Chapter 3: Modeling the binary task

Table of contents

1. Introduction
2. Decision variable and decision threshold
3. Changing the decision threshold: Example I
4. Changing the decision threshold: Example II
5. The equal-variance binormal model
6. The normal distribution
7. Demonstration of the concepts of sensitivity and specificity
8. Inverse variation of sensitivity and specificity
9. The ROC curve
10. Assigning confidence intervals to an operating point
11. Variability in sensitivity and specificity: the Beam et al study
12. Discussion
13. References

Online Supplementary Material

A. Online Appendix 3.A: R code demonstration of sensitivity/specificity concepts

B. Online Appendix 3.B: Calculating a confidence interval

C. Online Appendix 3.C: Introduction to R/RStudio, part II

D. Online Appendix 3.D: Plotting in R

E. Online Appendix 3.E: Getting help in R – Part I

F. Online Appendix 3.F: Getting help in R – Part II

G. Online Appendix 3.G: What if a package is missing

H. Online Appendix 3.H: Shaded distributions in R

I. Online Appendix 3.I: Numerical integration in R

3.1: Introduction

Chapter 02 introduced measures of performance associated with the binary decision task. Described in this chapter is a 2-parameter statistical model for the binary task, in other words it shows how one can predict quantities like sensitivity and specificity based on the values of the parameters of a statistical model. It introduces the fundamental concepts of a *decision variable* and a *decision threshold* (the latter is one of the parameters of the statistical model) that pervade this book, and shows how the decision threshold can be altered by varying experimental conditions. The receiver-operating characteristic (ROC) plot is introduced which shows how the dependence of sensitivity and specificity on the decision threshold is exploited by a measure of performance that is independent of decision threshold, namely the area AUC under the ROC curve. AUC turns out to be related to the other parameter of the model.

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

The final aim of this chapter is to introduce **R** in somewhat greater depth so that the reader can take advantage of it in later chapters. **Online Appendix 3.C** contains the 2nd part of the **R** tutorial; the first part was in **Chapter 01**. The intent is not to make an **R**-programmer out of the reader; rather it is to get the reader to a level to appreciate its utility in demonstrating abstract formulae and concepts. Since little statistical expertise is assumed of the reader, these demonstrations take on added importance. For example, **Online Appendix 3.C** has detail on the normal distribution, how to sample from it, how to get the probability density function, the quantile function, etc. Also important is the ability to visualize data. "*A picture is worth a thousand words*": this cliché is not only true it is particularly relevant to learning this area of science. A plot summarizes a lot of information into one "visual nugget". Considerable emphasis is placed in this book on visualizing data using **R** and almost every displayed plot has a statement in the caption naming the **R** file that generated it. Finally, the Online Appendix illustrates numerical integration in **R**.

The starting point is the important concepts of *decision variable* and *decision threshold*.

3.2: Decision variable and decision threshold

The model[1] for the binary task involves three assumptions: (i) the existence of a *decision variable* associated with each case, (ii) the existence of a case-independent *decision threshold* for reporting individual cases as non-diseased or diseased and (iii) the adequacy of training session(s) in getting the observer to a steady state. In addition, common to all models is that the observer is "blinded" to the truth, while the researcher is not.

3.2.1: Existence of a decision variable

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

- By *sensory variable* one means one that is sensed internally by the observer (in the cognitive system, associated with the brain) and as such is not directly measureable in the traditional physical sense. A physical measurement, for example, might consist of measuring a voltage difference across two points with a voltmeter. The term "*latent*" is often used to describe the sensory variable because it turns out that transforming this variable by an arbitrary monotonic non-decreasing transformation has no effect on the ROC – this will become clearer later. Alternative terms are "psychophysical variable", "perceived variable", "perceptual variable" or "confidence level". The last term is the most common. It is a *subjective* variable since its value is expected to depend on the observer: the same case shown to different observers could evoke different values of the sensory variable. Since one cannot measure it anyway, it would be a very strong assumption to assume that the two sensations are identical. In this book the term "latent decision variable", or simply "*decision variable*" is used, which hopefully gets away from the semantics and focuses instead on what the variable is used for, namely *making decisions*. The symbol Z will be used for it and specific realized values are termed z-samples. It is a *random* in the sense that it varies randomly from case to case; unless the cases are similar in some respect, for example, two variants of the same case under different image processing conditions, or images of twins; in these instances the corresponding decision variables are expected to be correlated. In the binary paradigm model to be described, the decision variables corresponding to different cases are assumed mutually independent.
- The latent decision variable *rank-orders* cases with respect to evidence for *presence* of disease. Unlike a traditional rank-ordering scheme, where "1" is the highest rank, the scale is inverted with larger values corresponding to greater evidence of disease. Without loss of generality, one assumes that the decision variable ranges from -∞ to +∞, with large positive values indicative of strong evidence for presence of disease, and large negative values indicative of strong evidence for absence of disease. The zero value indicates no evidence for presence or absence of disease. [The -∞ to +∞ scale is not an assumption. The

decision variable scale could just as well range from *a* to *b*, where *a* < *b*; with appropriate rescaling of the decision variable, there will be no changes in the rank-orderings, and the scale will extend from -∞ to +∞.] Such a decision scale, with increasing values corresponding to increasing evidence of disease, is termed *positive-directed*.

3.2.2: Existence of a decision threshold

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

- The decision threshold is a *fixed value* used to separate cases reported as diseased from cases reported as non-diseased.
- Unlike the random *Z*-sample, which varies from case to case, the decision threshold is held fixed for the duration of the study. In some of the older literature[2] the decision threshold is sometimes referred to as "response bias". The author hesitates to use the term "*bias*" which has a negative connotation, whereas, in fact, the choice of decision threshold depends on rational assessment of costs and benefits of different outcomes.
- The choice of decision threshold depends on the *conditions* of the study: perceived or known disease prevalence, cost-benefit considerations, instructions regarding dataset characteristics, personal interpreting style, etc. There is a transient "learning curve" during which observer is assumed to find the optimal threshold and henceforth holds it constant for the duration of the study. The learning is expected to stabilize during a sufficiently long training interval.
- Data should only be collected in the fixed threshold state, i.e., at the end of the training session.
- If a second study is conducted under different conditions, the observer will determine, after a new training session, the optimal threshold for the new conditions and henceforth hold it constant for the duration of the second study, etc.

From assumption #2, it follows that:

$$1 - Sp = FPF = P(Z \geq \zeta \mid T = 1) \quad . \qquad \textbf{(3.1)}$$

$$Se = TPF = P(Z \geq \zeta \mid T = 2) \quad . \qquad \textbf{(3.2)}$$

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

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

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

3.2.3: Adequacy of the training session

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

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

3.3: Changing the decision threshold: Example I

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

Next, suppose the study is repeated with the same cases but this time the payoff is $1 for each TP decision and $10 for each TN decision. Suppose, further, that sufficient time has elapsed between the two study sessions that memory effects can be neglected. Now the roles of sensitivity and specificity are reversed. The radiologist's incentive is to be correct on actually non-diseased cases without worrying too much about missing actually diseased cases. The radiologist is expected to set the decision threshold at a large value so that considerable evidence of disease-presence is required to result in a "case is diseased" decision, but even slight evidence of absence of disease is enough to result in a "case is non-diseased" decision. This radiologist is expected to achieve relatively low sensitivity but specificity will be higher. Assume the radiologist makes 90 TN decisions and 10 TP decisions, earning $910 for the second study. The corresponding sensitivity is 10% and specificity is 90%. The numbers in this example are summarized in Table 3.1.

Table 3.1: This table illustrates the dependence of the number of counts in a 2x2 table on the payoffs. Reversal of the payoff scheme causes the observer to reverse the roles of sensitivity and specificity to achieve the same payoff.

Decision	TP earns $10, TN earns $1		TP earns $1, TN earns $10	
	T = 0	T = 1	T = 0	T = 1

D = 0	#TN = 10	#FN = 10	#TN = 90	#FN = 90
D = 1	#FP = 90	#TP = 90	#FP = 10	#TP = 10
Se, Sp, Payoff	Se=0.9, Sp = 0.1, Payoff = $910		Se=0.1, Sp = 0.9, Payoff = $910	

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.

3.4: Changing the decision threshold: Example II

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

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

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

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

3.5: The equal-variance binormal model

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

1. The Z-samples for non-diseased cases are distributed $N\left(0,1\right)$.
2. The Z-samples for diseased cases are distributed $N\left(\mu,1\right);\mu \geq 0$.
3. A case is diagnosed as diseased if its Z-sample ≥ a constant threshold ζ , 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[a].

