's sensitivity, formerly smaller, could (and here is the ambiguity because if might not happen) now be greater than that of Observer 1.

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.

3.10 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:

$$\left.\begin{aligned} \mathrm{FPF}\,(\zeta) &\equiv 1 - \mathrm{Sp}\,(\zeta) \\ &= \Phi\,(-\zeta) \\ \mathrm{TPF}\,(\zeta) &\equiv \mathrm{Se}\,(\zeta) \\ &= \Phi\,(\mu - \zeta) \end{aligned}\right\} \tag{3.18}$$

Specifying ζ selects a particular operating point on this curve 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). Note that as ζ increases the operating point moves down the curve. The operating point $O(\zeta|\mu)$ for the equal variance binormal model is:

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

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. 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 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.

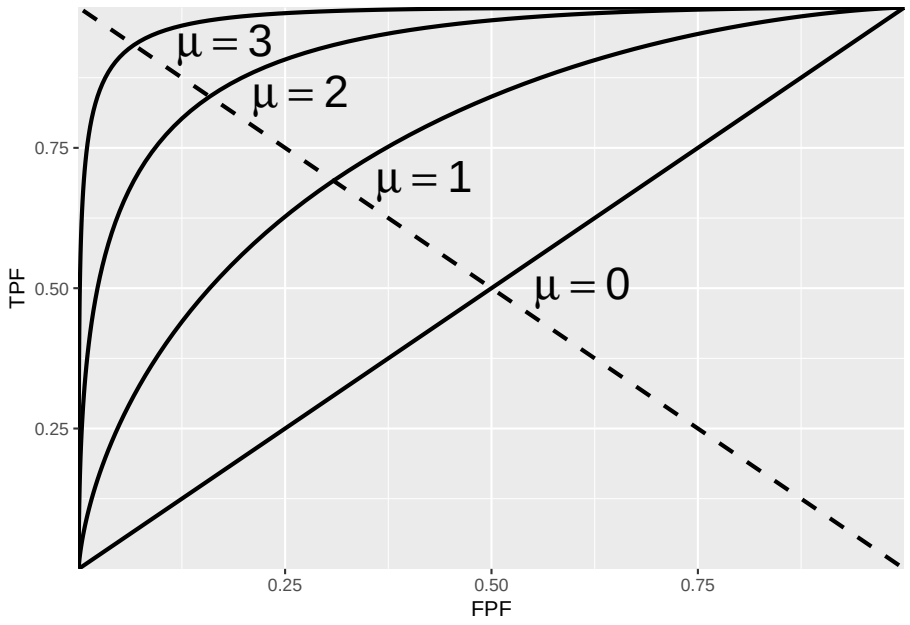


Figure 3.4: ROC plots predicted by the equal variance binormal model for different values of μ. As μ increases the intersection of the curve with the negative diagonal moves closer to the ideal operating point, (0,1) at which sensitivity and specificity are both equal to unity.

3.10.1 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. 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:

$$\text{TPF}(\zeta) = \text{FPF}(\zeta) = \Phi(-\zeta)$$

In this 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 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.

3.10.2 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: 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). [4] 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.

3.10.3 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, i.e., the 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:

$$\left.\begin{array}{l} \text{FPF}(\zeta) = \Phi(-\mu/2) \\ \text{TPF}(\zeta) = \Phi(\mu/2) \end{array}\right\} \tag{3.20}$$

[4] Many cases are assumed as otherwise, due to sampling variability, the operating point will not be on the theoretical ROC curve.

The first equation implies:

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

Therefore,

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

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

3.10.4 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 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 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 unequal-variance binormal model predicted ROC curve. The more expressive term area under curve or AUC is used to include this and other methods of estimating the area under the ROC curve.

In Eqn. (3.22), the extra subscript $\sigma = 1$ is necessary to distinguish it from another that 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, A_z must be between zero and one. If μ is non-negative, $A_{z;\sigma=1}$ 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 $\mu = \infty$ yields $A_{z;\sigma=1} = 1$.

> Since it is a scalar quantity, A_z can be used to unambiguously quantify performance than is possible using sensitivity - specificity pairs.

3.10.5 Properties of the equal-variance binormal model ROC curve

1. The ROC curve is completely contained within the unit square. This follows from the fact that both axes of the plot are probabilities.

2. The operating point rises monotonically from (0,0) to (1,1).

3. 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.

4. A_z is a monotone increasing function of μ. It varies from 0.5 to 1 as μ varies from zero to infinity.

3.10.6 Comments

Property 2: since the operating point can both be expressed in terms of Φ functions, which are monotone in their arguments, and in each case the argument ζ appears with a negative sign it follows that as ζ is lowered both TPF and FPF increase. The operating point corresponding to $\zeta - d\zeta$ is to the upper right of that corresponding ζ to (assuming $d\zeta > 0$).

Property 3: 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 $\mu > 0$, 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 $\mu = 3$. The entire curve is defined by varying ζ. Specifying a particular value of ζ corresponds to specifying a particular point on the ROC curve. In Fig. TBA 3.5 the open circle corresponds to the operating point (0.159, 0.977) defined by $\zeta = 1$: `pnorm(-1)` = 0.159; pnorm(3-1) = 0.977. The operating point lies exactly on the curve as this is a predicted operating point.

3.10.7 Physical interpretation of the mu-parameter

The μ parameter is equivalent (Macmillan and Creelman, 2004) 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 ease of detectability of the disease. SNR is a term widely used in engineering, specifically in signal detection theory (Green et al., 1966; Egan, 1975). It dates to the early 1940s when one had the problem (Marcum, 1947, Marcum (1960)) of detecting faint radar reflections from a plane against a background of noise. The radar radio "receiver" is the origin of the term in Receiver Operating Characteristic.

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.

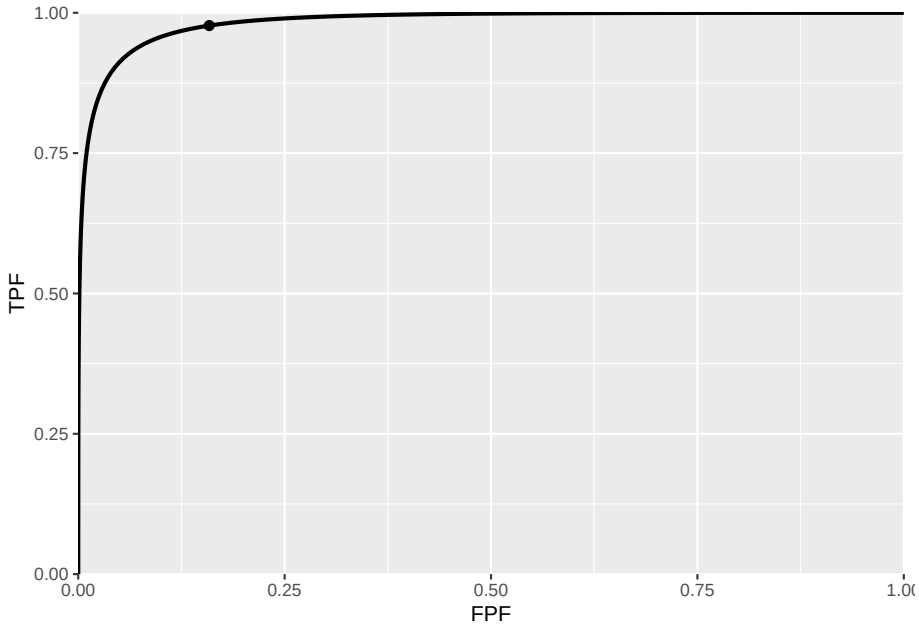


Figure 3.5: ROC curve predicted by equal variance binormal model for $\mu = 3$. The circled operating point corresponds to $\zeta = 1$. The operating point falls exactly on the curve, as these are analytical curves. With finite numbers of cases this is not observed in practice.

```
#> pnorm(3/sqrt(2)) =  0.983
```

For electrical signals, SNR can be measured with instruments but, in the context of decisions made by humans, what is 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 expertise.

> To this day I find it remarkable that an objective SNR-like quantity can be teased out of subjective observer decisions.

3.11 Confidence intervals for an operating point

- A $(1-\alpha)$ confidence interval (CI) of a statistic is the range that is expected to contain the true value with probability $(1-\alpha)$.
- It should be clear that a 99 percent CI is wider than a 95 percent CI, and that a 90 percent CI 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 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" represents 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.

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

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

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

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):

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

For a discrete distribution, one has probability *mass* function in contrast to a continuous distribution where 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:

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

It follows that FPF has mean $\widehat{\text{FPF}}$ and variance σ^2_{FPF} given by (since $Var(aX) = a^2 Var(X)$ where a is a constant):

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

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

$$FPF \sim N\left(\widehat{\text{FPF}}, \sigma^2_{\text{FPF}}\right)\Big\} \tag{3.31}$$

Eqn. (3.31) allows us to write down the approximate symmetric confidence interval for $\widehat{\text{FPF}}$, i.e., $\pm z_{\alpha/2} \times \sigma_{\text{FPF}}$ around $\widehat{\text{FPF}}$.

$$CI^{\text{FPF}}_{1-\alpha} = \left(\widehat{\text{FPF}} - z_{\alpha/2}\sigma_{\text{FPF}}, \widehat{\text{FPF}} + z_{\alpha/2}\sigma_{\text{FPF}}\right) \tag{3.32}$$

In Eqn. (3.32) $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the unit normal distribution: it is defined such that the area to the *right* under the unit normal distribution

pdf from $z_{\alpha/2}$ to $+\infty$ equals $\alpha/2$. For example $z_{0.025} = 1.96$, see Fig. 3.2. In general $z_{\alpha/2} = -\Phi^{-1}(\alpha/2)$. For example `-qnorm(0.025)` $= 1.96$.

These relations involving $z_{\alpha/2}$ follow:

$$\left.\begin{aligned} z_{\alpha/2} &= \Phi^{-1}(1-\alpha/2) \\ &= -\Phi^{-1}(\alpha/2) \\ \alpha/2 &= \int_{z_{\alpha/2}}^{\infty} \phi(z)dz \\ &= 1-\Phi(z_{\alpha/2}) \\ &= \Phi(-z_{\alpha/2}) \end{aligned}\right\} \tag{3.33}$$

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

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

$$CI_{1-\alpha}^{\text{TPF}} = \left(\widehat{\text{TPF}} - z_{\alpha/2}\sigma_{\text{TPF}}, \widehat{\text{TPF}} + z_{\alpha/2}\sigma_{\text{TPF}}\right) \tag{3.34}$$

In Eqn. (3.34),

$$\sigma_{\text{TPF}}^2 = \frac{\widehat{\text{TPF}}\left(1-\widehat{\text{TPF}}\right)}{K_2} \tag{3.35}$$

The confidence intervals are largest when the probabilities (FPF or TPF) are close to 0.5 and decrease inversely as the square root of the relevant number of cases. The symmetric binomial distribution based estimates can stray outside the allowed range (0 to 1). Exact confidence intervals 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:

```
#> alpha =  0.05
#> K1 =  99
#> K2 =  111
#> mu =  5
#> zeta =  2.5
#> Specificity =  0.99
#> Sensitivity =  0.991
#> Approx 95 percent CI for Specificity =  0.97 1.01
#> Exact 95 percent CI for Specificity =  0.945 1
#> Approx 95 percent CI for Sensitivity =  0.973 1.01
#> Exact 95 percent CI for Sensitivity =  0.951 1
```

Table 3.1: 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 21 percent compared to 53 percent for sensitivity and 63 percent for specificity).

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

Note that the approximate confidence intervals can stray outside the allowed range but the exact confidence intervals do not.

3.12 Variability: the Beam 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 sampled (stratified sampling) enriched screening cases containing 45 cases with cancer and the rest with benign lesions. Ground truth for these women had been established either by biopsy or by 2-year follow-up.

The observed range of sensitivity (TPF) was 53 percent and the range of FPF was 63 percent; the corresponding range for the AUC was 21 percent, Table 3.1. Empirical AUC was estimated using a 5-point BIRADS ratings of the images (the zero category was not allowed). Explanation of empirical AUC is deferred to Chapter 5.