The model described above is termed the *equal-variance binormal model*. [If the common variance is not unity, one can rescale the decision axis to achieve unit-variance without changing the predictions of the model.] A more general model termed the *unequal-variance binormal model* is generally used for modeling human observer data, **Chapter 06**, but for the moment, one does not need that complication. The equal-variance binormal model is defined by:

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

In Eqn. (3.3) the subscript *t* denotes the truth, sometimes referred to as the "gold standard", with *t* = 1 denoting a non-diseased case and *t* = 2 denoting a diseased case. The variable $Z_{k_t t}$ denotes the random Z-sample for case

[a] In his teaching experience, this example invariably elicits laughter from the audience. It also reminds the author, in the current (Aug. 2016) political context, of a particular prognosticator (Bill Kristol) whose political predictions are so bad that he is considered "good", but not the way he would like it. Figuratively, he is a Kristol-Ball; see for example https://www.youtube.com/watch?v=UmmGHueOpEs.

$k_t t$, where k_t is the index for cases with truth state *t*; for example $k_1 1 = 21$ denotes the 21st non-diseased case and $k_2 2 = 3$ denotes the 3rd diseased case. To explicate $k_1 1 = 21$ further, the label k_1 indexes the case while the label 1 indicates the truth of the case. The label k_t ranges from $1,2,...,K_t$, where K_t is the total number of cases with disease state *t*.

The author departs from usual convention, which labels the cases with a single index k_t, which ranges from 1 to $\left(K_1 + K_2\right)$, and one is left guessing as to the truth-state of each case. Also, the proposed notation extends more readily to the FROC paradigm where two states of truth have to be distinguished, one at the case level and one at the location level, **Chapter 13**.

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

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

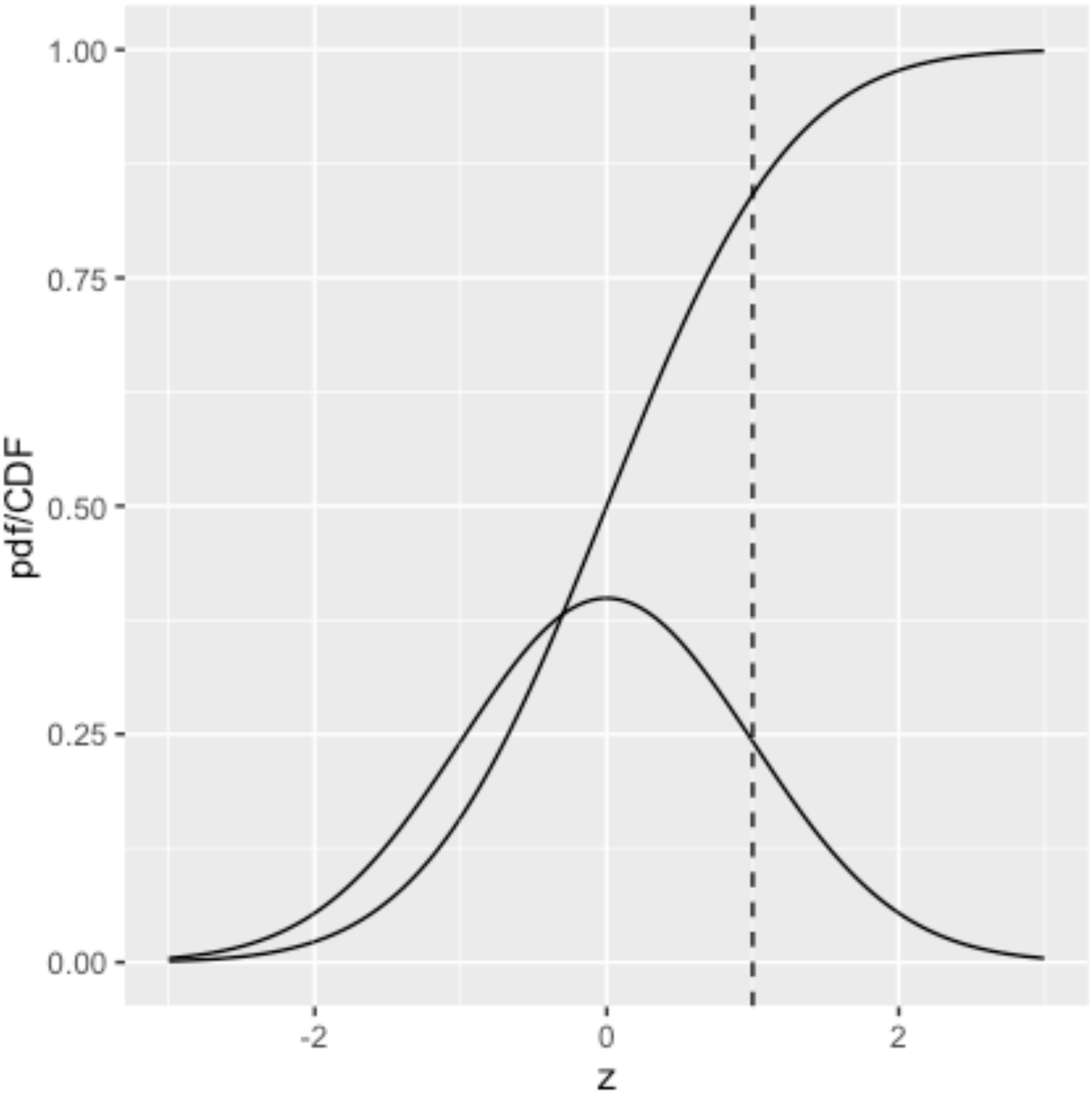

Fig. 3.1: The sigmoid shaped curve is the *CDF*, or cumulative distribution function, of the *N(0,1)* distribution, while the bell-shaped curve is the corresponding *pdf*, or probability density function. The dashed line corresponds to the reporting threshold $\zeta = 1$. The area under the *pdf* to the left of ζ equals the value of CDF at the selected ζ, i.e., 0.841 (**`pnorm(1)`** = 0.841). The code for this figure is in **`mainUnitNormalPdfCdf.R`**.

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

$$\phi\left(z \mid \mu,\sigma\right)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(z-\mu)^2}{2\sigma^2}} \qquad . \qquad \mathbf{(3.4)}$$

By definition,

$$\phi\left(z \mid \mu,\sigma\right)dz = P\left(z < Z \le z + dz \middle| Z \sim N\left(\mu,\sigma^2\right)\right) \qquad . \qquad \mathbf{(3.5)}$$

The right hand side of Eqn. (3.5) is the probability that the random variable *Z*, sampled from $N\left(\mu,\sigma^2\right)$, is between the fixed limits *z* and *z* + *dz*. For this reason $\phi\left(z \mid \mu,\sigma\right)$ is termed the probability *density* function. The special case $N\left(0,1\right)$ is referred to as the *unit normal distribution*; it has zero mean and unit variance and the corresponding notation is $\phi\left(z\right)$. The defining equation for the *pdf* of this distribution is:

$$\phi\left(z\right)=\frac{1}{\sqrt{2\pi}}e^{\frac{-z^2}{2}} \qquad . \qquad \mathbf{(3.6)}$$

The integral of $\phi\left(z\right)$ from negative infinity to z is the probability that a sample from the unit normal distribution is less than z. Regarded as a function of z, this is termed the *cumulative distribution function* (*CDF*) and is denoted, in this book, by the symbol Φ. The function $\Phi\left(z\right)$, specific to the unit normal distribution, is defined by:

$$\Phi\left(z\right) \equiv P\left(Z \le z \middle| Z \sim N\left(0,1\right)\right) = \int_{-\infty}^{z}\phi\left(t\right)dt \qquad . \qquad \mathbf{(3.7)}$$

Fig. 3.1 shows plots, as functions of *z,* of the *CDF* and the *pdf* for the unit normal distribution. Since *z*-samples outside ±3 are unlikely, the plotted range, from -3 to +3 includes most of the distribution. The *pdf* is the familiar bell-shaped curve, centered at zero; the corresponding **R** function is **dnorm()**, i.e., density of the normal distribution. The CDF $\Phi\left(z\right)$ increases monotonically from 0 to unity as *z* increases from -∞ to +∞. It is the sigmoid-shaped curve in Fig. 3.1; the corresponding **R** function is **pnorm()**. A related function is the inverse of Eqn. (3.7). Suppose the left hand side of Eqn. (3.7) is denoted *p*, which is a probability in the range 0 to 1, i.e.,

$$p \equiv \Phi(z) = \int_{-\infty}^{z} \phi(t)\,dt \quad . \tag{3.8}$$

The inverse of $\Phi(z)$ is that function which when applied to *p* yields the upper limit z in Eqn. (3.8), i.e.,

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

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

$$\Phi^{-1}(\Phi(z)) = z \quad . \tag{3.10}$$

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

To summarize, **norm** implies the unit normal distribution, **p** denotes a probability distribution function[b] or *CDF*, **q** denotes a quantile function and **d** denotes a density function; this convention is used with all distributions in **R**; there is a method to the madness.