In Fig. 3.6 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 μ parameters) of radiologists B and C are similar. The only difference between them is that radiologist B is using lower threshold than 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. 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 superior performance to B or C. No change in threshold is going to get the other two to a similar level of performance

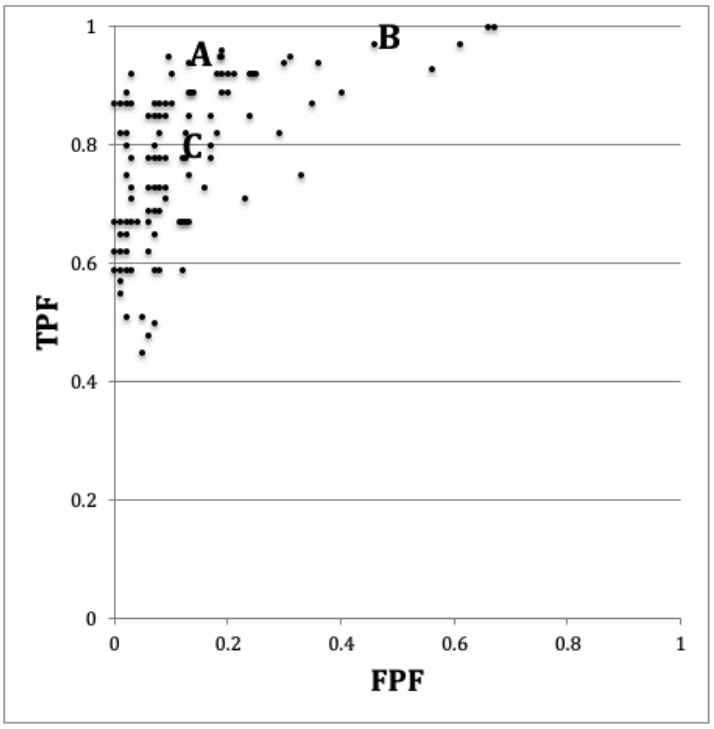


Figure 3.6: Schematic patterned from the Beam et al study showing the ROC operating points of 108 mammographers. Wide variability in sensitivity and specificity are evident while AUC is less variable. See text below.

as radiologist A. Extensive training will be needed to bring the under-performing radiologists to the expert level of radiologist A.

Fig. 3.6 and Table 3.1 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 differences between readers.

3.13 Discussion

Sensitivity and specificity are widely used in the medical imaging literature. It is important to realize that they do not provide a complete picture of diagnostic performance, since they represent performance at a particular observer-dependent threshold. As demonstrated in Fig. 3.6 expert observers can and do operate at different thresholds. If using sensitivity and specificity the dependence on reporting threshold often makes it difficult to unambiguously compare observers. An additional source of variability is introduced by the varying thresholds.

The ROC curve and AUC are completely defined by the μ parameter of the equal variance binormal model. Since both are independent of reporting threshold they overcome the ambiguity inherent in comparing sensitivity/specificity pairs. 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 μ parameter. The latter has the physical meaning of a perceptual signal to noise ratio.

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 AUCs. 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.

3.14 Appendix I

3.14.1 Estimates from a finite sample

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

```
#> seed =  100
#> mu true =  1.5
#> zeta true =  0.75
#> 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 equal to the true value. In fact, all of the estimates, sensitivity, specificity and μ are subject to sampling variability.

3.14.2 Changing the seed variable

No matter how many times one runs the above code, one always sees the same output shown above. This is because 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. Changing the `seed` to 101 yields:

```
#> seed =  101
#> mu true =  1.5
```

```
#> zeta true =  0.75
#> 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 cases. The effect is quite large (estimated sensitivity falls from 0.909 to 0.545 and estimated μ falls from 2.56 to 0.879) because the size of the relevant case set, $K_2 = 11$ for sensitivity, is small.

3.14.3 Increasing the numbers of cases

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

```
#> seed =  100
#> mu true =  1.5
#> zeta true =  0.75
#> K1 =  90
#> K2 =  110
#> Specificity =  0.778
#> Sensitivity =  0.836
#> Est. of mu =  1.74
```

Next we change `seed` to 101.

```
#> seed =  101
#> mu true =  1.5
#> zeta true =  0.75
#> 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 3.2 illustrates this trend with increasing sample size.

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.

Table 3.2: 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

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

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 3.2, 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 3.2, 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 large 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 and hence yields less variability. One uses the ratings paradigm

which is more efficient than the binary paradigm used in this chapter. Finally, one takes advantage of correlations that exist between the interpretations using matched-case matched-reader interpretations in two modalities; this tends to decrease variability in the AUC-difference even further (most applications of ROC methods involved detecting differences in AUCs not absolute values).

3.15 Chapter References

Chapter 4

Ratings Paradigm

4.1 How much finished 90%

4.2 Introduction

In Chapter 2 the binary paradigm and associated concepts, primarily sensitivity and specificity, were introduced. Chapter 3 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 - true and false positives. 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 as obtained using the binary task, and moreover, it is more efficient.

In this method the observer assigns a rating (i.e., an ordered label) to each case. Described first is a typical ROC counts table and how operating points (i.e., pairs of FPF and TPF values) are obtained 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 formulae for calculating operating points are presented. 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 labels is illustrated with an example. A method of collecting ROC data on a 6-point scale is described that has the advantage of also yielding an unambiguous single

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

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.

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 table introduced in Table 2.1. [1]

Table 4.1 is a representative counts table for a 5-rating study. 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:

- 1 received the "5" rating,
- 2 the "4" rating,
- eight the "3" rating,
- 19 the "2" rating and

[1] This type of data representation is sometimes called a frequency table, but frequency means a rate of number of events per some unit of time, so I prefer the clearer term "counts".

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

- 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,
- 5 the "3" rating,
- 6 the "2" rating and
- 5 the "1" rating.

A little thought should convince the reader that the observed tilting of the counts is reasonable.

> The spread of the counts 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 the ratings are ordered labels, and further modeling is required, see Chapter 6, that uses only the ordering information implicit in the labels, not the actual values, to reach quantitative conclusions about the spread of each distribution.

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) 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 $\text{FPF}_{\geq 5} = 1/60 = 0.017$. 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 $\text{TPF}_{\geq 5} = 22/50 = 0.440$. 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., $\text{FPF}_{\geq 4} = 3/60 = 0.05$. 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., $\text{TPF}_{\geq 4} = 34/50 = 0.680$. 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 assumptions they are called **empirical** 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 $K_{1;r}$ denote the number of non-diseased cases rated r, and $K_{2;r}$ the number of diseased cases rated r. Define dummy counts $K_{1;R+1} = K_{2;R+1} = 0$, where R is the number of ROC bins (and $R = 5$ in the current example). This allows inclusion of the origin (0,0) in the formulae. The new range of r is defined as $r = 1, 2, ..., (R + 1)$. Within each truth-state, non-diseased or diseased, the individual bin counts sum to the total number of non-diseased and diseased cases, respectively. The following equations summarize these statements:

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

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

$$K_{1;R+1} = K_{2;R+1} = 0$$

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

The operating points are defined by:

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

4.4.1 Labeling the points

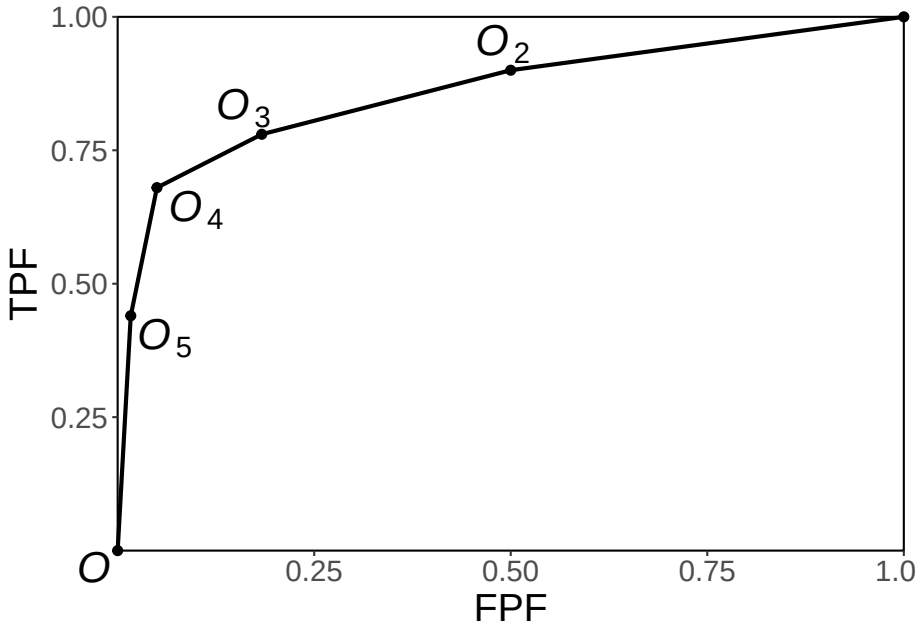


Figure 4.1: Operating point labeling convention and the empirical ROC plot.

The labeling O_n of the points follows the following convention, Fig. 4.1: from Eqn. (4.1) the point O_1 corresponding to $r = 1$ would be 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_2, which corresponds to $r = 2$; the next lower-left point is labeled O_3, corresponding to $r = 3$, etc., and the point labeled O_5 is the lowest non-trivial operating point corresponding to $r = 5$ and finally O_{R+1} corresponding to $r = R + 1 = 6$ 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_2, and working down the curve each time increasing the subscript by one. The maximum number of non-trivial points is $R - 1$.

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 and finally this 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 Implemention in code

It is useful to replace the preceding detailed explanation with a simple algorithm, as in the following code (lines 1 - 5):

```
1  options(digits = 3)
2  FPF <- OpPts[1,]
3  TPF <- OpPts[2,]
4  df <- data.frame(FPF = FPF, TPF = TPF)
5  df <- t(df)
6  print(df)
7  #>       [,1] [,2]  [,3] [,4] [,5]
8  #> FPF 0.0167 0.05 0.183  0.5    1
9  #> TPF 0.4400 0.68 0.780  0.9    1
10 # ev = equal variance model
11 mu_ev <- qnorm(.5)+qnorm(.9);sigma_ev <- 1
12 Az_ev <- pnorm(mu_ev/sqrt(2))
13 cat("uppermost point based estimate of mu_ev = ", mu_ev, "\n")
14 #> uppermost point based estimate of mu_ev =  1.28
15 cat("corresponding estimate of Az_ev = ", Az_ev, "\n")
16 #> corresponding estimate of Az_ev =  0.818
```

Notice that the values of the arrays `FPF` and `TPF` are identical to those listed in Table 4.2. Regarding lines 10 and 11 of code it was shown in Chapter 2 that in the equal variance binormal model the operating point determines the parameters $\mu_{ev} = 1.282$, Eqn. (3.17), or equivalently $A_{z;\sigma=1} = 0.818$, Eqn. (3.23). These 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 panel, and for comparison, the right panel is the unequal variance model fit to all operating points (the unequal variance model is described in the next chapter).

```
# equal variance fit to uppermost operating point
p1 <- plotROC (mu_ev, sigma_ev, FPF, TPF)
# the following values are from unequal-variance model fitting
# to be discussed in the next chapter
mu <- 2.17;sigma <- 1.65
# following formula is discussed in the next chapter
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)
```

It should come as no surprise that the uppermost operating point in panel A is *exactly* on the predicted curve: after all, this point was used to calculate μ_{ev} = 1.282. The corresponding value of ζ can be calculated from Eqn. (3.15), namely:

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

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

These are coded below:

```
# Sp = 1 - FPF = 1 - 0.5
# Se = 0.9
cat("zeta from Sp = ", qnorm(1-0.5),"\n")
#> zeta from Sp =  0
cat("zeta from Se = ", mu_ev-qnorm(0.9),"\n")
#> zeta from Se =  0
```

Either way, one gets the same result: $\zeta = 0$. This should make sense as FPF = 0.5 is consistent with half of the (symmetrical) unit-normal non-diseased distribution being above $\zeta = 0$.

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

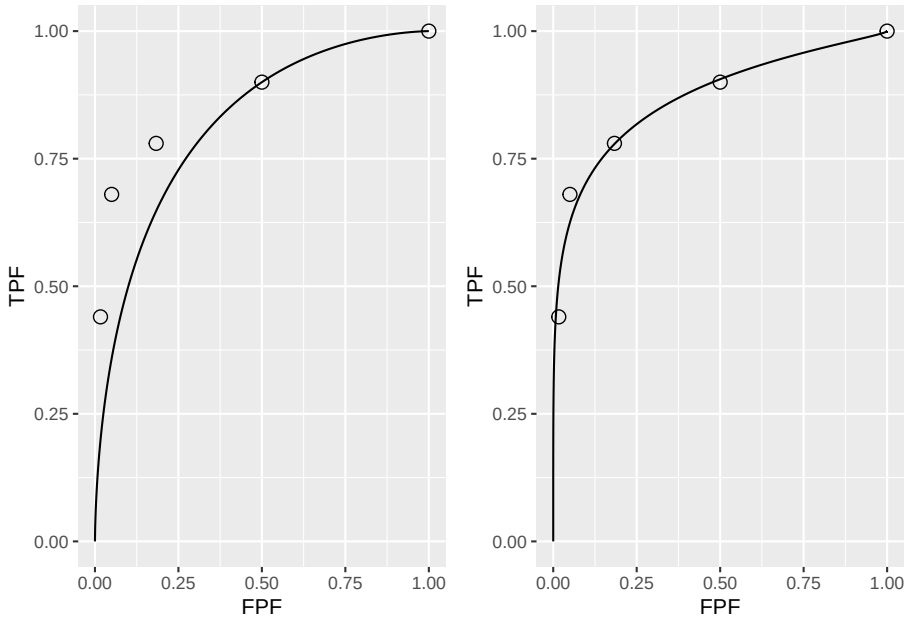


Figure 4.2: A: This panel is the predicted ROC curve for $\mu = 1.28$ superposed on the operating points. B: This panel is the same data fitted with a unequal variance model described in a following chapter.

- In Fig. 4.2 panel A show that 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_{ev} and $A_{z;\sigma=1}$ the values would have been different and the new curve would pass exactly through the new selected point. No single-point based choice of μev would yield a satisfactory visual fit to all the observed operating points. This is the reason one needs a modified model, namely the unequal variance binormal model, to fit radiologist data.
- Fig. 4.2 panel B shows the predicted ROC curve by the unequal variance binormal model, to be introduced in a following chapter. The corresponding parameter values are $\mu = 2.17$ and $\sigma = 1.65$.
- Notice the improved visual quality of the fit, especially if one keeps in mind that each observed point is subject to sampling variability, see Eqn. (3.32) and Eqn. (3.34). [2]

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 binning rule Eqn. (4.2): if $z > \zeta$ assign a rating of 2 and otherwise assign a rating of 1.

> 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. 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 in previous sessions.

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:

[2] The estimates in the preceding chapter were for a single operating point. Since the multiple operating points discussed here are correlated — some of the counts used to calculate them are common to two or more operating points — the method described in the previous chapter overestimates the confidence intervals. A modeling approach accounts for these correlations and yields narrower confidence intervals.

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

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

$$\left.\begin{array}{l} \zeta_r = r + 1 \\ r = 0, 1, ..., R-1 \\ \zeta_0 = -\infty \\ \zeta_R = \infty \end{array}\right\} \tag{4.3}$$

The empirical thresholds are integers, not floating point values predicted by Eqn. (4.5). Either way one gets the same operating points. This is a subtle and important distinction explained further in the next section.

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 $\zeta_4 \le 5 < \zeta_5$, 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 $\zeta_3 \le 4 < \zeta_4$. 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 cumulates counts in bins 3, 2 and 1, respectively. The 1-bin yields a trivial operating point (1,1). 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.

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 very 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 increased 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, which they do not.

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, 2004). The first question is "is the case diseased?" The binary (Yes/No) response to this question allows unambiguous calculation of the operating point. 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. 4.3. 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 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 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 Chapter 1.2, patient images are displayed one patient at a time. A fifth paradigm involves presentation of multiple images simultaneously to the observer, where one image is from a diseased patient, and the rest are from non-diseased patients. The observer's task is to pick the image 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

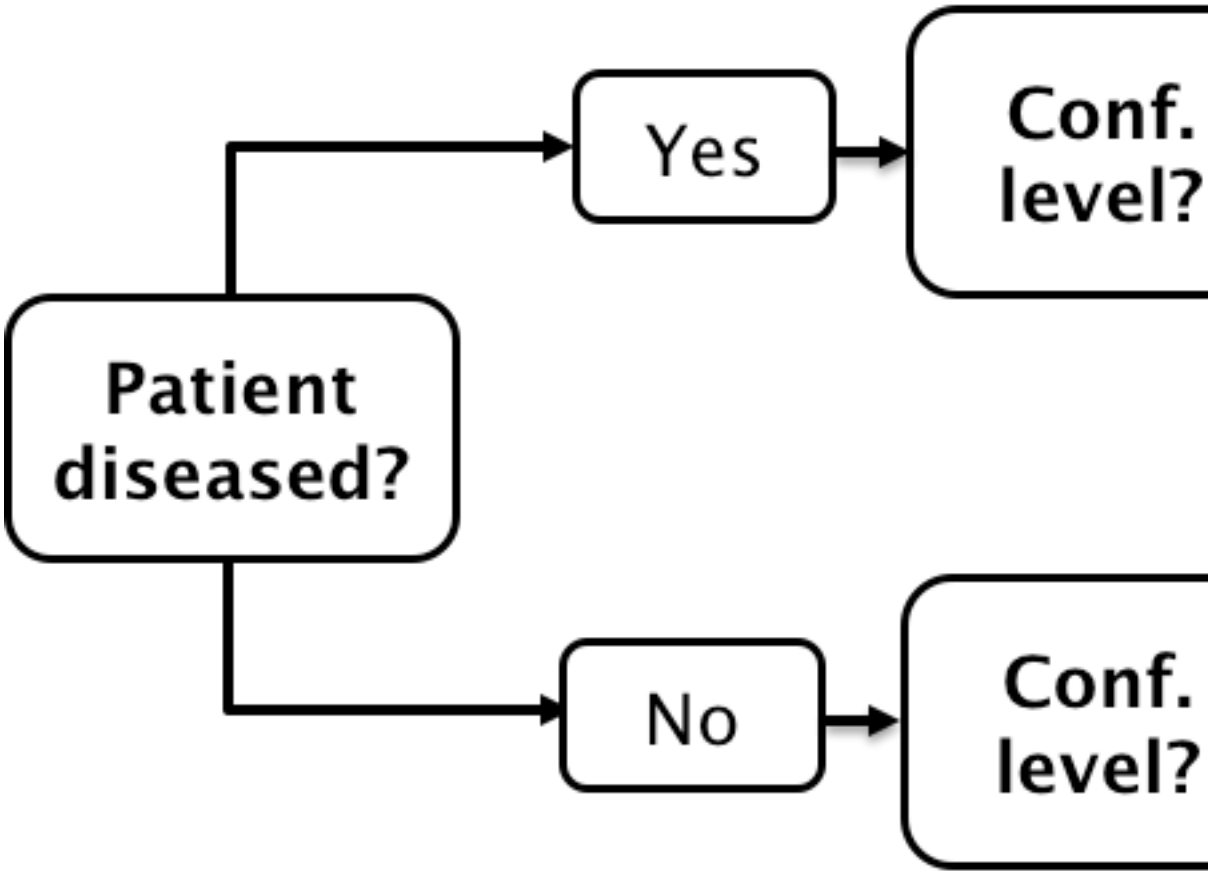


Figure 4.3: A method for acquiring ROC data on an effectively 6-point scale that also yields an unambiguous single operating point for declaring patients diseased. Note the reversal of the ordering of the final ratings in the last "column" in the lower half of the figure.

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 et al., 1966). If only two cases are presented, one diseased and the other non-diseased, then n = 2 and the task is 2AFC. In Fig. 4.4, in the left image a Gaussian nodule is superposed on a square region extracted from a non-diseased mammogram. The right image is a region extracted from a different non-diseased mammogram (one should not use the same background in the two images – the analysis assumes that different, i.e., independent images, are shown). If the observer clicks on the left image, a correct choice is recorded. [In some 2AFC-studies, the backgrounds are simulated non-diseased images. They resemble mammograms; the resemblance depends on the expertise of the observer: expert radiologists can tell that they are not true mammograms. They are actually created by filtering the random white noise with a $1/f^3$ spatial filter (Burgess, 2011).]

The 2AFC paradigm is popular because its analysis is straightforward and there is a theorem (Green et al., 1966) that $P(C)$, the probability of a correct choice in the 2AFC task equals the 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, I cannot recommend it for clinical studies because it does not resemble any clinical task. Additionally, the forced-choice paradigm is wasteful of known-truth images, often a difficult/expensive resource to come by, because narrower confidence intervals are obtained using the ratings ROC method or by utilizing location specific extensions of the ROC paradigm.

Fig. 4.4: 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 to the true area under the ROC curve.

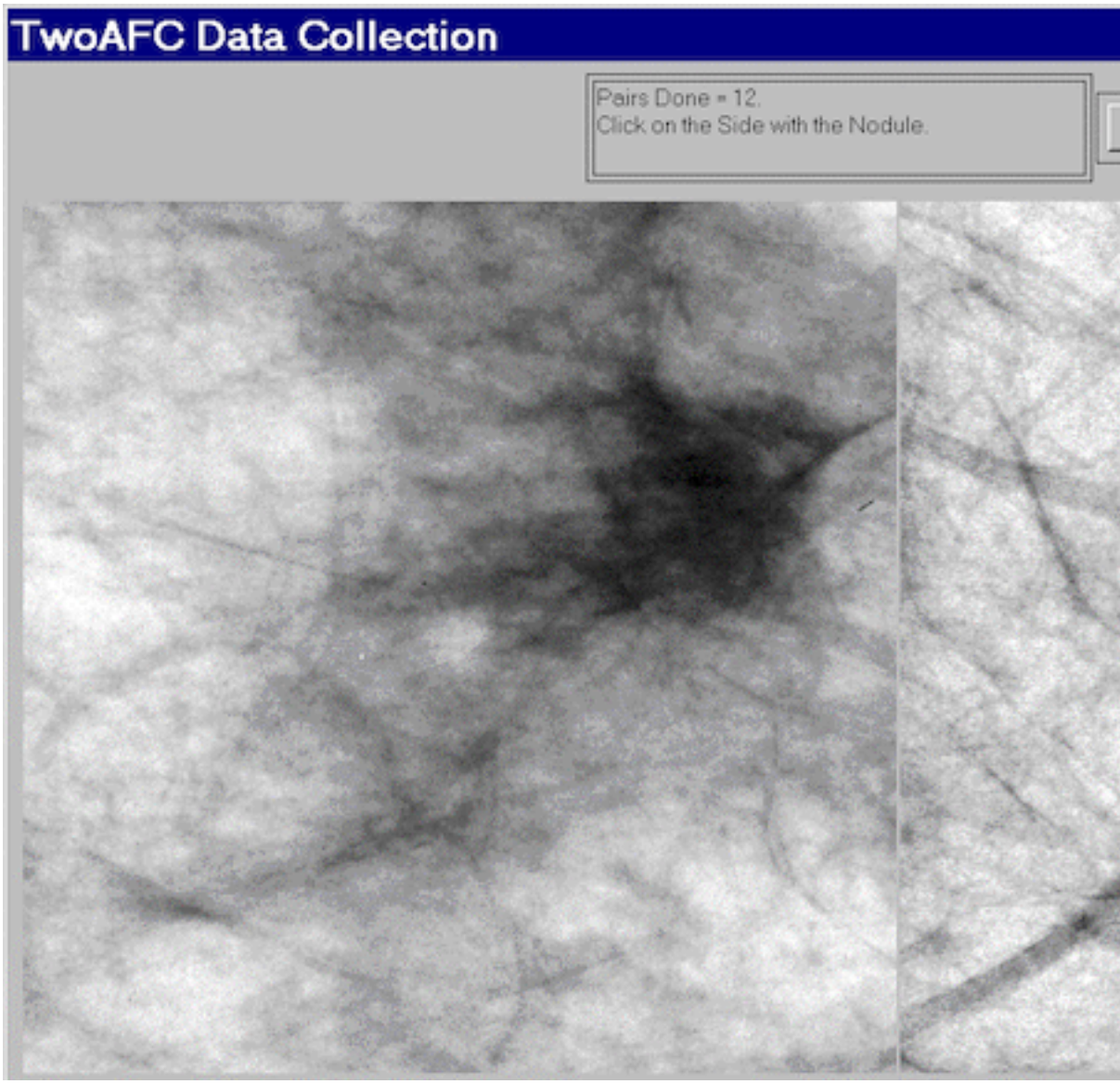


Figure 4.4: Example of image presentation in a 2AFC study.

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.

- 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. Usually laboratory ROC studies are conducted on retrospectively acquired images. Patients, whose images are 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. The correlation is taken to be an axiomatic truth by researchers when, in fact, it is an assumption.

- This section should not be interpreted as expressing my lack of trust in laboratory studies. 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. 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 radiologists' brings their full expertise to bear on each image interpretation) and real clinical images.

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). 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) and new ROC study designs sometimes tend to follow their advice. I recommend a 6-point rating scale as outlined in Fig. 4.3. This section provides the background for the recommendation.

- A widely cited (39,704 citations at the time of writing, ca. March 2023) 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 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."

- DPC comment: 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.
- 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."

- DPC comment: In the medical imaging context, a trivial way to increase the number of ratings would be to color-code the images: e.g., 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 colors encode 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, addressed by Miller (see above). 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. 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 study designer. Expert radiologists will not modify their reading style and one should be suspicious when they do 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. [3]

- 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 cannot fit it (in simple terms, the program "crashes" or reports unphysical parameter values). This occurs 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 thus reduce instances of degeneracy (this is actually done in some ROC software to overcome this problem). Another possibility is to use the empirical (trapezoidal) area under the ROC curve, which can always be calculated. Actually, fitting methods now exist that are robust to data degeneracy, such as discussed in Chapter `TempComment \@ref(proper-roc-models)` and in the RSM fitting chapter in TBA `RJafrocFrocBook`, so this reason for acquiring continuous data is no longer valid.

- 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. 4.3.

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

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, the American College of Radiology has developed the Breast Imaging Reporting and

[3] Dr. Fraser famously termed – off the record – our statistical analysis in (Chakraborty et al., 1986) "gobbledygook".

Data System (BI-RADS) to standardize mammography reporting (Liberman and Menell, 2002). There are six assessment categories: category 0 indicates need for additional imaging; category 1 is a negative (clearly non-diseased); category 2 is a benign finding; category 3 is probably benign, with short-interval follow-up suggested; category 4 is a suspicious abnormality for which biopsy should be considered; category 5 is highly suggestive of malignancy and appropriate action should be taken. The 4th edition of the BI-RADS manual 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.

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 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, the 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.

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

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 non-zero 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 Chapter `TempComment \@ref(proper-roc-models)`, 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. The fitting algorithm implemented in PROPROC software **reorders** confidence levels assumed by the binormal model, Chapter `TempComment \@ref(proper-roc-models)`, paragraph following TBA 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 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 are expected to be predictive of cancer.

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 explained. 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, 2004) to the two-question six-point scale in Fig. 4.3 was for a rating study on performance in recognizing odors.

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 Chapter 6.

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 – and the rating paradigm 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.

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 Chapter 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 to a parametric model is to derive analytical expressions for the separation parameter μ of the model and the area AUC under the curve. It was shown, see Fig. 4.2, that the equal variance binormal model did not fit a clinical dataset and that an unequal variance binormal model yielded a better visual fit. This turns out to be an almost universal finding. Before getting into the complexity of the unequal variance binormal model curve fitting, it is appropriate to introduce a simpler empirical approach, which is very popular with researchers in this field.

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 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 statistic known as the Wilcoxon is described.

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				

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

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 using 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.

To address any cell (i.e., case) in Table 5.1 one needs two indices: the row number t and the column number $k_t t$. Since the column number depends on the value of t one needs two indices to specify it: specifically, $k_t t$ denotes the column number k_t of a case with truth index t. [1]

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(ζ) and empirical true positive fraction TPF(ζ) are defined by:

[1] Alternative notation commonly 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 easily to more complex data structures, e.g., FROC, I prefer the two-index notation.

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

Here $I(x)$ is the indicator function which 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:

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

The difference between Eqn. (5.1) and Eqn. (3.18) is that the former is nonparametric while the latter is parametric.

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:

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

Its coordinates are defined by:

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

For example,

$$\left.\begin{aligned} \mathrm{FPF}_4 \equiv \mathrm{FPF}\left(\zeta_4\right) &= \frac{1}{K_1}\sum_{k_1=1}^{K_1} I\left(z_{k_1 1} \geq \zeta_4\right) \\ \mathrm{TPF}_4 \equiv \mathrm{TPF}\left(\zeta_4\right) &= \frac{1}{K_2}\sum_{k_2=1}^{K_2} I\left(z_{k_2 2} \geq \zeta_4\right) \\ O_4 \equiv \left(\mathrm{FPF}_4, \mathrm{TPF}_4\right) &= (0.017, 0.44) \end{aligned}\right\} \tag{5.5}$$

Fig. 4.1 is the empirical ROC plot. It illustrates the convention used to label the operating points introduced earlier is, i.e., O_2 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 $O_0 \equiv (\mathrm{FPF}_0, \mathrm{TPF}_0) = (1, 1)$ and $O_R \equiv (\mathrm{FPF}_R, \mathrm{TPF}_R) = (0, 0)$, where $R = 5$.

5.5 AUC under the empirical ROC plot

Fig. 5.1 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 can write equations to calculate the total area but there is a theorem (see below) that the empirical area is equal to a 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 summing the areas of the triangle and trapezoids or performing planimetry.

5.6 The Wilcoxon statistic

A statistic is any value calculated from observed data. The Wilcoxon statistic is defined by:

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

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

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

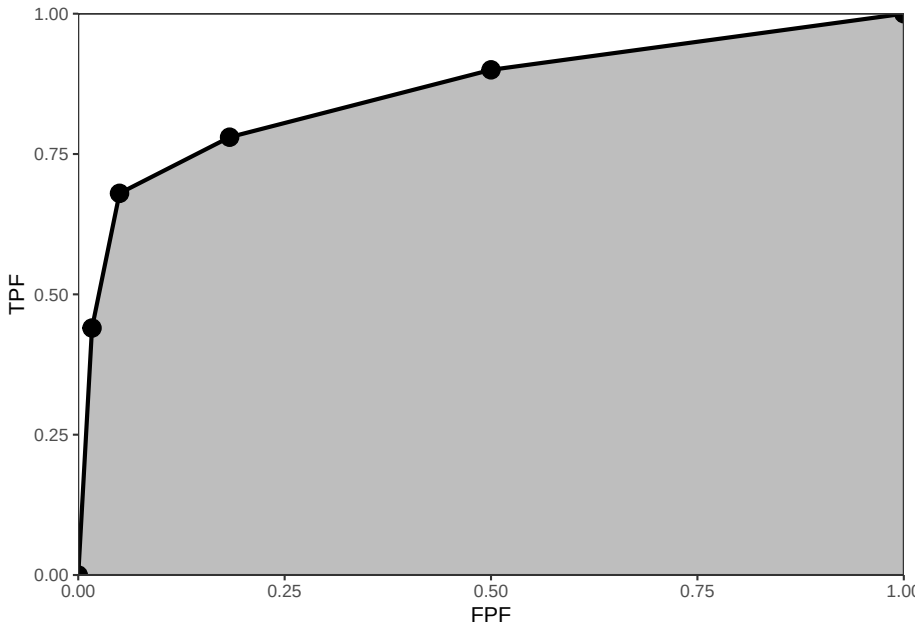


Figure 5.1: The empirical ROC plot corresponding to Table 4.1; the shaded area is the empirical AUC.

The function $\psi(x, y)$ 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 $0.5 \leq \mathrm{W} \leq 1$. 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:

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

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 $x \equiv$ FPF and $y \equiv$ TPF. One has to be careful to include the end-points as otherwise the area will be underestimated.

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


RocOperatingPoints <- function( K1, K2 ) {

  nOpPts <- length(K1) - 1 # number of op points
```

```
  FPF <- array(0,dim = nOpPts)
  TPF <- array(0,dim = nOpPts)

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

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

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

zk1  <- rep(1:length(RocCountsTable[1,]),RocCountsTable[1,])#convert frequency table to array
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. 5.2.

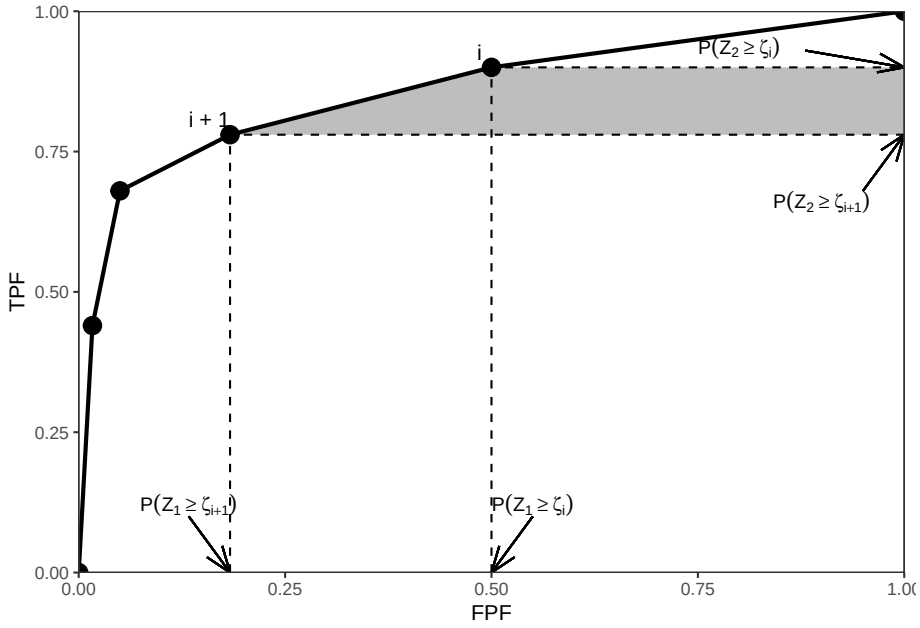


Figure 5.2: :Illustration of the derivation of Bamber's equivalence theorem. Shows an empirical ROC plot for R = 5; the shaded area is due to points labeled i and i + 1.

The 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. 5.2, 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