Open the **software.prj** file corresponding to this chapter. The following **Console** window code snippet shows their usage (the reader should type the first line into the **RStudio** window to confirm):

3.6.1: Code Snippet

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

[b] In the statistical literature this is also referred to as the probability distribution function, which unfortunately has the same abbreviation as the probability density function, excepting for a change is case; for clarity, the author will always refer to it as the cumulative distribution function (*CDF*).

```
[1] -1.96
> options(digits = 3)
> qnorm(0.025);qnorm(1-0.025);pnorm(qnorm(0.025));qnorm(pnorm(-1.96))
[1] -1.96
[1] 1.96
[1] 0.025
[1] -1.96
```

Multiple **R** commands can be placed on a line as long as they are separated by semi-colons. The first command **`qnorm(0.025)`** demonstrates the identity:

$$\Phi^{-1}\left(0.025\right) = -1.959964 \quad . \tag{3.11}$$

Use **`options(digits = 3)`** to display fewer digits. The next command **`qnorm(1-0.025)`** demonstrates the identity:

$$\Phi^{-1}\left(1-0.025\right) = 1.959964 \quad . \tag{3.12}$$

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 0.025. Similarly, Eqn. (3.12) mans 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% of the unit-normal *pdf,* these two values can be used to estimate 95% 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% confidence interval, and the observed value plus 1.96 defines the upper limit of its 95% confidence interval.

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

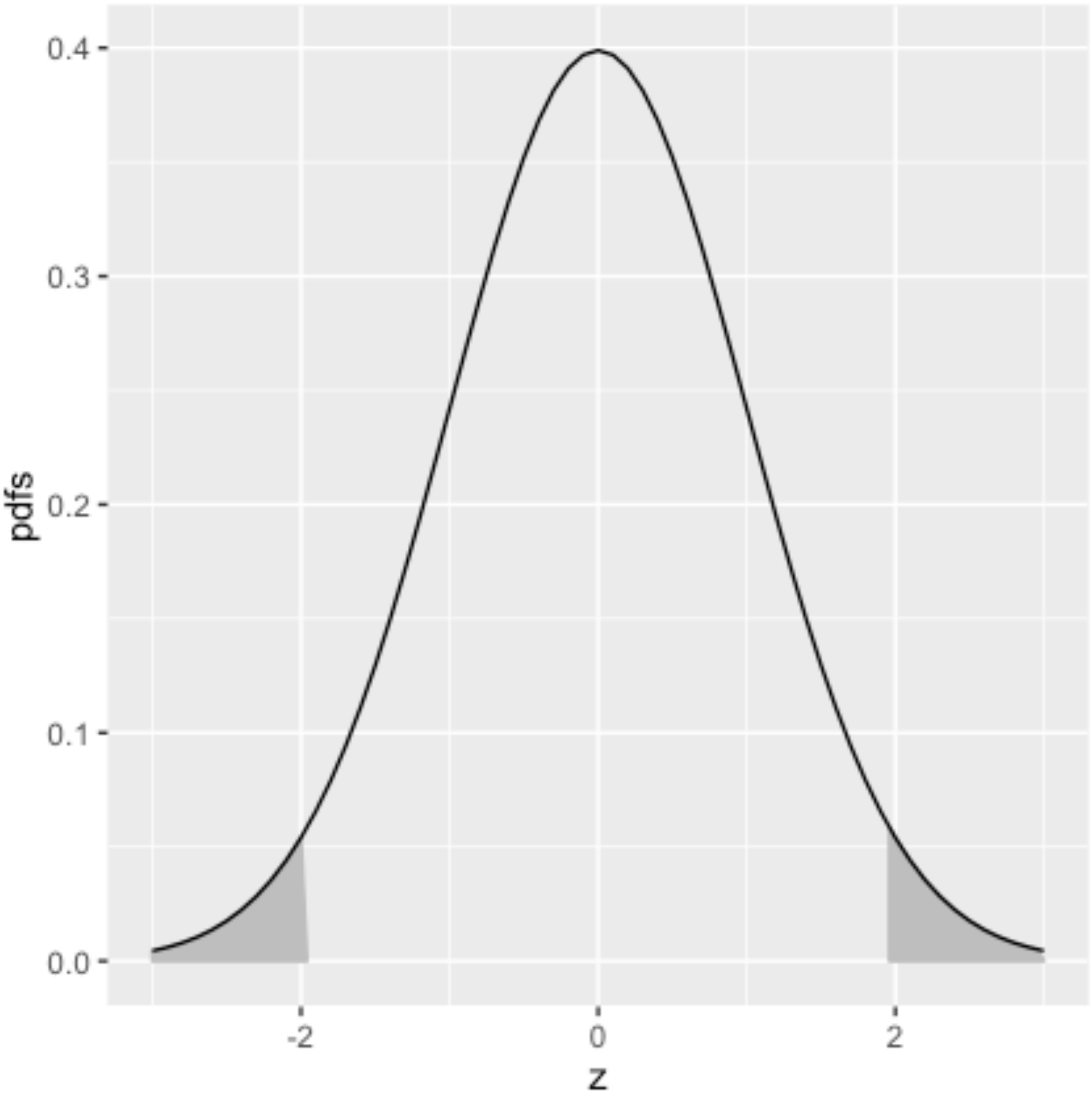

Fig. 3.2: This plot illustrates the fact that 95% of the total area under the unit normal pdf is contained in the range $|Z| < 1.96$, which can be used to construct a 95% confidence interval for an estimate of a suitably normalized statistic. The area contained in each shaded tail is 2.5%. This figure was constructed using the code in **`mainShadedTails.R`**.

3.6.2: Analytic expressions for specificity and sensitivity

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

$$Sp(\zeta) = P\left(Z_{k_1 1} < \zeta \middle| Z_{k_1 1} \sim N(0,1)\right) = \Phi(\zeta) \quad . \qquad \textbf{(3.13)}$$

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

$$\left.\begin{aligned} Se(\zeta) &= P\left(Z_{k_2 2} \geq \zeta\right) = P\left(\left(Z_{k_2 2} - \mu\right) \geq (\zeta - \mu)\right) \\ &= 1 - P\left(\left(Z_{k_2 2} - \mu\right) < (\zeta - \mu)\right) = 1 - \Phi(\zeta - \mu) \end{aligned}\right\} . \tag{3.14}$$

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

$$1 - \Phi(\zeta) = \Phi(-\zeta) . \tag{3.15}$$

$$1 - \Phi(\zeta - \mu) = \Phi(\mu - \zeta) . \tag{3.16}$$

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:

$$Sp(\zeta) = \Phi(\zeta) . \tag{3.17}$$

$$Se(\zeta) = \Phi(\mu - \zeta) . \tag{3.18}$$

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

As probabilities, both sensitivity and specificity are restricted to the range 0 to 1. The observer's performance could be characterized by specifying sensitivity *and* specificity, i.e., a *pair* of numbers. If both sensitivity and specificity of an imaging system are greater than the corresponding values for another system, then the 1st system is unambiguously better than the 2nd. But what if sensitivity is greater for the 1st but specificity is greater for the 2nd? Now the comparison is ambiguous. It is difficult to unambiguously compare two pairs of performance indices. Clearly, a scalar measure is desirable that combines sensitivity and specificity into a single measure of diagnostic performance.

The parameter μ satisfies the requirements of a scalar figure of merit (FOM). Eqn. (3.17) and Eqn. (3.18) can be solved for μ as follows. Inverting the equations yields:

$$\zeta = \Phi^{-1}\left(Sp\right) \quad . \tag{3.19}$$

$$\mu - \zeta = \Phi^{-1}\left(Se\right) \quad . \tag{3.20}$$

Eliminating ζ yields:

$$\mu = \Phi^{-1}\left(Sp\right) + \Phi^{-1}\left(Se\right) \quad . \tag{3.21}$$

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

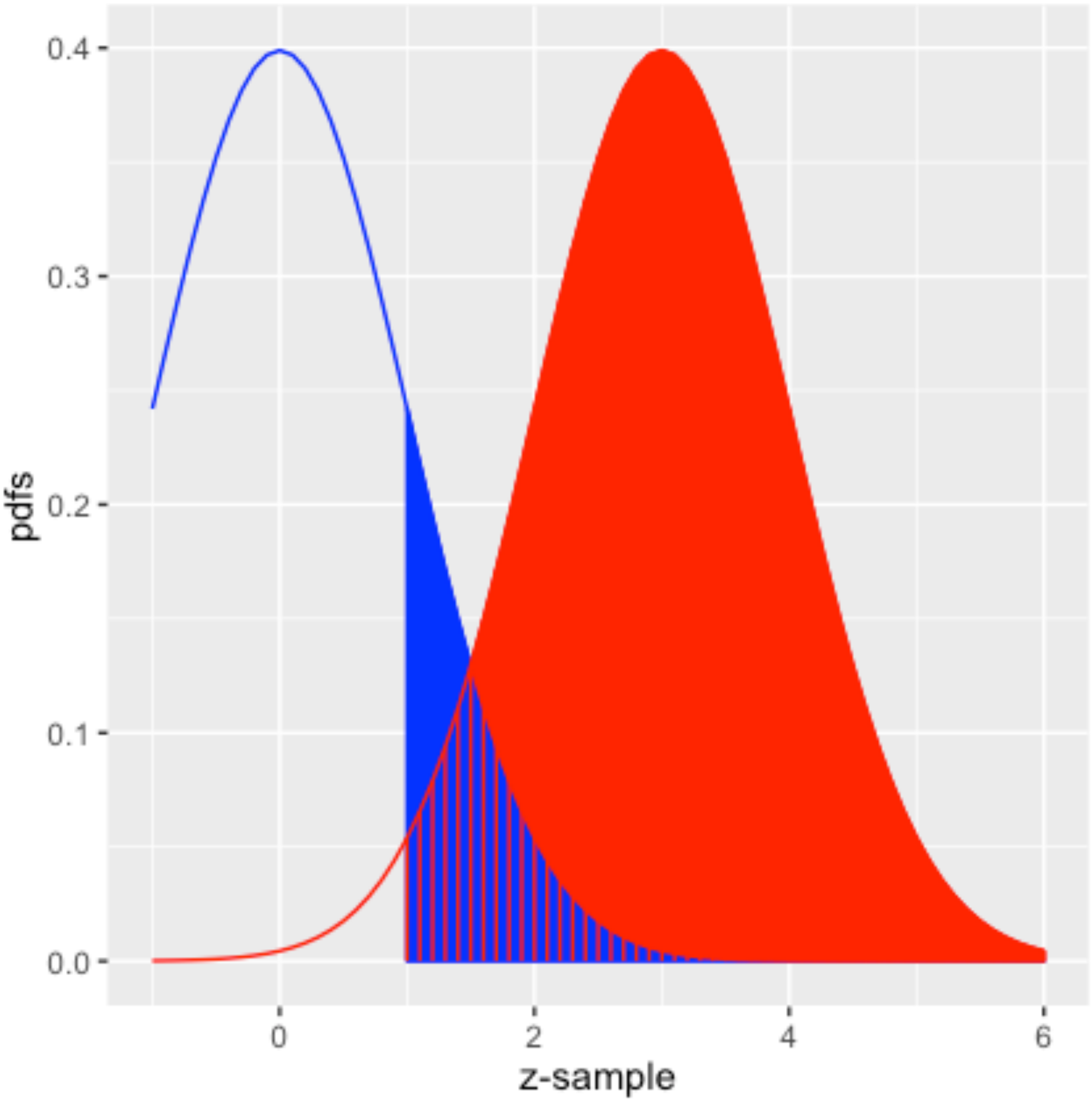

Fig. 3.3: The equal-variance binormal model for $\mu = 3$ and $\zeta = 1$; the blue curve, centered at zero, corresponds to the *pdf* of non-diseased cases and the red one, centered at $\mu = 3$, corresponds to the *pdf* of diseased cases. The left edge of the blue shaded region represents the threshold ζ, currently set at unity. The red shaded area, including the "common" portion with the vertical red lines, is sensitivity. The blue shaded area including the "common" portion with the vertical red lines is 1-specificity. The colors may not reproduce well in the book, but they can be viewed in the online material. This figure was generated using **mainShadedPlots.R**.

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* ζ, which is the *complement* of specificity, i.e., it is false positive fraction, which is 1 - 0.841 = 0.159. The 0.841 comes from **pnorm(1)** = 0.841. 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* ζ, which is sensitivity or true positive fraction, which is $\Phi(3-1)=\Phi(2)=0.977$, because **`pnorm(2)`** = 0.977.

Demonstrated next are these concepts using **`R`** examples.

3.7: Demonstration of the concepts of sensitivity and specificity

The code for this demonstration is **`mainBinaryRatings.R`** and an explanation is in Online Appendix 3.A. Ensure that line 4 reads **`seed <- 100;K1 <- 9;K2 <- 11`** and **`Source`** the file. The following output appears in the **`Console`** window.

3.7.1: Code Output

```
> source('~/book2/02 A ROC analysis/A3 Modeling Binary Paradigm/software/mainBinaryRatings.R')
seed =  100 , K1 =  9 , K2 =  11 Specificity =  0.889 , Sensitivity =  0.909 , Est. mu =  2.56
```

The estimate of μ (2.56) was calculated using Eqn. (3.21), i.e., **`mu <- qnorm(Sp) + qnorm(Se)`**.

3.7.2: Changing the seed variable: case-sampling variability

No matter how many times one clicks the **`Source`** button one always sees the same output shown in §3.7.1. This is because at line 4 one sets the **`seed`** of the random number generator to a fixed value, namely 100. This is like having a perfectly reproducible reader repeatedly interpret the same cases – one always gets the same results. Change the **`seed`** to 101 and click on the **`Source`** button. One should see:

3.7.2.1: Code Output

```
> source('~/book2/02 A ROC analysis/A3 Modeling Binary Paradigm/software/mainBinaryRatings.R')
seed =  101 , K1 =  9 , K2 =  11 Specificity =  0.778 , Sensitivity =  0.545 , Est. mu =  0.879
```

Changing the seed is equivalent to sampling a completely new set of patients. *This is an example of case sampling variability*. The effect is quite large (**`Se`** fell from 0.909 to 0.545 and estimated mu fell from 2.56 to 0.88!) because the size of the relevant case set, 11 for sensitivity, is small, leading to large variability.

3.7.3: Increasing the numbers of cases

Increase K_1 and K_2, by a factor of 10 each, and return **`seed`** to 100. Clicking on the **`Source`** button yields:

3.7.3.1: Code Output

```
> source('~/book2/02 A ROC analysis/A3 Modeling Binary Paradigm/software/mainBinaryRatings.R')
seed =  100 , K1 =  90 , K2 =  110 Specificity =  0.778 , Sensitivity =  0.836 , Est. mu =  1.74
```

Change the **seed** variable to 101 and click on **Source**:

3.7.3.2: Code Output

```
> source('~/book2/02 A ROC analysis/A3 Modeling Binary Paradigm/software/mainBinaryRatings.R')
seed =  101 , K1 =  90 , K2 =  110 Specificity =  0.811 , Sensitivity =  0.755 , Est. mu =  1.57
```

Notice that now the values are less sensitive to **seed**. Table 3.2 illustrates this trend with ever increasing sample sizes (the reader should confirm the listed values).

Table 3.2: Effect of sample size on case-sampling variability of estimates of sensitivity, specificity and the separation parameter; $\widehat{Se}$ = estimate of sensitivity, $\widehat{Sp}$ = estimate of specificity, $\hat{\mu}$ = estimate of separation parameter; K_1 and K_2 are the numbers of non-diseased and diseased cases, respectively. Different values of **seed** generate different case samples. The parameters of the model are $\mu = 1.5$ and $\zeta = \mu / 2$. The values for infinite numbers of cases are from the analytical expressions Eqn. (3.17) and Eqn. (3.18).

K_1	K_2	**seed**	$\left(\widehat{Se}, \widehat{Sp}\right)$	$\hat{\mu}$
9	11	100	(0.889, 0.909)	2.56
		101	(0.778, 0.545)	0.879
90	110	100	(0.778, 0.836)	1.74
		101	(0.811, 0.755)	1.57
900	1100	100	(0.764, 0.761)	1.43
		101	(0.807, 0.759)	1.57
9000	11000	100	(0.774, 0.772)	1.5
		101	(0.771, 0.775)	1.5
∞	∞	NA	(0.773, 0.773)	1.5

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. If one types **pnorm(0.75)** in the **Console** window one sees:

3.7.3.3: Code Snippet

```
> pnorm(0.75) # example 1
[1] 0.7733726
> 2*qnorm(pnorm(zeta)) # example 2
[1] 1.5
```

Because the threshold is halfway between the two distributions, 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. (3.21): 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 four numbers are involved. Comparing $\hat{\mu}$ values is easier, as only two numbers are involved. The tendency to become independent of case sample is discernible with fewer cases with $\hat{\mu}$, around 90/110 cases, than with sensitivity and specificity pairs. The numbers in the table might appear disheartening in terms of the implied numbers of cases needed to detect a difference in specificity. Even with 200 cases, the difference in specificity for two seed values is 0.081, which is actually a large effect considering that the scale extends from 0 to 1.0. A similar comment applies to differences in sensitivity. The situation is not quite that bad. One uses an area measure that combines sensitivity and specificity yielding less variability. One uses the ratings paradigm, which is more efficient than the binary one in this chapter. Finally, one takes advantage of correlations that exist between the interpretations in matched-case matched-reader interpretations in two modalities that tend to decrease variability in the AUC-difference even further (most applications of ROC methods involved detecting *differences* in AUCs not absolute values).