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

The area A_i of the shaded trapezoid in Fig. 5.2 is (the steps are shown explicitly):

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

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

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

It is shown in the Appendix that the term A_0 corresponds to the triangle at the upper right corner of Fig. 5.2, 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:

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

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 shown. This was Bamber's essential contribution. In the absence of this theorem, the Wilcoxon statistic would be "just another statistic", at least in the context of ROC analysis. The theorem is so important that a 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: 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:

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

The height is:

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

Multiplication of the last two expressions yields A_4.

5.11 Chapter References

Chapter 6

Binormal model

6.1 How much finished

97%

6.2 Introduction

The equal variance binormal model was described in Chapter 2. The ratings method of acquiring ROC data and calculation of operating points was discussed in Chapter 4. It was shown there that for a clinical dataset the unequal-variance binormal model visually fitted the data better than the equal-variance binormal model.

This chapter deals with the unequal-variance binormal model, often abbreviated to **binormal model**. It is applicable to univariate datasets in which there is *one 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 Chapter `TempComment \@ref(bivariate-binormal-model)` a bivariate model will be described where each case yields two ratings, as in a single observer interpreting cases in two modalities, or the similar problem of two observers interpreting the same cases in a single modality.

6.3 Binormal model

The binormal model is defined by (capital letters indicate random variables lower-case are realized values and t denotes the truth state):

$$\left.\begin{array}{c} Z_{k_t t} \sim N\left(\mu_t, \sigma_t^2\right) \\ t = 1, 2 \end{array}\right\} \tag{6.1}$$

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 ($t = 1$) are distributed as a $N(0, 1)$ distribution, i.e., the unit normal distribution, while the z-samples for diseased cases ($t = 2$) are distributed as a $N(\mu, \sigma^2)$ distribution, i.e., a normal distribution with mean μ and variance σ^2. 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 $\sigma > 1$.[1]

6.3.1 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 presence of disease. Define $R - 1$ ordered cutoffs ζ_i where $i = 1, 2, ..., R-1$ and $\zeta_1 < \zeta_2, ... < \zeta_{R-1}$. Also define two 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)):

$$\left.\begin{array}{r} \text{if}\left(\zeta_{r-1} \le z < \zeta_r\right) \Rightarrow \text{rating} = r \\ r = 1, 2, ..., R \end{array}\right\} \tag{6.3}$$

The above figure, generated with $\mu = 1.5$, $\sigma = 1.5$, $\zeta_1 = -2$, $\zeta_2 = -0.5$, $\zeta_3 = 1$ and $\zeta_4 = 2.5$, illustrates how realized z-samples are converted to ratings, i.e.,

[1] A more complicated version of this model would allow the mean of the non-diseased distribution to be non-zero and its variance different from unity. The resulting 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).

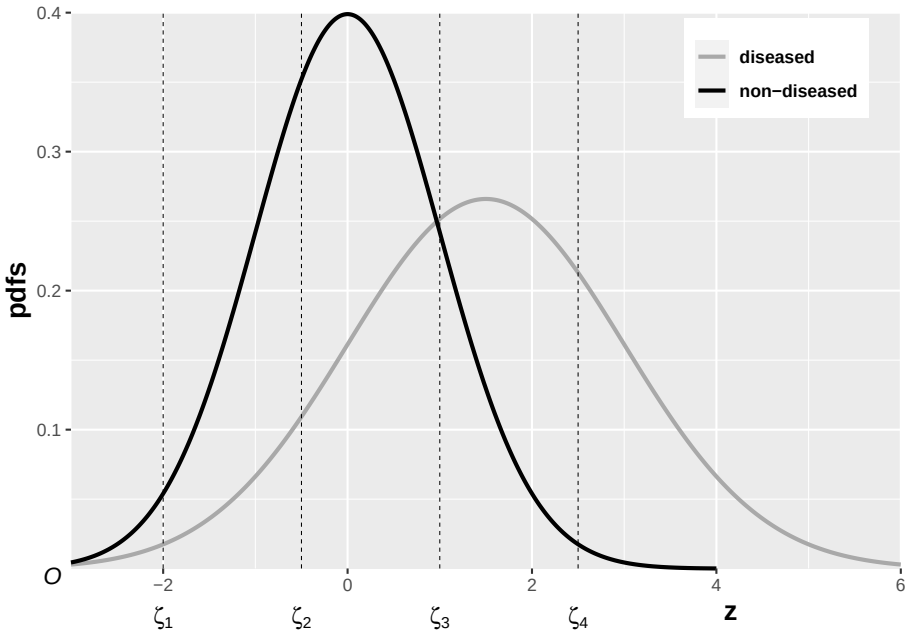


Figure 6.1: The pdfs of the two binormal model distributions for $\mu = 1.5$ and $\sigma = 1.5$. Four thresholds $\zeta_1, \zeta_2, \zeta_3, \zeta_4$ are shown corresponding to a five-rating ROC study. The rating assigned to a case is determined by its z-sample according to the binning rule.

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".

6.3.2 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 in Chapter 3 applies:

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

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, abbreviated to Se, is defined by $\mathrm{Se} \equiv P(Z_2 > \zeta)$. It follows, because σ is positive, that:

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

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

$$\left.\begin{aligned}\mathrm{Se}(\zeta|\mu, \sigma) &= 1 - P\left(\frac{Z_2 - \mu}{\sigma} \leq \frac{\zeta - \mu}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{\zeta - \mu}{\sigma}\right) = \Phi\left(\frac{\mu - \zeta}{\sigma}\right)\end{aligned}\right\} \tag{6.5}$$

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

$$\left.\begin{aligned}\mathrm{Sp}(\zeta) &= \Phi(\zeta) \\ \mathrm{Se}(\zeta|\mu, \sigma) &= \Phi\left(\frac{\mu - \zeta}{\sigma}\right)\end{aligned}\right\} \tag{6.6}$$

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

$$\left.\begin{aligned}\mathrm{FPF}(\zeta) &= 1 - \mathrm{Sp}(\zeta) \\ &= 1 - \Phi(\zeta) \\ &= \Phi(-\zeta)\end{aligned}\right\} \tag{6.7}$$

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

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):

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

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

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

This equation will be put into conventional notation next.

6.3.3 Binormal model in conventional notation

The (μ,σ) notation just described makes sense when extending the binormal model to newer models described later (see Chapter `TempComment \@ref(proper-roc-models)`). However, it was not the way the binormal model was originally parameterized. Instead the following notation is widely used in the literature:

$$\left.\begin{aligned} a &= \frac{\mu}{\sigma} \\ b &= \frac{1}{\sigma} \end{aligned}\right\} \tag{6.11}$$

> The reason for the (a,b) instead of the (μ,σ) notation is historical. (Dorfman and Alf Jr, 1969) assumed that the diseased distribution had unit variance, and the non-diseased distribution had standard deviation b and their separation was a, see Plot A in Fig. 6.2.

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, see plot B. Accordingly the inverses of Eqn. (6.11) are:

$$\left.\begin{aligned} \mu &= \frac{a}{b} \\ \sigma &= \frac{1}{b} \end{aligned}\right\} \tag{6.12}$$

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

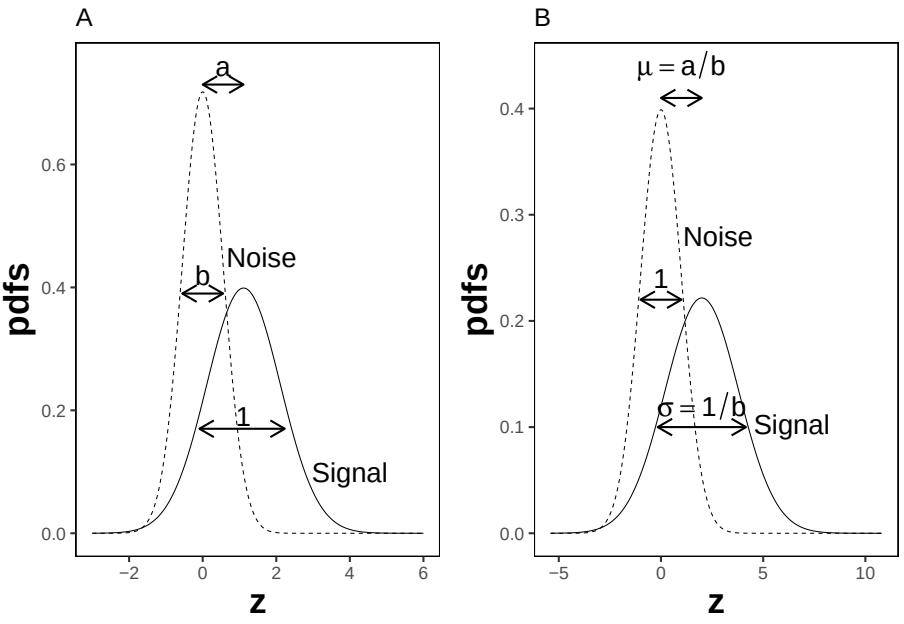


Figure 6.2: Plot A shows the definitions of the (a,b) parameters of the binormal model. In plot B the x-axis has been rescaled so that the noise distribution has unit variance; this illustrates the difference between the (a,b) and the (μ, σ) parameters. In this figure $\mu = 2$ and $\sigma = 1.8$ which correspond to $a = 1.11$ and $b = 0.556$.

6.4 ROC curve

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

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

Since $\Phi^{-1}(\text{FPF})$ 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 to (1,1).

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

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

Solve for ζ from the equation for FPF:

$$\zeta = -\Phi^{-1}\left(\text{FPF}\right) \tag{6.15}$$

6.5 Density functions

According to Eqn. (6.1) the probability that a non-diseased case z-sample is smaller than ζ, i.e., the cumulative distribution function (CDF) function for non-diseased cases, is:

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

Likewise, the CDF for diseased case z-samples is:

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

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

$$\left.\begin{aligned} pdf_N(\zeta) &= \frac{\partial \Phi(\zeta)}{\partial \zeta} \\ &= \phi(\zeta) \\ &\equiv \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\zeta^2}{2}\right) \end{aligned}\right\} \tag{6.16}$$

$$\left.\begin{aligned} pdf_D(\zeta) &= \frac{\partial \Phi\left(\frac{\zeta-\mu}{\sigma}\right)}{\partial \zeta} \\ &= \frac{1}{\sigma}\phi\left(\frac{\zeta-\mu}{\sigma}\right) \\ &\equiv \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\zeta-\mu)^2}{2\sigma}\right) \end{aligned}\right\} \tag{6.17}$$

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

$$\left.\begin{aligned} pdf_D(\zeta) &= b\phi(b\zeta - a) \\ &= \frac{b}{\sqrt{2\pi}} \exp\left(-\frac{(b\zeta-a)^2}{2}\right) \end{aligned}\right\} \tag{6.18}$$

6.6 Invariance property of pdfs

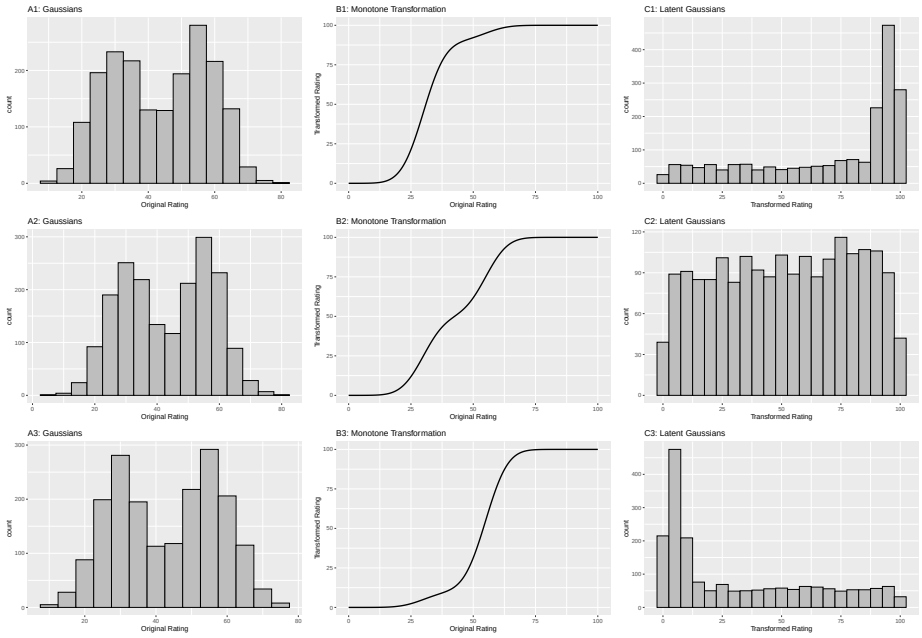
The binormal model is not as restrictive as might appear at first sight. Any monotone increasing transformation $Y = f(Z)$ applied to the observed z-samples, and the associated thresholds, will yield the same observed data, e.g., Table 4.1. This is because such a transformation leaves the ordering of 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.



This figure illustrates 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. Each of the latent Gaussian plots C1, C2 and C3 appears to be not binormal. However, using 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. 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.

- Plot A2 is for a different seed value, plot B2 is the transformation and now the transformed histogram is almost flat, plot C2. For plots A2 and C2 the corresponding AUCs are identical.

- Plot A3 is for a different seed value, B3 is the transformation 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.

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.

6.7 Az and d-prime measures

The (full) area under the ROC, denoted A_z, is derived in (Thompson and Zucchini, 1989):

$$A_z = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right) \tag{6.19}$$

The binormal fitted AUC increases as a increases or as b decreases. Equivalently, it increases as μ increases or as σ decreases.

The reason for the name A_z is that historically (prior to maximum likelihood estimation) this quantity was estimated by converting the probabilities FPF and TPF to *z-deviates* (see TBA), which of-course assumes normal distributions. The z-subscript is meant to emphasize that this is a binormal model derived estimate.

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:

$$d' = \sqrt{2}\Phi^{-1}(A_z) \tag{6.20}$$

The d' index can be regarded as a perceptual signal-to-noise-ratio.

6.8 Fitting the binormal model

(Dorfman and Alf Jr, 1969) were the first to fit ratings data to the binormal model. The details of the procedure are in Appendix 6.12. While historically very important in showing how statistically valid quantitative analysis is possible using ROC ratings data, the fitting procedure suffers from what are termed "degeneracy issues" and "fitting artifacts" discussed in Appendix 6.13. Degeneracy is when the fitting procedure yields unreasonable parameter values. Fitting artifacts occur when the fitted curve predicts worse than chance level performance in some region of the fitted ROC curve. Because of these issues usage of this method is now discouraged as it has largely been supplanted by other software such as the CBM fitting method, the proper ROC fitting method implemented in PROPROC and the RSM (radiological search model) based fitting method. These are discussed in later chapters.

6.9 Partial AUC measures

Two partial AUC measures have been defined. The idea is to have an AUC-like measure that emphasizes some region of the ROC curve, one that is argued to be clinically more significant, instead of A_z which characterizes the whole curve. In the following two definitions are considered, one that emphasizes the high specificity region of the ROC curve and one which emphasizes the high sensitivity region of the curve.

Shorthand: denote $A \equiv A_z$, $x \equiv$ FPF and $y \equiv$ TPF. The two partial AUC measures correspond to a partial integral along the x-axis starting from the origin (high specificity) and the other to a partial integral along the y-axis ending at (1,1) corresponding to high sensitivity. These are denoted by X and Y superscripts.

6.9.1 Measure emphasizing high specificity

The partial area under the ROC, A_c^X, is defined as that extending from $x = 0$ to $x = c$, where $0 \le c \le 1$ (in our notation c always means a cutoff on the x-axis of the ROC):

$$\left.\begin{aligned} A_c^X &= \int_{x=0}^{x=c} y\,dx \\ &= \int_{x=0}^{x=c} \Phi\left(a + b\,\Phi^{-1}(x)\right)\,dx \end{aligned}\right\} \tag{6.21}$$

The second form follows from Eqn. (6.13).

(Thompson and Zucchini, 1989) derive a formula for the partial-area in terms of the binormal model parameters a and b:

$$A_c^X = \int_{z_2=-\infty}^{\Phi^{-1}(c)} \int_{z_1=-\infty}^{\frac{a}{\sqrt{1+b^2}}} \phi\left(z_1, z_2; \rho\right) dz_1 dz_2 \tag{6.22}$$

On the right hand side the integrand $\phi\left(z_1, z_2; \rho\right)$ is the standard bivariate normal density function with correlation coefficient ρ. It is defined by:

$$\left.\begin{aligned} \phi\left(z_1, z_2; \rho\right) &= \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{z_1^2 - 2\rho z_1 z_2 + z_2^2}{2\left(1-\rho^2\right)}\right) \\ \rho &= -\frac{b}{\sqrt{1+b^2}} \end{aligned}\right\} \tag{6.23}$$

As demonstrated later the integrals occurring on the right hand side of Eqn. (6.22) can be evaluated numerically.

As an area measure the partial AUC A_c^X has a simple *geometric* meaning. A *physical* meaning is as follows:

> An ROC curve[2] can be defined over the truncated dataset where all z-samples **smaller** than $-\Phi^{-1}(c)$ are ignored. The maximum area of this curve is that defined by the rectangle with corners at $(0, 0)$ and $(c, \mathrm{TPF}\left(c\right))$: c is the abscissa at the upper limit of the integration interval along the x-axis and $\mathrm{TPF}\left(c\right)$ is the corresponding ordinate: see Eqn. (6.13). Dividing A_c^X by $\mathrm{TPF}\left(c\right) \times c$ yields a normalized partial area measure, denoted A_c^{XN}, where $0 \leq A_c^{XN} \leq 1$. **This is the classification accuracy between diseased and non-diseased cases measured over the truncated dataset.** If $a \geq 0$ it is constrained to (0.5, 1).

$$A_c^{XN} = \frac{A_c^X}{\mathrm{TPF}\left(c\right) \times c} \tag{6.24}$$

6.9.2 Measure emphasizing high sensitivity

Since the integral in Eqn. (6.21) is from $x = 0$ to $x = c$ this partial AUC measure emphasizes the *high specificity* region of the ROC curve (since $x = 0$ corresponds to unit, i.e. highest, specificity).

An alternative partial AUC measure has been defined (Jiang et al., 1996) that emphasizes the *high sensitivity* region of the ROC as follows:

[2] This curve is not binormal as the truncation destroys the normality of the two distributions

$$A_c^Y = \int_{y=\mathrm{TPF}(c)}^{y=1} (1-x)\, dy \tag{6.25}$$

A_c^Y is the (un-normalized) area below the ROC extending from $y = \mathrm{TPF}(c)$ to $y = 1$. The superscript Y denotes that the integral is over part of the y-axis. The maximum value of this integral is the area of the rectangle defined by the corner points $(c, \mathrm{TPF}(c))$ and $(1, 1)$. Therefore the normalized area is defined by (our normalization differs from that in the cited reference):

$$A_c^{YN} = \frac{A_c^Y}{(1-\mathrm{TPF}(c)) \times (1-c)} \tag{6.26}$$

A *physical* meaning is as follows:

> An ROC curve can be defined over the truncated dataset where all z-samples **greater** than $-\Phi^{-1}(c)$ are ignored. A_c^{YN} **is the classification accuracy between diseased and non-diseased cases measured over the truncated dataset.** By definition the normalized area ranges between 0 and 1.

6.9.3 Numerical examples

Fig. 6.3 shows the two un-normalized areas.

The following code illustrates calculation of the partial-area measure using the function `pmvnorm` in the `R` package `mvtnorm` (Genz et al., 2021). The parameter values were: $a = 1.8$, $b = 1$ and $c = 0.3$ (see lines 1-3 below).