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

The inverse variation of sensitivity and specificity is modeled in the binormal model by the threshold parameter ζ. From Eqn. (3.17), specificity at threshold ζ is $\Phi(\zeta)$ and the corresponding expression for sensitivity is $\Phi(\mu-\zeta)$. Since the threshold ζ appears with a *minus* sign, the dependence of sensitivity on ζ will be the *opposite* of the corresponding dependence of specificity. In Fig. 3.3, the left edge of the blue shaded region represents the threshold ζ, currently set at unity. As ζ is moved towards the left specificity decreases but sensitivity increases. Specificity decreases because less of the non-diseased distribution lies to the left of the new threshold, in other words fewer non-diseased cases are correctly diagnosed as non-diseased. Sensitivity increases because more of the diseased distribution lies to the right of the new threshold, in other words more diseased cases are correctly diagnosed as diseased. If an observer has higher sensitivity than another observer, but lower specificity, it is difficult to unambiguously compare them. It is not impossible[3,4]. The unambiguous comparison is difficult for the following reason. Assuming the second observer can be coaxed into adopting a lower threshold, thereby decreasing specificity to match that of the first observer, and then it is possible that the second observer's sensitivity, formerly smaller, could now be greater than that of the first observer. A single figure of merit is desirable to the sensitivity - specificity analysis. It is possible to leverage the inverse variation of sensitivity and specificity by combing them into a single scalar measure, as was done with the μ parameter in the previous section, Eqn. (3.21). An equivalent way is by using the area under the ROC plot, discussed next.

3.9: 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. (3.14), Eqn. (3.17) and Eqn. (3.18) it follows that

$$FPF\left(\zeta\right)=1-Sp\left(\zeta\right)=1-\Phi\left(\zeta\right)=\Phi\left(-\zeta\right) \quad . \qquad \textbf{(3.22)}$$

$$TPF\left(\zeta\right)=Se\left(\zeta\right)=\Phi\left(\mu-\zeta\right) \quad . \qquad \textbf{(3.23)}$$

Specifying ζ selects a particular *operating point* on this plot and varying ζ from +∞ to -∞ causes the operating point to trace out the ROC *curve from the origin to (1,1)*. Specifically, as ζ is *decreased* from +∞ to -∞, the operating point *rises* from the *origin* (0,0) to *the end-point* (1,1). In general, as ζ *increases* the operating point moves *down* the curve, and conversely, as ζ decreases the operating point moves up the curve. The operating point $O\left(\zeta \mid \mu\right)$ for the equal variance binormal model is (the notation assumes the μ parameter is fixed and ζ is varied by the observer in response to interpretation conditions):

$$O\left(\zeta \mid \mu\right)=\left(\Phi\left(-\zeta\right),\Phi\left(\mu-\zeta\right)\right) \quad . \qquad \textbf{(3.24)}$$

The operating point predicted by the above equation lies *exactly* on the theoretical ROC curve. This condition can only be achieved with very large numbers of cases, so that sampling variability is very small. In practice, with finite datasets, the operating point will almost never be exactly on the theoretical curve. The author leaves it as an exercise for the reader to come up with an exception.

The ROC curve is the locus of the operating point for fixed μ *and variable* ζ. Fig. 3.4 shows examples of equal-variance binormal model ROC curves for different values of μ. Each curve is labeled with the corresponding value of μ . Each has the property that TPF is a monotonically increasing function of FPF and the slope decreases monotonically as the operating point moves up the curve. As μ increases the curves get progressively *upward-left shifted*, approaching the top-left corner of the ROC plot. In the limit $\mu \to \infty$ 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.

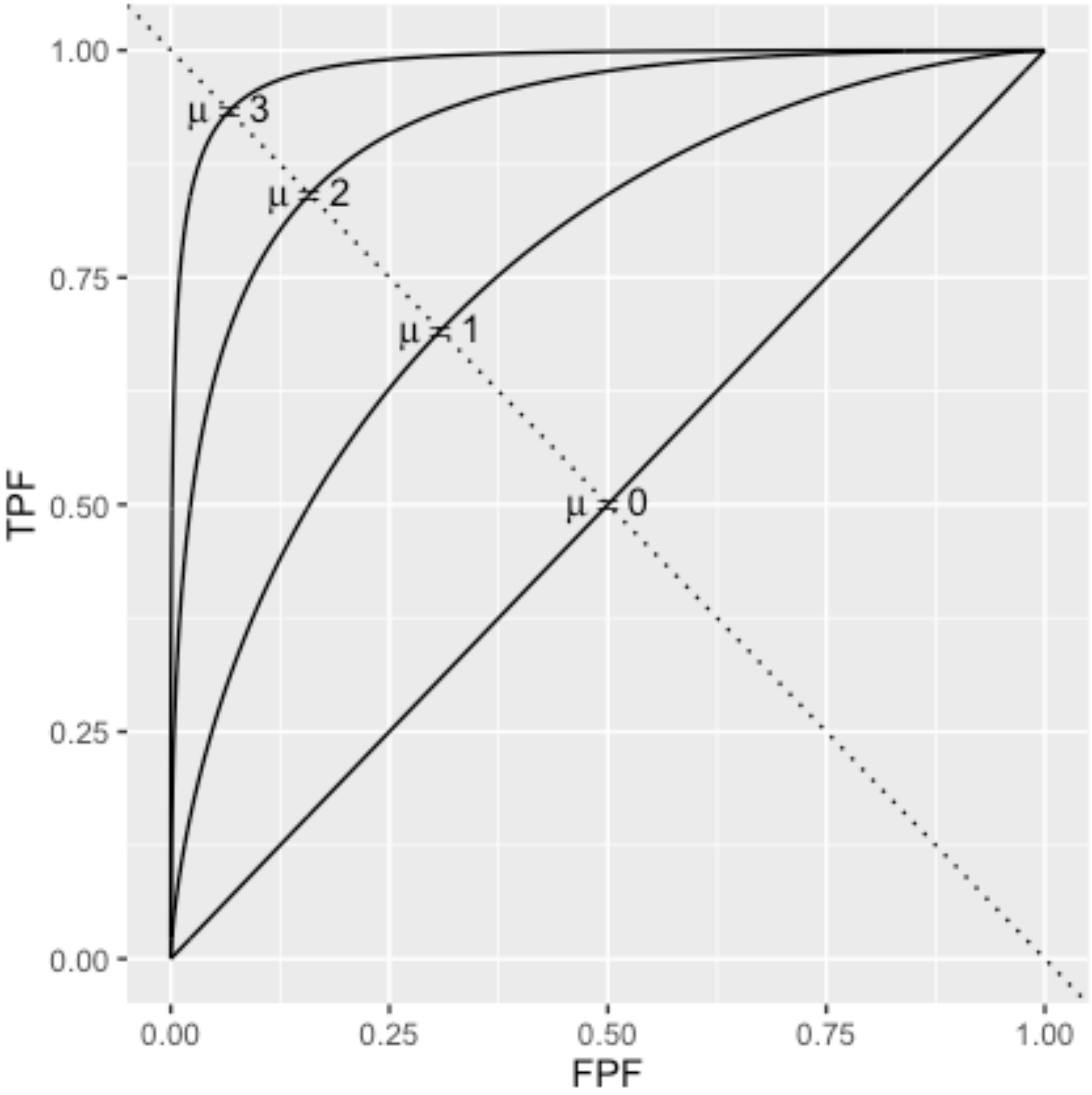

Fig. 3.4: ROC plots predicted by the equal variance binormal model for different values of μ . As μ increases the intersection of the curve with the negative diagonal moves closer to the ideal operating point, (0,1) at which sensitivity and specificity are both equal to unity. [The curves were generated using **`mainRocCurveEqualVarianceModel.R`**.]

3.9.1: The chance diagonal

In Fig. 3.4 the ROC curve for $\mu = 0$ is the *positive diagonal* of the ROC plot, termed the *chance diagonal*. Along this curve TPF = FPF and the observer's performance is at *chance level*. In the equal variance binormal model, for $\mu = 0$, the *pdf* of the diseased distribution is identical to that of the non-diseased distribution: both are centered at the origin. Therefore, no matter the choice of threshold ζ, TPF = FPF. Setting $\mu = 0$ in Eqn. (3.22) and (3.23) yields:

$$TPF\left(\zeta\right) = FPF\left(\zeta\right) = \Phi\left(-\zeta\right) \quad . \qquad \textbf{(3.25)}$$

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

3.9.1.1: The guessing observer

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

3.9.2: Symmetry with respect to negative diagonal

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

$$\left.\begin{aligned} FPF(\zeta) &= \Phi\left(-\frac{\mu}{2}\right) \\ TPF(\zeta) &= \Phi\left(\frac{\mu}{2}\right) \end{aligned}\right\} . \qquad \textbf{(3.26)}$$

The first equation implies:

$$1 - FPF(\zeta) = 1 - \Phi\left(-\frac{\mu}{2}\right) = \Phi\left(\frac{\mu}{2}\right) . \qquad \textbf{(3.27)}$$

Therefore,

$$TPF(\zeta) = 1 - FPF(\zeta) . \qquad \textbf{(3.28)}$$

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

3.9.3: 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:

$$A_{z;\sigma=1} = \int_0^1 TPF(\zeta)\, d\left(FPF(\zeta)\right) = \int_0^1 FPF(\zeta)\, d\left(TPF(\zeta)\right) . \qquad \textbf{(3.29)}$$

Eqn. (3.29) has the following equivalent interpretations:

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

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

By convention, the symbol A_z is used for the area under the *binormal model predicted* ROC curve. In Eqn. (3.29), the subscript $\sigma = 1$ is necessary to distinguish it from another one corresponding to the unequal variance binormal model to be derived in **Chapter 06**. It can be shown that (the proof is in **Chapter 06**):

$$A_{z;\sigma=1} = \Phi\left(\frac{\mu}{\sqrt{2}}\right) \quad . \qquad \textbf{(3.30)}$$

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

3.9.4: Properties of the equal-variance binormal model ROC curve

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

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

c) Since μ is positive, the slope of the equal-variance binormal model curve at the origin (0,0) is infinite and the slope at (1,1) is zero, and the slope along the curve is always non-negative and decreases monotonically as the operating point moves up the curve.

d) AUC is a monotone increasing function of μ. It varies from 0.5 to 1 as μ varies from zero to infinity.

3.9.5: Comments

Property (b): since the operating point coordinates can both be expressed in terms of Φ functions, which are monotone in their arguments, and in each case the argument ζ appears with a negative sign, it follows that as

ζ is lowered both *TPF* and *FPF* increase. In other words, the operating point corresponding to $\zeta - d\zeta$ is to the upper right of that corresponding to ζ (assuming $d\zeta > 0$)

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

$$\frac{d\left(TPF\right)}{d\left(FPF\right)} = \frac{d\left(\Phi\left(\mu-\zeta\right)\right)}{d\left(\Phi\left(-\zeta\right)\right)} = \frac{\phi\left(\mu-\zeta\right)}{\phi\left(-\zeta\right)} = e^{\mu\left(\zeta-\mu/2\right)} \propto e^{\mu\zeta} \geq 0 \quad . \tag{3.31}$$

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

$$d\Phi\left(\zeta\right) = P\left(\zeta < Z < \zeta + d\zeta\right) = \phi\left(\zeta\right)d\zeta \quad . \tag{3.32}$$

Since the slope of the ROC curve can be expressed as a power of *e*, it is always non-negative. Provided $\mu > 0$, then in the limit $\zeta \to \infty$, the slope at the origin approaches ∞. Eqn. (3.31) also implies that in the limit $\zeta \to -\infty$ that the slope of the ROC curve at the end-point (1,1) approaches zero. For constant $\mu > 0$, the slope is a monotone increasing function of ζ. As ζ *decrease* from +∞ to -∞, the slope decreases monotonically from +∞ to 0.

Fig. 3.4 is the ROC curve for the equal-variance binormal model for $\mu = 3$. The entire curve is defined by μ. Specifying a particular value of ζ corresponds to specifying a particular point on the ROC curve. In Fig. 3.4 the open circle corresponds to the operating point (0.159, 0.977) defined by ζ = 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.

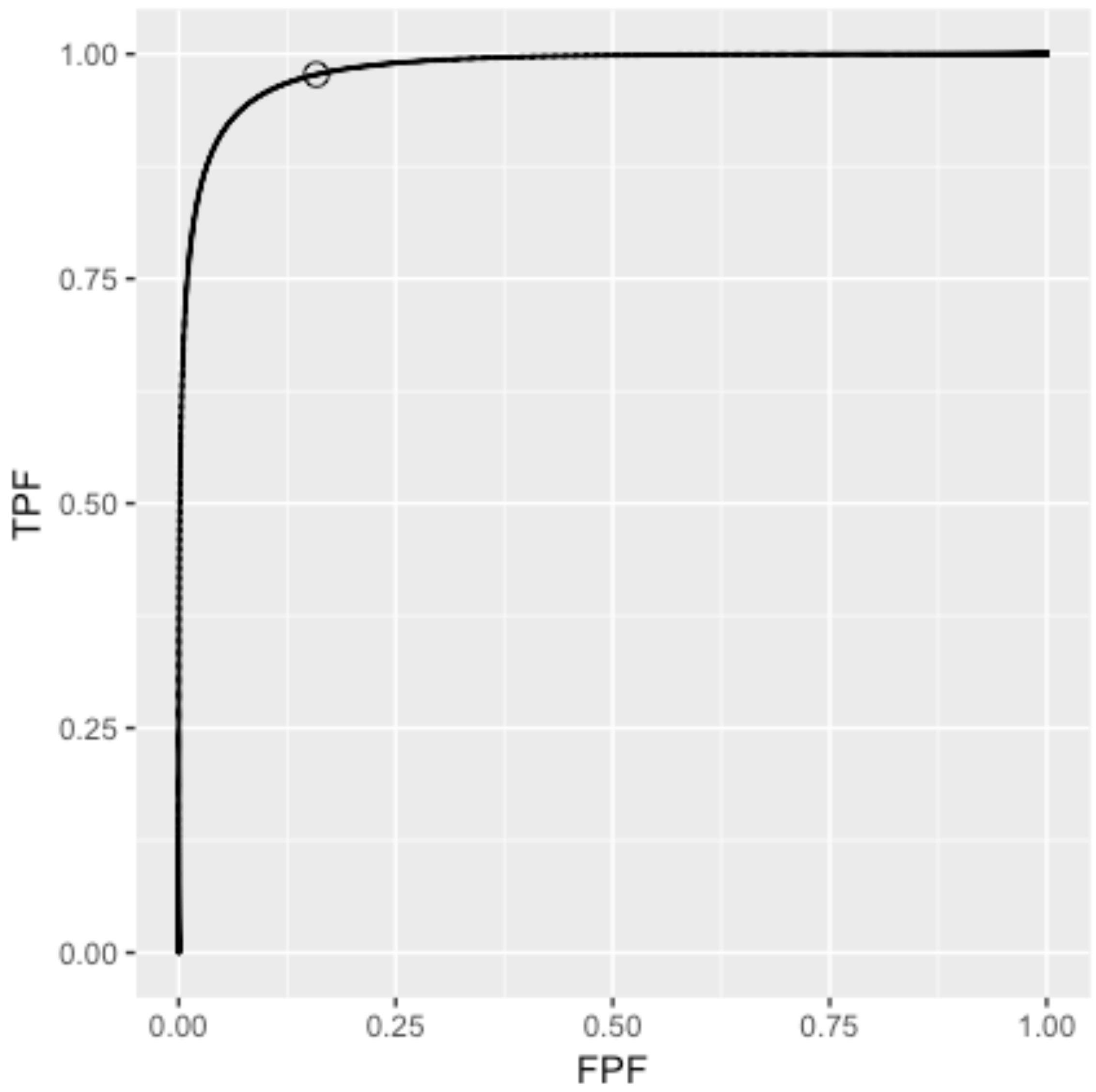

Fig. 3.5: ROC curve predicted by equal variance binormal model for $\mu = 3$. The circled operating point corresponds to $\zeta = 1$. The operating point falls exactly on the curve, as these are analytical results. Due to sampling variability, with finite numbers of cases, this is not observed in practice.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

3.9.6: Physical interpretation of μ

As a historical note, μ is equivalent[2] 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[1,5], it dates to the early 1940s when one had the problem[6] of detecting faint radar reflections from a plane against a background of noise. The reader may be aware of the "rule-of-thumb" that if SNR exceeds three the target is likely to be detected. It will be shown later that the area under the ROC curve is the probability that a diseased case Z-sample is greater than that of a non-diseased one, **Chapter 05**. It is also the probability that a correct choice will be made if a diseased and non-diseased case are shown simultaneously and the observer is asked to pick the diseased case, which is the 2 alternative forced choice (2AFC) paradigm[1,7]. The following code snippet shows that for $\mu = 3$, the probability of detection is 98.3%.