```
1  a <- 1.8
2  b <- 1
3  fpf_c <- 0.3 # cannot use c as variable name
4  tpf_c <- pnorm(a + b * qnorm(fpf_c))
5  A_z <- pnorm(a/sqrt(1+b^2))
6  rho <- -b/sqrt(1+b^2)
7  Lower1 <- -Inf
8  Upper1 <- qnorm(fpf_c)
9  Lower2 <- -Inf
10 Upper2 <- a/sqrt(1+b^2)
11 sigma <- rbind(c(1, rho), c(rho, 1))
12 A_x <- as.numeric(pmvnorm(
13   c(Lower1, Lower2),
14   c(Upper1, Upper2),
15   sigma = sigma))
```

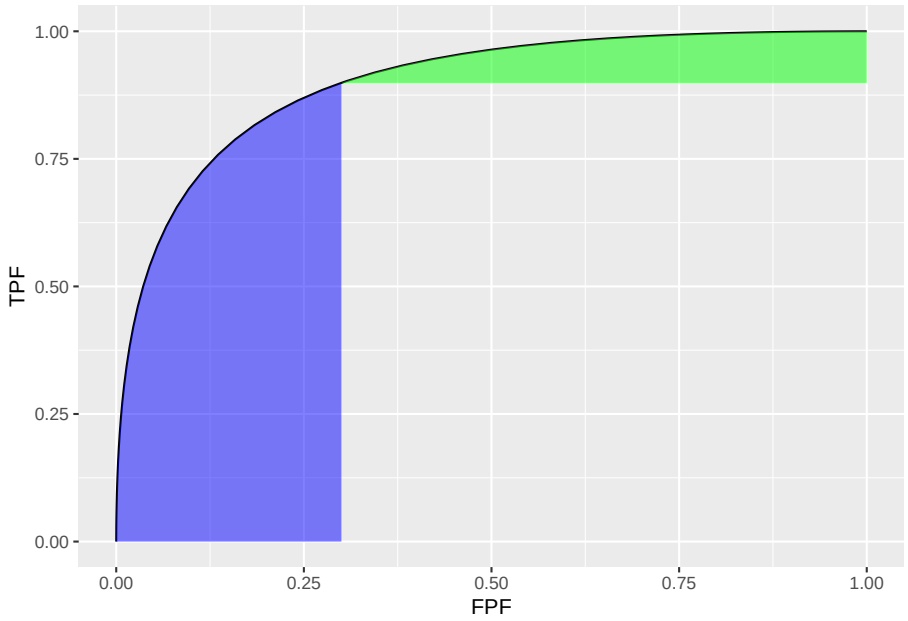


Figure 6.3: Un-normalized partial AUC measures: the blue shaded area is A_c^X, the partial area below the ROC; the green shaded area is A_c^Y the partial area above the ROC. Parameters are $a = 1.8$, $b = 1$ and $c = 0.3$.

```
16
17 # divide by area of rectangle
18 A_xn <- A_x/fpf_c/tpf_c
```

The function `pmvnorm` is called at line 12. The un-normalized partial-area measure $A_c^X = 0.216$. The corresponding full AUC measure is $A_z = 0.898$. The normalized measure is $A_c^{XN} = 0.802$. This is the classification accuracy between non-diseased and diseased cases in the truncated dataset defined by ignoring cases with z-samples smaller than $-\Phi^{-1}(c) = 0.524$. This measure emphasizes specificity.

A_c^Y can be calculated using geometry. One subtracts A_c^X from A_z to get the area under the ROC to the right of FPF $= c$. Next one subtracts from this quantity the area of the rectangle with base $(1 - c)$ and height TPF_c. This yields the area if the green shaded region A_c^Y. To normalize it one divides by the area of the rectangle defined by the corner points (c, TPF_c) and (1,1).

```
# implement geometrical logic
A_y <- (A_z - A_x)-(1-fpf_c)*(tpf_c)
A_yn <- A_y/(1-tpf_c)/(1-fpf_c)
```

The un-normalized partial-area measure $A_c^Y = 0.053$. The normalized measure is $A_c^{YN} = 0.748$. This is the classification accuracy between non-diseased and diseased cases in the truncated dataset defined by ignoring cases with z-samples greater than $-\Phi^{-1}(c) = 0.524$. This measure emphasizes sensitivity.

The variation with a of the two normalized AUC measures is shown next. The function `normalizedAreas` encapsulates the above calculations and is called for different values of a.

```
a_arr = seq(0,8)
A_xn_arr <- array(dim = length(a_arr))
A_yn_arr <- array(dim = length(a_arr))
for (i in 1:length(a_arr)) {
  x <- normalizedAreas(a_arr[i], 1, 0.1) # c = 0.1
  A_xn_arr[i] <- x$A_xn
  A_yn_arr[i] <- x$A_yn
}
```

Table 6.1 shows A_c^{XN} and A_c^{YN} partial AUCs for different values of the a parameter for $b = 1$ and $c = 0.1$. It demonstrates that the normalized areas are constrained between 0.5 and 1 (as long as a in non-negative). For numerical reasons (basically a zero-divided-by-zero condition) it is difficult to show that A_c^{YN} approaches 1 in the limit of very large a-parameter (since the green shaded area shrinks to zero).

Table 6.1: Summary of normalized A_c^{XN} and A_c^{YN} partial AUCs for different values of the a parameter, where $b = 1$ and $c = 0.1$.

a	A_c^{XN}	A_c^{YN}
0	0.5000	0.5000
1	0.6260	0.7015
2	0.7785	0.8208
3	0.9144	0.8842
4	0.9822	0.9189
5	0.9981	0.9393
6	0.9999	0.9521
7	1.0000	0.9608
8	1.0000	0.9670

6.10 Comments on partial AUC measures

There are several issues with the adoption of either partial AUC measure.

1. Since a partial area measure corresponds to classification accuracy measured over a **truncated** dataset a fundamental correspondence between A_z and classification accuracy measured over the **entire** dataset is lost. A basic statistical principle of the desirability of an estimate valid for the entire population is being violated.

2. The choice of the truncation cutoff is arbitrary and subject to bias on the part of the investigator. This is similar to the type of bias that is inherent in a single point (sensitivity-specificity) based approach to analysis: this was the very reason for adoption of a measure such as A_z that averages over the whole curve, as argued so eloquently in (Metz, 1978).

3. Then there is the issue of possible loss of statistical power. If A_z is estimated from the whole dataset and either Eqn. (6.24) or Eqn. (6.26) is used to estimate partial AUC, then one expects no loss in statistical power, as these equations represent noiseless mathematical transformations using the (a, b) parameters estimated over the entire dataset. However, if an empirical partial AUC measure is used there will surely be loss of statistical power resulting from ignoring some of the data. Due to degeneracy issues usage of the empirical partial AUC is often unavoidable. This is because performing significance testing requires that the dataset be re-sampled many times and the parametric fit may not work every time.

The second point is illustrated by the study reported in (Jiang et al., 1996). The ROC curves of a developmental-stage CAD system and that of radiologists cross each other: at high specificity the radiologists were better but the reverse was

true at high sensitivity. By choosing the latter region the authors demonstrated statistically significant superiority of CAD over radiologists. Analysis using A_z failed to reach statistical significance.

Two very large clinical studies (Fenton et al., 2007, Fenton et al. (2011)) using 222,135 and 684,956 women, respectively, showed that a commercial CAD can actually have a detrimental effect on patient outcome(Philpotts, 2009). A more recent study has confirmed the negative view of the efficacy of CAD(Lehman et al., 2015) and there has even been a call for ending Medicare reimbursement for CAD interpretations(Fenton, 2015). I have not followed the field since ca. 2016 and it is likely that newer versions of CAD now being used in the clinic are better than those evaluated in the cited studies. But the point is that even using a ca. 1996 developmental-stage CAD the authors were able to claim, using a partial AUC measure, that CAD outperformed radiologists, a result clearly not borne out by later large clinical studies while the A_z measure did not allow this conclusion.

6.11 Discussion

The binormal model is historically very important and the contribution by Dorfman and Alf (Dorfman and Alf Jr, 1969) was seminal. Prior to their work, there was no statistically valid way of estimating AUC from observed ratings counts. Their work and a key paper (Lusted, 1971) accelerated research using ROC methods. The number of publications using their algorithm, and the more modern versions developed by Metz and colleagues, is probably well in excess of 500. Because of its key role, I have endeavored to take out some of the mystery about how the binormal model parameters are estimated. In particular, a common misunderstanding that the binormal model assumptions are violated by real datasets, when in fact it is quite robust to apparent deviations from normality, is addressed (details are in Section 6.6).

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

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 1969.

6.12 Appendix: Fitting an ROC curve

One aim of this chapter is to demystify statistical curve fitting. With the passing of Profs. Donald Dorfman, Charles Metz and Richard Swensson, parametric modeling is much neglected. Researchers have instead focused on nonparametric analysis using the empirical AUC defined in Chapter 5. A claimed

advantage (overstated in my opinion, see Section 6.6) of non-parametric analysis is the absence of distributional assumptions. Non-parametric analysis yields no insight into what is limiting performance. Binormal model based curve fitting described in this chapter will allow the reader to appreciate a later chapter (see RSM fitting chapter in `RJafrocFrocBook`) that describes a more complex fitting method which yields important insights into the factors limiting human observer (or artificial intelligence algorithm) performance.

6.12.1 JAVA fitted ROC curve

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

6.12.2 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.

6.12.2.1 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):

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

This suggests that a plot of $y = \Phi^{-1}(\text{TPF})$ 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 Jr, 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):

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

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
```

```
#> Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
#> i Please use `linewidth` instead.
```

Fig. 6.4 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.

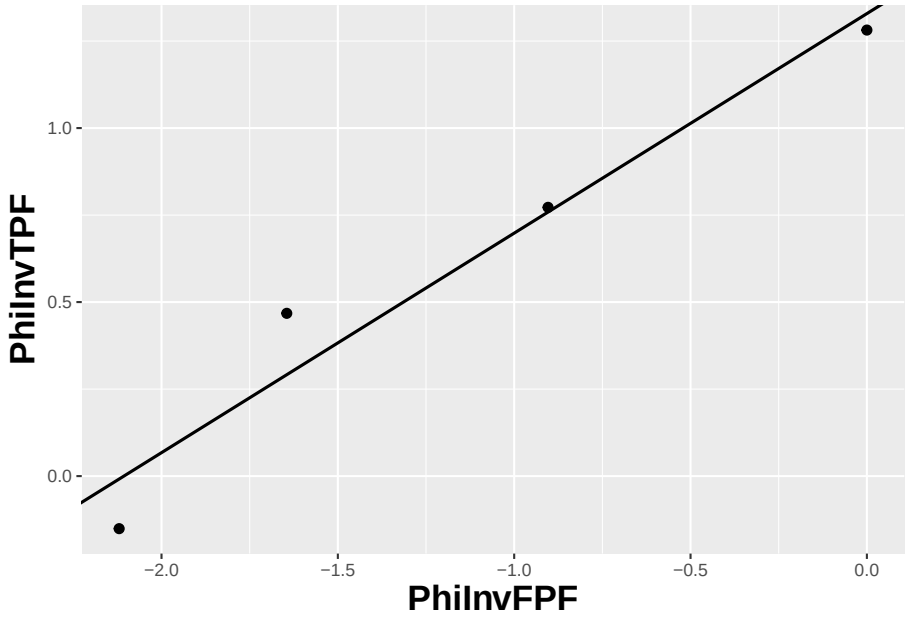


Figure 6.4: The straight line fit method of estimating parameters of the fitting model.