3.9.6.1: Code snippet

```
> pnorm(3/sqrt(2))
[1] 0.9830526
```

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

3.10: Assigning confidence intervals to an operating point

[The notation in the following equations follows that introduced in **Chapter 02**.] A $\left(1-\alpha\right)$ confidence interval (CI) of a statistic is that range[c] expected to contain the true value of the statistic with probability $\left(1-\alpha\right)$. It should be clear that a 99% CI is wider than a 95% CI, and a 90%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 in **Chapter 07**). With a distributional assumption, the method being described now, the parameters of the distribution can be estimated, and since the distribution accounts for variability, the needed confidence interval estimate follows. With quantities *TPF* and *FPF*, each of which involves a ratio of two integers, it is convenient to assume a binomial distribution for the following reason: the diagnosis "non-diseased" vs. "diseased" is a Bernoulli trial, i.e., one whose outcome is binary. A Bernoulli trial is like a coin-toss, a special coin whose probability of landing "diseased" face up is *p*, which is not necessarily 0.5 as with a real coin. The limits on *p* are $0 \le p \le 1$. It is a theorem in statistics[8] that the total number of Bernoulli outcomes of one type, e.g., $\#FP$, is a binomial-distributed random variable, with success probability $\widehat{FPF}$ and trial size K_1. The circumflex denotes an estimate.

[c] Since the observed value is a realization of a random variable, the 95% confidence interval is also a random *range* variable. If the trial were repeated many times, the true value would be included in 95% of the confidence intervals. An individual estimated range is *not* guaranteed to contain the true value with 95% probability.

$$\# FP \sim B\left(K_1, \widehat{FPF}\right) \quad . \tag{3.33}$$

In Eqn. (3.33), $B\left(n,p\right)$ denotes the binomial distribution[8] with success probability *p* and trial size *n*:

$$\left.\begin{array}{l} k \sim B\left(n,p\right) \\ k = 0,1,2,...,n \end{array}\right\} \quad . \tag{3.34}$$

Eqn. (3.34) states that *k* is a random sample from the binomial distribution $B\left(n,p\right)$. For reference, the *probability mass function* (*pmf*) of $B\left(n,p\right)$ is defined by (the subscript denotes a binomial distribution):

$$pmf_{Bin}\left(k;n,p\right) = \left(\begin{array}{c} n \\ k \end{array}\right) p^k \left(1-p\right)^{n-k} \quad . \tag{3.35}$$

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

The binomial coefficient $\left(\begin{array}{c} n \\ k \end{array}\right)$ appearing in Eqn. (3.35), to be read as "*n pick k*", is defined by:

$$\left(\begin{array}{c} n \\ k \end{array}\right) = \frac{n!}{k!\left(n-k\right)!} \quad . \tag{3.36}$$

From the properties of the binomial distribution the variance of $\# FP$ is given by:

$$\sigma_{\# FP}^2 = Var\left(\# FP\right) = K_1 \widehat{FPF}\left(1 - \widehat{FPF}\right) \quad . \tag{3.37}$$

Therefore, the distribution of *FPF* is:

$$FPF \sim B\left(\widehat{FPF}, \frac{\widehat{FPF}\left(1-\widehat{FPF}\right)}{K_1}\right) \quad . \tag{3.38}$$

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

$$\sigma_{FPF}^2 \equiv \frac{1}{K_1}\widehat{FPF}\left(1-\widehat{FPF}\right) \quad . \tag{3.39}$$

For large n the binomial distribution $B(n,p)$ asymptotically approaches[8] a normal distribution with mean np and variance $np(1-p)$:

$$B(n,p) \rightarrow N\left(np, np(1-p)\right) \quad . \tag{3.40}$$

Replacing n with K_1 and p with $\widehat{FPF}$, the binomial distribution of $\# FP$ (the hash symbol denotes total number of counts of the quantity appearing to its right) asymptotically approaches a normal distribution with mean $K_1\widehat{FPF}$ and variance $K_1\widehat{FPF}\left(1-\widehat{FPF}\right)$:

$$\# FP \sim N\left(K_1\widehat{FPF}, K_1\widehat{FPF}\left(1-\widehat{FPF}\right)\right) \quad . \tag{3.41}$$

Eqn. (3.41) implies that for large K_1, FPF follows the normal distribution (the variance has to be divided by the square of the number of non-diseased cases):

$$FPF \sim N\left(\widehat{FPF}, \widehat{FPF}\left(1-\widehat{FPF}\right)/K_1\right) \quad . \tag{3.42}$$

It follows that:

$$\left.\begin{array}{l} \widehat{\sigma_{FPF}^2} = \widehat{FPF}\left(1-\widehat{FPF}\right)/K_1 \\ FPF \sim N\left(\widehat{FPF}, \widehat{\sigma_{FPF}^2}\right) \end{array}\right\} \quad . \tag{3.43}$$

Translating the mean to zero and dividing by the square root of the variance, it follows that:

$$\frac{FPF - \widehat{FPF,}}{\widehat{\sigma}_{FPF}} \sim N\left(0,1\right) \quad . \qquad \textbf{(3.44)}$$

In practice, the normal approximation is adequate if *both* of the following two conditions are *both* met (i.e., $\widehat{FPF}$ is not too close to zero or 1):

$$K_1\widehat{FPF} > 10 \qquad \& \qquad K_1\left(1 - \widehat{FPF}\right) > 10 \quad . \qquad \textbf{(3.45)}$$

From the properties of the normal distribution, it follows that an approximate symmetric $\left(1-\alpha\right)\times 100\%$ confidence interval for FPF is (with this definition a 95% confidence interval corresponds to choosing $\alpha = 0.05$):

$$CI_{1-\alpha}^{FPF} = \left(\widehat{FPF} - z_{\alpha/2}\widehat{\sigma}_{FPF}, \widehat{FPF} + z_{\alpha/2}\widehat{\sigma}_{FPF}\right) \quad . \qquad \textbf{(3.46)}$$

In Eqn. (3.46) $z_{\alpha/2}$ is the *upper* $\alpha/2$ quantile of the unit normal distribution, i.e., the area to the *right* under the unit normal distribution *pdf* from $z_{\alpha/2}$ to infinity equals $\alpha/2$. It is the complement of the Φ^{-1} introduced earlier; the difference is that the latter uses the area to the *left*:

$$\left.\begin{aligned} z_{\alpha/2} &= \Phi^{-1}\left(1-\alpha/2\right) \\ \alpha/2 &= \int_{z_{\alpha/2}}^{\infty} \phi\left(z\right)dz = 1 - \Phi\left(z_{\alpha/2}\right) \end{aligned}\right\} \quad . \qquad \textbf{(3.47)}$$

The reader should be convinced that the two equations in Eqn. (3.47) are consistent.

Similarly, an approximate symmetric $\left(1-\alpha\right)\times 100\%$ confidence interval for TPF is:

$$CI_{1-\alpha}^{TPF} = \left(\widehat{TPF} - z_{\alpha/2}\widehat{\sigma}_{TPF}, \widehat{TPF} + z_{\alpha/2}\widehat{\sigma}_{TPF}\right) \quad . \qquad \textbf{(3.48)}$$

In Eqn. (3.48),

$$\widehat{\sigma^2_{TPF}} \equiv \frac{1}{K_2}\widehat{TPF}\left(1-\widehat{TPF}\right) \quad . \tag{3.49}$$

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[9] 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 in "**`mainConfidenceIntervals.R`**" in Online Appendix 3.B. Ensure that **`mu`** is set to 1.5 at line 5. **`Source`** the code to get:

3.10.1: Code Output

```
> source('~/book2/A ROC analysis/A3 ModelingBinaryTask/software/mainConfidenceIntervals.R')
alpha =  0.05 K1 =  99 K2 =  111 mu =  1.5 zeta =  0.75
Specificity =  0.778 Sensitivity =  0.847
approx 95% CI on Sp =  0.696 0.86
Exact 95% CI on Sp =  0.683 0.855
approx 95% CI on Se =  0.78 0.914
Exact 95% CI on Sp =  0.766 0.908
```

The exact and approximate confidence intervals are close to each other. Now change **`mu`** to five, thereby assuring that both sensitivity and specificity will be close to unity; **`source`** the code to get:

3.10.2: Code Output

```
> source('~/book2/A ROC analysis/A3 ModelingBinaryTask/software/mainConfidenceIntervals.R')
alpha =  0.05 K1 =  99 K2 =  111 mu =  5 zeta =  2.5
Specificity =  0.99 Sensitivity =  0.991
approx 95% CI on Sp =  0.97 1.01
Exact 95% CI on Sp =  0.945 1
approx 95% CI on Se =  0.973 1.01
Exact 95% CI on Sp =  0.951 1
```

The approximate confidence interval is clearly incorrect as it extends beyond unity. The exact confidence in calculated by avoiding the normal approximation, Eqn. (3.40). Instead, one numerically calculates a confidence interval $\left(CI_{1-\alpha}^{lower}, CI_{1-\alpha}^{upper}\right)$ such that at most $\alpha/2$ of the binomial distribution *pmf* is below the lower limit $CI_{1-\alpha}^{lower}$ and at most $\alpha/2$ of the binomial distribution *pmf* is above the upper limit $CI_{1-\alpha}^{upper}$.