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, see Appendix, 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, 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.

6.12.3 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 $\Phi(\zeta_r) - \Phi(\zeta_{r-1})$; we are simply subtracting two expressions for specificity, Eqn. (6.2.5).

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

Similarly, the probability of a diseased case yielding a count in the 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 $1 - \text{TPF}$ at that threshold:

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

The area between the two thresholds is:

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

Let K_{1r} denote the number of non-diseased cases in the 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 $\Phi(\zeta_{r+1}) - \Phi(\zeta_r)$, the probability of the observed number of counts, assuming the counts are independent, is $\left(\Phi(\zeta_{r+1}) - \Phi(\zeta_r)\right)^{K_{1r}}$. Similarly, the probability of observing counts in diseased case bin r is $\left(\Phi(b\zeta_{r+1} - a) - \Phi(b\zeta_r - a)\right)^{K_{2r}}$, 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):

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

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:

$$\prod_{r=1}^{R} \left[\left(\Phi(\zeta_{r+1}) - \Phi(\zeta_r)\right)^{K_{1r}} \left(\Phi(b\zeta_{r+1} - a) - \Phi(b\zeta_r - a)\right)^{K_{2r}} \right] \tag{6.27}$$

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):

$$\frac{K_1!}{\prod_{r=1}^{R} K_{1r}!}\frac{K_2!}{\prod_{r=1}^{R} K_{2r}!} \tag{6.28}$$

The likelihood function is the product of Eqn. (6.27) and Eqn. (6.28):

$$L\left(a,b,\vec{\zeta}\right) = \left(\frac{K_1!}{\prod_{r=1}^{R} K_{1r}!}\frac{K_2!}{\prod_{r=1}^{R} K_{2r}!}\right) \times \prod_{r=1}^{R}\left[\left(\Phi\left(\zeta_{r+1}\right)-\Phi\left(\zeta_r\right)\right)^{K_{1r}}\left(\Phi\left(b\zeta_{r+1}-a\right)-\Phi\left(b\zeta_r-a\right)\right)^{K_{2r}}\right] \tag{6.29}$$

The left hand side of Eqn. (6.29) 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. (6.29) is forbidding but here comes a simplification. The difference of probabilities such as $\Phi\left(\zeta_r\right)-\Phi\left(\zeta_{r-1}\right)$ 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:

$$LL\left(a,b,\vec{\zeta}\right) = \log\left(L\left(a,b,\vec{\zeta}\right)\right) \tag{6.30}$$

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:

$$LL\left(a, b, \vec{\zeta}\right) \propto \sum_{r=1}^{R} K_{1r} \log\left(\Phi\left(\zeta_{r+1}\right) - \Phi\left(\zeta_r\right)\right) + \sum_{r=1}^{R} K_{2r} \log\left(\Phi\left(b\zeta_{r+1} - a\right) - \Phi\left(b\zeta_r - a\right)\right) \tag{6.31}$$

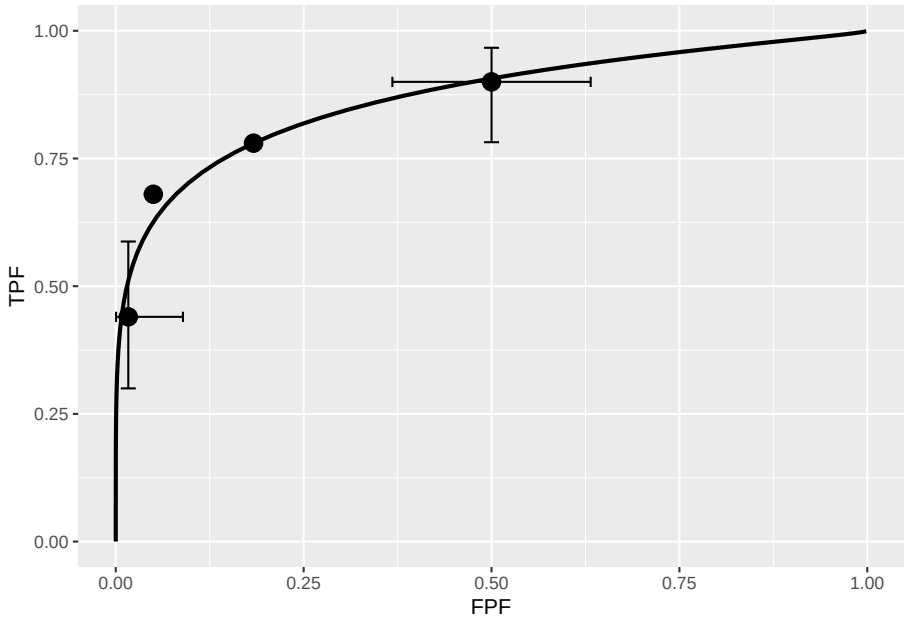
The left hand side of Eqn. (6.31) is a function of the model parameters $a, b, \vec{\zeta}$ and the observed data, the latter being the counts contained in the vectors $\overrightarrow{K_1}$ and $\overrightarrow{K_2}$, where the vector notation is used as a compact form for the counts $K_{11}, K_{12}, ..., K_{1R}$ and $K_{21}, K_{22}, ..., K_{2R}$, respectively. The right hand side of Eqn. (6.31) 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.

6.12.4 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 and Zhai, 2022). 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.

6.12.5 Validating the fit

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, 2005):

$$C^2 = \sum_{t=1}^{2} \sum_{r=1}^{R} \frac{\left(K_{tr} - \langle K_{tr} \rangle\right)^2}{\langle K_{tr} \rangle} K_{tr} \geq 5 \tag{6.32}$$

The expected values are given by:

$$\begin{aligned} \langle K_{1r} \rangle &= K_1 \left(\Phi\left(\zeta_{r+1}\right) - \Phi\left(\zeta_r\right)\right) \\ \langle K_{2r} \rangle &= K_2 \left(\Phi\left(a\zeta_{r+1} - b\right) - \Phi\left(a\zeta_r - b\right)\right) \end{aligned} \tag{6.33}$$

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:

$$C^2 \sim \chi^2_{df} \tag{6.34}$$

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

$$df = (R-1) + (R-1) - (2 + R - 1) = (R-3) \tag{6.35}$$

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).

6.12.6 Estimating the covariance matrix

TBA See book chapter 6.4.3. This is implemented in `RJafroc`.

6.12.7 Estimating the variance of Az

TBA See book chapter 6.4.4. This is implemented in `RJafroc`.

6.13 Appendix: Binormal model degeneracy and artifacts

Two helper functions are introduced here, `BMPoints` for binormal model predicted operating points and `CBMPoints` for for CBM (contaminated binormal model) operating points. The latter will become clearer in Chapter `TempComment \@ref(proper-roc-models)`. As always, to view the hidden code one needs to `fork` the repository.

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 non-lesion locations. The resulting mixture distribution will then appear to have larger variance than samples from non-diseased regions. 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 `TempComment \@ref(proper-roc-models)` – 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 `TempComment \@ref(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.

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.

6.13.1 Degenerate datasets

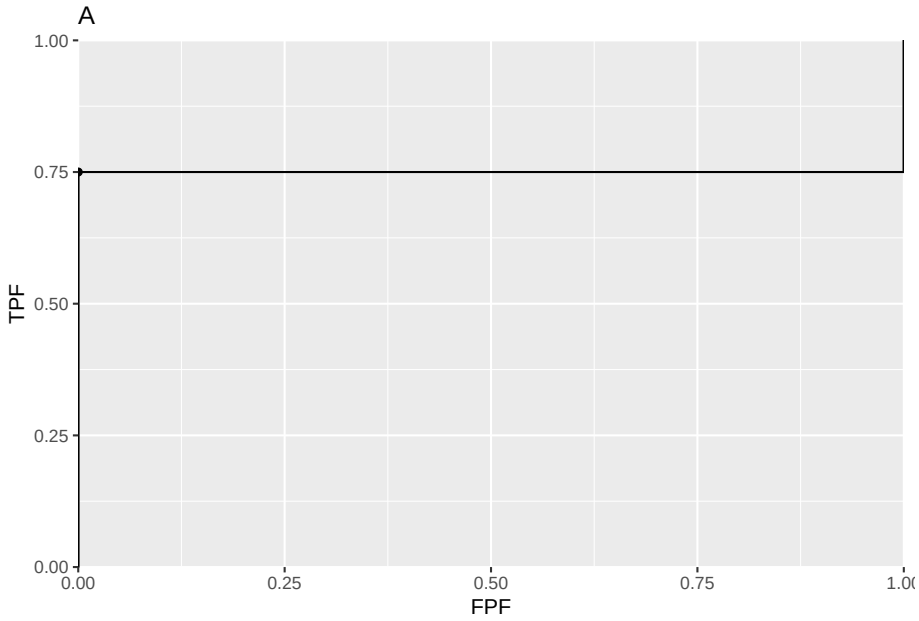
Metz defined binormal degenerate data sets as those that result in exact-fit binormal ROC curves of inappropriate shape consisting of a series of horizontal and/or vertical line segments in which the ROC "curve" crosses the chance line. The crossing of the chance line occurs because the degenerate data sets can be fitted exactly by infinite or zero values for the model slope parameter `b`, and infinite values for the decision thresholds, or both.

6.13.2 Understanding degenerate datasets

To understand this, consider that the non-diseased distribution is a Dirac delta function centered at zero (by definition such a function integrates to unity) and the unit variance diseased distribution is centered at 0.6744898. In other words this binormal model is characterized by `a = 0.6744898` and `b = 0`. What is the expected ROC curve? As the threshold ζ is moved from the far right, gradually to the left, TPF will increase but FPF is stuck at zero until the threshold reaches zero. Just before reaching this point, the coordinates of the ROC operating point are (0, 0.75). The 0.75 is due to the fact that `z = 0` is -0.6744898 units relative to the center of the diseased distribution, so the area under the diseased distribution below `z = 0` is 0.249999984. Since `pnorm` is the probability *below* the threshold, TPF must be its complement, namely 0.75. This explains the operating point (0,0.75), which lies on the y-axis. As the threshold crosses the zero-width delta function, FPF shoots up from 0 to 1, but TPF stays constant. Therefore, the operating point has jumped from (0, 0.75) to (1, 0.75). When

the threshold is reduced further, the operating point moves up vertically, along the right side of the ROC plot, until the threshold is so small that virtually all of diseased distribution exceeds it and the operating point reaches (1, 1). The ROC curve is illustrated in plot A.

```
plotOP <- data.frame(FPF = 0, TPF = 0.75)
a <- 0.6744898; b <- 0
plotCurve <- BMPoints(a, b)
figA <- ggplot(mapping = aes(x = FPF, y = TPF)) +
  geom_line(data = plotCurve) +
  geom_point(data = plotOP)  +
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0)) +
  ggtitle("A")
print(figA)
```



This is an extreme example of an ROC curve with a "hook". If the data is such that the only operating point provided by the observer is (0,0.75) then this curve will be an exact fit to the operating point.

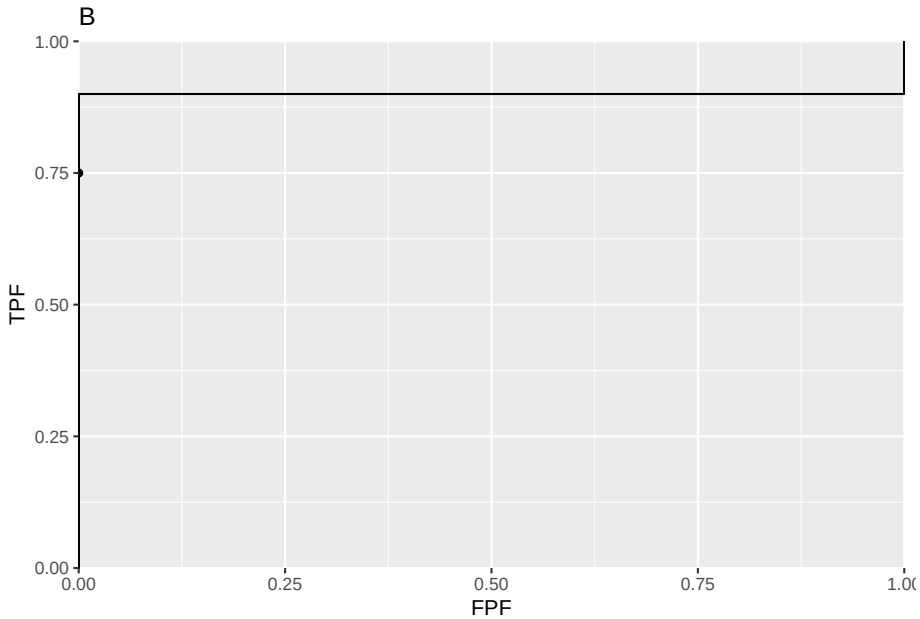
6.13.3 The exact fit is not unique

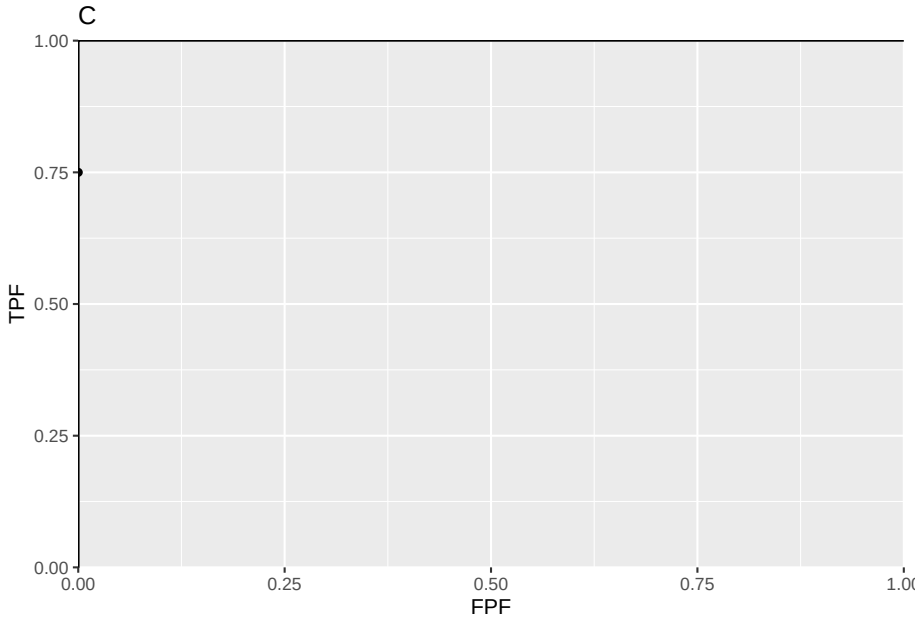
Actually, given one operating point (0, 0.75) the preceding fit is not even unique. If the diseased distribution is shifted appropriately to the right of its previous

position, and one can determine the necessary value of a, then the ROC curve will shoot upwards through the operating point (0, 0.75) to (0, 0.9), as in plot B, before proceeding horizontally to (1, 0.9) and then completing the curve to (1, 1). If the diseased distribution is shifted well to the right, i.e., a is very large, then the ROC curve will shoot upwards past the operating point, as in plot C, all the way to (0,1) before proceeding horizontally to (1, 1).

```
a <- 1.281552; b <- 0
plotCurve <- BMPoints(a, b)
figB <- ggplot(mapping = aes(x = FPF, y = TPF)) +
  geom_line(data = plotCurve) +
  geom_point(data = plotOP)  +
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0)) +
  ggtitle("B")

a <- Inf; b <- 0
plotCurve <- BMPoints(a, b)
figC <- ggplot(mapping = aes(x = FPF, y = TPF)) +
  geom_line(data = plotCurve) +
  geom_point(data = plotOP)  +
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0)) +
  ggtitle("C")
print(figB);print(figC)
```





All of these represent exact fits to the observed operating point, with `b = 0` and different values of `a`. None of them is reasonable.

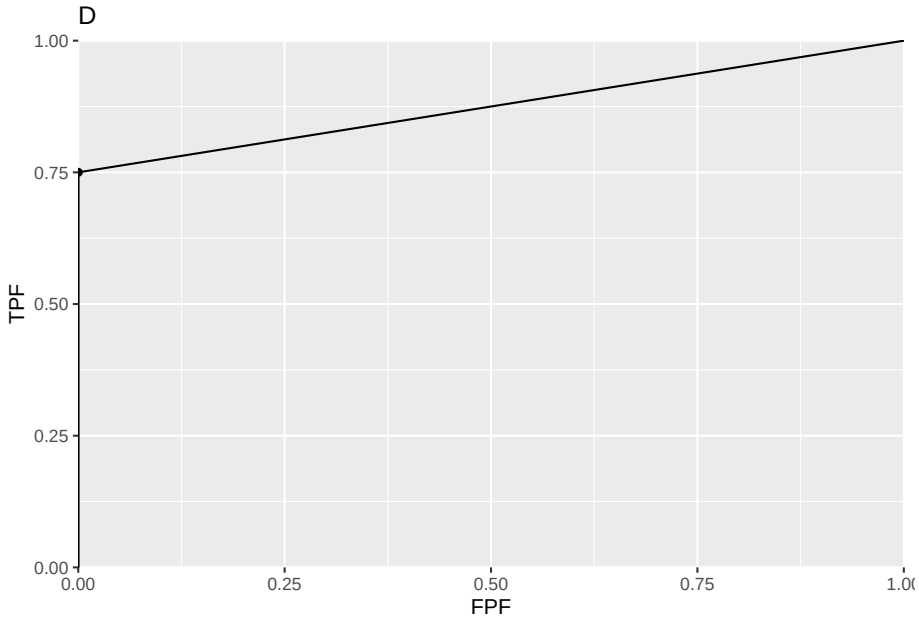
6.13.4 Comments on degeneracy

Degeneracy occurs if the observer does not provide any interior operating points. So why worry about it? Perhaps one has a non-cooperating observer, who is not heeding the instructions to *spread the ratings, use all the bins*. A simple example shows that the observer could if fact be cooperating fully and is still unable to provide any interior data points. Consider 100 diseased cases consisting of 75 easy cases and 25 difficult ones and 100 easy non-diseased cases. The observer is expected to rate the 75 easy diseased cases as *fives*, the difficult ones as *ones* and the 100 non-diseased cases are rated *ones*. No amount of coaxing *please, please spread your ratings* is going to convince this observer to rate with twos, threes and fours any of the 75 easy diseased cases. If the cases are obviously diseased, and that is what is meant by *easy cases*, they are supposed to be rated fives: *definitely diseased*. Forcing them to rate some of them as *probably diseased* or *possibly diseased* would be irrational and guilty of bending the reading paradigm to fit the convenience of the researcher (early in his research career, the author used to believe in the existence of non-cooperating observers, so Metz's advice to *spread the ratings* did not seem unreasonable at that time).

6.13.5 A reasonable fit to the degenerate dataset

If the dataset yields a single operating point (0, 0.75), what is a reasonable ROC plot? There is a theorem that given an observed operating point, the line connecting that point to (1, 1) represents a lower bound on achievable performance by the observer. The observer using a guessing mechanism to classify the remaining cases achieves the lower bound. Here is an explanation of this theorem. Having rated the 75 easy diseased cases as fives, the observer is left with 25 diseased cases and 100 non-diseased cases, all of which appear definitely non-diseased to the observer. Suppose the observer randomly rates 20% of the remaining cases as fours. This would pick up five of the actually diseased cases and 20 non-diseased ones. Therefore, the total number of diseased cases rated four or higher is 80, and the corresponding number of non-diseased cases is 20. The new operating point of the observer is (0.20, 0.80). Now, one has two operating points, the original one on the y-axis at (0, 0.75) and an interior point (0.20, 0.80). Next, instead of randomly rating 20% of the remaining cases as fours, the observer rates 40% of them as fours, then the interior point would have been (0.40, 0.85). The reader can appreciate that simply by increasing the fraction of remaining cases that are randomly rated fours, the observer can move the operating point along the straight line connecting (0, 0.75) and (1, 1), as in plot D. Since a guessing mechanism is being used, this must represent a lower bound on performance. The resulting ROC curve is proper and the net AUC = 0.875.

```
mu <- Inf; alpha <- 0.75
plotCurve <- CBMPoints(mu, alpha)
figD <- ggplot(mapping = aes(x = FPF, y = TPF)) +
  geom_line(data = plotCurve) +
  geom_point(data = plotOP)  +
  scale_x_continuous(expand = c(0, 0)) +
  scale_y_continuous(expand = c(0, 0)) +
  ggtitle("D")
print(figD)
```



For this dataset this is in fact the fit yielded by the contaminated binormal model (CBM) and the radiological search model (RSM). Why should one select the lowest possible performance consistent with the data? Because it yields a *unique* value for performance: any higher performance would not be unique.

6.14 Chapter 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):209–213.

Berbaum, K. S., Dorfman, D. D., Franken Jr, E., and Caldwell, R. T. (2002). An empirical comparison of discrete ratings and subjective probability ratings. *Academic Radiology*, 9(7):756–763.

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. and Zhai, X. (2022). *RJafroc: Artificial Intelligence Systems and Observer Performance*. R package version 2.1.2.9000.

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*, pages 837–845.

Dorfman, D. D. and Alf Jr, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating-method data. *Journal of mathematical psychology*, 6(3):487–496.

Dorfman, D. D., Berbaum, K. S., Metz, C. E., Lenth, R. V., Hanley, J. A., and Dagga, H. A. (1997). Proper receiver operating characteristic analysis: the bigamma model. *Academic Radiology*, 4(2):138–149.

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. J. (2015). Is it time to stop paying for computer-aided mammography? *JAMA Internal Medicine*, 175(11):1837–1838.

Fenton, J. J., Abraham, L., Taplin, S. H., Geller, B. M., Carney, P. A., D'Orsi, C., Elmore, J. G., Barlow, W. E., and Consortium, B. C. S. (2011). Effectiveness of computer-aided detection in community mammography practice. *Journal of the National Cancer institute*, 103(15):1152–1161.

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., et al. (2007). Influence of computer-aided detection on performance of screening mammography. *New England Journal of Medicine*, 356(14):1399–1409.

Fryback, D. G. and Thornbury, J. R. (1991). The efficacy of diagnostic imaging. *Medical decision making*, 11(2):88–94.

Genz, A., Bretz, F., Miwa, T., Mi, X., and Hothorn, T. (2021). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.1-3.

Green, D. M., Swets, J. A., et al. (1966). *Signal detection theory and psychophysics*, volume 1. Wiley 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. *Medical decision 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.

Jiang, Y., Metz, C. E., and Nishikawa, R. M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*, 201(3):745–750.

Kundel, H., Berbaum, K., Dorfman, D., Gur, D., Metz, C., and Swensson, R. (2008). Receiver operating characteristic analysis in medical imaging (icru report 79). *Report, International Commission on Radiation Units & Measurments*.

Larsen, R. J. and Marx, M. L. (2005). *An introduction to mathematical statistics*. Prentice Hall Hoboken, NJ.

Lehman, C. D., Wellman, R. D., Buist, D. S., Kerlikowske, K., Tosteson, A. N., Miglioretti, D. L., Consortium, B. C. S., et al. (2015). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine*, 175(11):1828–1837.

Liberman, L. and Menell, J. H. (2002). Breast imaging reporting and data system (bi-rads). *Radiologic Clinics*, 40(3):409–430.

Lusted, L. B. (1971). Signal detectability and medical decision making. *Science*, 171:1217–1219.

Macmillan, N. A. and Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

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.

Marcum, J. (1947). A statistical theory of target detection by pulsed radar. Technical report, RAND CORP SANTA MONICA CA.

Marcum, J. (1960). A statistical theory of target detection by pulsed radar. *IRE Transactions on Information Theory*, 6(2):59–267.

Metz, C. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.

Metz, C. E. (1986). ROC methodology in radiologic imaging. *Investigative Radiology*, 21(9):720–733.

Metz, C. E. (1989). Some practical issues of experimental design and data analysis in radiological roc studies. *Investigative radiology*, 24(3):234–245.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81.

Nishikawa, R. (2012). Estimating sensitivity and specificity in an ROC experiment. *Breast Imaging*, pages 690–696.

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. D., Gatsonis, C., Hendrick, E., Yaffe, M., Baum, J. K., Acharyya, S., Conant, E. F., Fajardo, L. L., Bassett, L., D'Orsi, C., et al. (2005). Diagnostic performance of digital versus film mammography for breast-cancer screening. *New England Journal of Medicine*, 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. E., Gur, D., and Metz, C. E. (1992). The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. *Investigative radiology*, 27(2):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.

Thompson, M. L. and Zucchini, W. (1989). On the statistical analysis of roc curves. *Statistics in medicine*, 8(10):1277–1290.

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.

Zhou, X.-H., Obuchowski, N. A., and McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine*. John Wiley and Sons, New York.