3.10.3: Exercises

The preceding demonstration should give the reader an idea of the power of **R**, and that one does not have to be a statistician or an expert programmer to benefit from it. Experiment with different values of parameters (e.g., try reducing the numbers of cases, or change the seed), and/or run the code in debug mode, until it makes sense. Change α appropriately to be convinced that a 99% CI is wider than a 95% CI, and a 90% CI is narrower. Finally, as a greater challenge, repeat the code without initializing seed, so that the samples are independent, and confirm that the 95% confidence intervals indeed include the correct analytical value with probability 95%.

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

In this study[10] 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 (ABR)[11,12]. One hundred eight (108) ABR-certified radiologists from these centers gave blinded interpretation to a common set of 79 randomly selected enriched screening cases containing 45 cases with cancer and the rest normal or with benign lesions. Ground truth for these women had been established either by biopsy or by 2-year follow-up (establishing truth is often the most time consuming part of conducting an ROC study). The observed range of sensitivity (TPF) was 53% and the range of FPF was 63%; the corresponding range for AUC was 21%, Table 3.3.

Table 3.3: This table illustrates the variability of sample of 108 board-certified 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% compared to 53% for sensitivity and 63% for specificity).

Measure	**Min%**	**Max%**	**Range%**
Sensitivity	46.7	100.0	53.3
Specificity	36.3	99.3	63.0
ROC AUC	0.74	0.95	0.21

In Fig. 3.6, a schematic of the data, if one looks at the points labeled (B) and (C) one can mentally construct a smooth ROC curve that starts at (0,0), passes roughly through these points and ends at (1,1). In this sense, the intrinsic performances (i.e., AUCs or equivalently the μ parameter) of the two radiologists are similar. The only difference between them is that radiologist (B) is using lower threshold relative to the radiologist (C). Radiologist (C) is more concerned with minimizing FPs while radiologist (B) is more concerned with maximizing sensitivity. By appropriate feedback radiologist (C) can perhaps be induced to change the threshold to that of radiologist (B), or they both could be induced to achieve a happy compromise. An example of feedback might be: "*you are missing too many cancers and this could get us all into trouble; worry less about reduced specificity and more about increasing your sensitivity*". In contrast, radiologist (A) has intrinsically greater performance (B) or (C). No change in threshold is going to get the other two to a similar level of performance as radiologist A. Extensive training will be needed to bring the underperforming radiologists to the expert level represented by radiologist A.

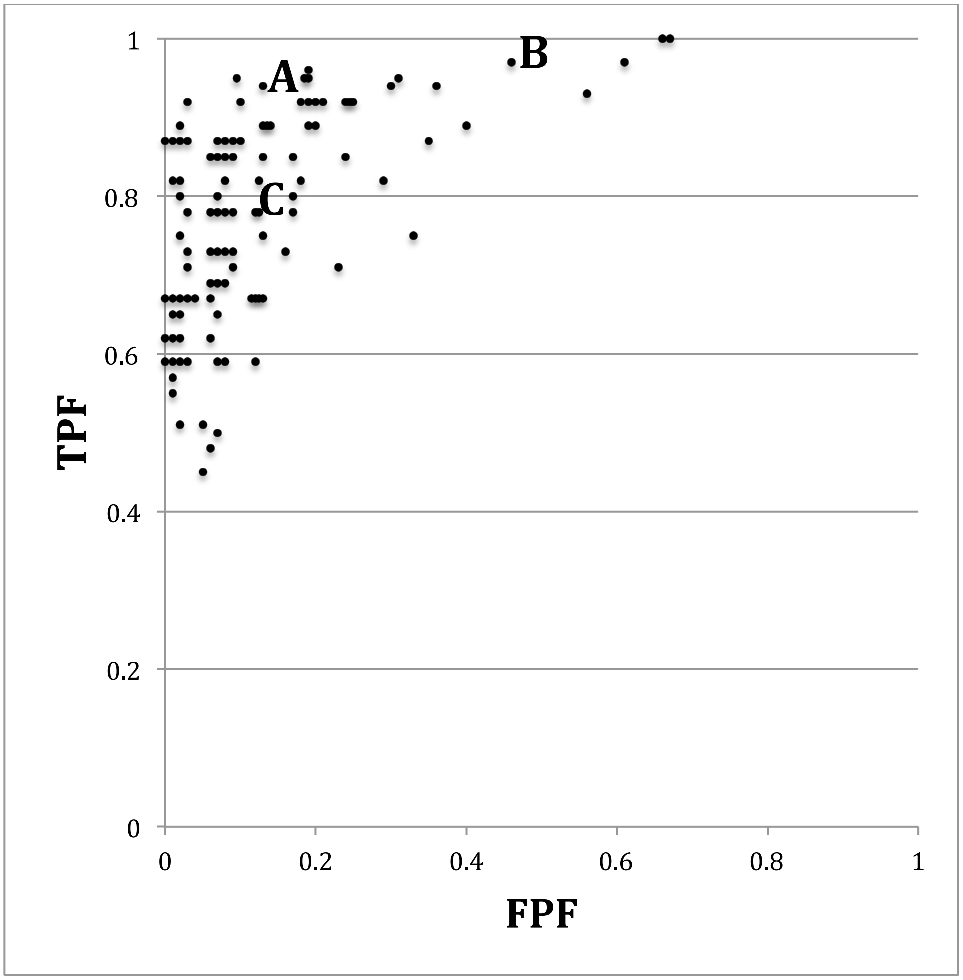

Fig. 3.6: Schematic, patterned from the Beam et al study, showing the ROC operating points of 108 mammographers. Wide variability in sensitivity (40%) and specificity (45%) are evident. Radiologists (B) and (C) appear to be trading sensitivity for specificity and vice versa, while radiologist A's performance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

Fig. 3.6 and Table 3.3 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 ζ, the area under the ROC plot is independent of ζ. 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% while that for specificity is 63%. In contrast, the range for AUC is only 21%. This means that much of the observed variations in sensitivity and specificity are due to variations in thresholds, and using AUC eliminates this source of variability. Decreased variability of a measure is a highly desirable characteristic as it implies the measurement is more precise, making it easier to detect genuine changes between readers and / or modalities.*

3.12 Discussion

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

The ROC curve is the locus of operating points as the threshold is varied. It and AUC are completely defined by the μ parameter of the equal variance binormal model. Since both are independent of reporting threshold ζ, they overcome the ambiguity inherent in comparing sensitivity/specificity pairs. Both are scalar measures of performance. AUC is widely used in assessing imaging systems. It should impress the reader that a subjective internal sensory perception of disease presence and an equally subjective internal threshold can be translated into an objective performance measure, such as the area under an ROC curve or equivalently, the μ parameter. The latter has the physical meaning of a perceptual signal to noise ratio.

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

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

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

3.13 References

1. Green DM, Swets JA. *Signal Detection Theory and Psychophysics.* New York: John Wiley & Sons; 1966.
2. Macmillan NA, Creelman CD. *Detection Theory: A User's Guide.* New York: Cambridge University Press; 1991.
3. Skaane P, Bandos AI, Gullien R, et al. Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population-based screening program. *Radiology.* 2013;267(1):47-56.

4. Bandos AI, Rockette HE, Gur D. Use of likelihood ratios for comparisons of binary diagnostic tests: Underlying ROC curves. *Medical physics.* 2010;37(11):5821-5830.

5. Egan JP. *Signal Detection Theory and ROC Analysis.* first ed. New York: Academic Press, Inc.; 1975.

6. A Statistical Theory of Target Detection by Pulsed Radar. Santa Monica, CA: U. S. Air Force; 1947.

7. Burgess AE. Comparison of receiver operating characteristic and forced choice observer performance measurement methods. *Med Phys.* 1995;22(5):643-655.

8. Larsen RJ, Marx ML. *An Introduction to Mathematical Statistics and Its Applications.* 3rd ed. Upper Saddle River, NJ: Prentice-Hall Inc; 2001.

9. Conover WJ. *Practical nonparametric statistics.* New York: John Wiley & Sons. ; 1971.

10. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample. *Archives of Internal Medicine.* 1996;156(2):209-213.

11. Hendrick RE, Bassett L, Botsco MA, al. e. *Mammography Quality Control Manual.* 4 ed: American College of Radiology, Committee on Quality Assurance in Mammography; 1999.

12. Barnes GT, Hendrick RE. Mammography accreditation and equipment performance. *Radiographics.* 1994;14(1):129-138.