ROC, Precision-Recall:
software with confidence limits and Bayes (predicted value) calculations.

Robert E. Wheeler

March 13, 2011

Contents

1 Introduction

Although often obscured by its presentation[1], ROC is really a graphical method for comparing two empirical distributions. Graphics are always useful as a method of summarizing data and often they can reveal aspects that are not obvious by other means, and of course this is ROC's greatest strength. One of the more important problems that is easy with ROC but difficult when looking at data from a pair of empirical distributions is the picking of a cut-off point that will discriminate between samples from the populations. This is an important problem that needs careful treatment. Unfortunately, a ROC chart is too often taken as a thing of permanence and conclusions are drawn without considering the statistical fluctuations that are natural in any real world activity. I suspect that some users of ROC feel that such fluctuations are minor because they have collected large samples – but do they really know?

A ROC chart is independent of what is called the "skew" of the data; that is the relative sizes of the two data sets. This has its advantages, but it can also fool the user. One really needs to take the prevalence of the effect into account in order to well judge the values read from a ROC graph. Without doing so, many foolish conclusions can be drawn. For example, a cut-off may appear from a ROC graph to discriminate well between two populations, but when it is used in actual practice it may be found that the number of false positives is extraordinarily high. This is the case with prostate-specific antigen tests and notoriously with the infant test PKS, which never misses a positive case, but has a false positive rate of 90%.

Precision-recall is a relabeling of a Bayes calculation, which is often referred to as a "predicted value" calculation. The Bayes calculation derives from a ROC calculation by the inclusion of prevalence. The difference between Bayes and Precision-recall, is that prevalence is expressed in terms of "skew," which is the ratio of the population sizes. The statistical properties of ROC calculations carry over to Precision-recall calculations, with the result that a Precision-recall chart is simply a different visual display of the data. Hopefully these relationships will become clear in the sequel, but first ROC must be discussed.

A ROC chart may be used to describe the efficacy of a classifier which allocates items into one of two categories depending on whether or not they exceed a threshold. Multiple classifiers may be used simultaneously and the decision about the item may be made by choosing one of the classifiers according to a random event such that the expectation of the choice is optimal. This involves examining the convex hull of the ROC curves, and picking thresholds and blending probabilities to achieve the expectation.

A good general survey may be found in [Fawcett, 2006]. The software rocplus() supports most of the things he describes.

[1] In terms of unobservable, continuous underlying variables.

2 What is ROC

First of all, one must have a problem involving two things and the desire to discriminate between them. ROC simply counts the number of times that data from one condition exceeds that from the second. A large number implies that the first condition is superior to the second: in other words it implies that the mass of the data from the first condition lies to the right of the second. This is a very sensible thing to do. The number of exceedences divided by the total number possible is called the AUC; but it has another name in the statistical community, the Mann-Whitney form of the Wilcoxon, Mann-Whitney statistic. This statistic is an old friend and there is very little that is not known about it's behavior. A ROC curve is simply a graphical representation of the values that make up the AUC calculation. It is easiest to illustrate this numerically.

Suppose one has two sets of data from two conditions, A and B, like in Table (1), which I have sorted into an decreasing order; and below the data in the Ex row are the exceedences of B over A. The 5 in the second column gives the number of B's that exceed 20 in A. Similarly 7 in the third column gives the number of B's that exceed 17. etc. The AUC is 0.75, which is the sum of the Ex row divided by 100, the maximum possible value. The ROC curve simply plots these values as in Table (2), where the top edge of the vertical B bars above the A values are marked with x's. It should be obvious from this plot that the "area under the curve" is AUC.

Table 1: Illustrative data from two conditions

A	31	20	17	16	14	10	8	7	1	0	
B	34	28	25	22	21	19	18	15	13	6	
Ex	1	5	7	7	8	9	9	9	10	10	AUC=75/100

Table 2: ROC curve

	10									x	x
	9						x	x	x	o	.
	8					x	.	.	o	.	.
	7			O	x	.	.	o	.	.	.
	6			.	.	.	o	.	.	.	.
B	5		x	.	.	o	.	.	.	.	.
	4		.	.	o	.	.	.	.	.	.
	3		.	o	.	.	.	.	.	.	.
	2		o	.	.	.	.	.	.	.	.
	1	x	.	.	.	.	.	.	.	.	.
		1	2	3	4	5	6	7	8	9	10
						A					

When all the B's are larger than the A's, the Ex row will contain nothing but 10's and the AUC will be unity (100/100), and the ROC curve will have x's only in the top row. On the other hand when the A's and B's are equal, the Ex row will contain nothing but 5's and the AUC will be 0.5 (50/100). A line for this is customarily shown on a ROC curve as in Table (2), where it is marked with o's.

Table 3: Distributions

B				1				1	1	1	1	2		1	1			1
A		2			2	1		1	1	1	1						1	
	0	2	4	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34

The axes are labeled from 1 to 10 because they represent the percentages of the values that exceed various thresholds. Table (3) shows the raw data aligned with increasing thresholds. For example, we see that 7 (70%) of the B values are 18 or larger, while 3 (3%) of the A values are 18 or larger. This produces the point (3,7) on the plot. Similarly, 90% of the B values and 60% of the A values are 10 or larger[2]. The values on the two axes are typically referred to as "Sensitivity" and "1-Specificity".

There is an optimum point on the curve. Assuming *a priori* that samples from the two distributions are equally probable, then the optimum point is the one furtherest from the 50-50 line, and is tangent to the curve at a 45° angle . In this illustration, it is the point (3,7), Marked in Table (2) with a large O, corresponding to the B threshold of 18. The optimum point is the point at which the error is a minimum. The points along the curve represent decision thresholds when the data is viewed as arriving in a stream, and the receiver makes a decision for each datum as to the distribution from which it came. If the datum value exceeds the threshold, it is classified one way, if it does not, the other. The optimum point is the threshold at which these errors are a minimum. When the samples are not equally probable, the optimum point is located elsewhere on the curve, which will be discussed later.

Figure (1) shows a plot produced by the software in which both the optimum point and a selection of threshold values are shown. Figure (2) shows the "smoothed" version, which is arguably a better representation of the conditions that generated the data. Smoothed curves always pass through the "rough" data. The sharp eyed observer will no doubt notice that the threshold values are points on the curves and not in the space.

The uncertainty in a ROC curve is shown by confidence bands as in Figure (3). The distance between these bands is large because the difference between the distributions is barely significant. The sample size is after all very small. The lower confidence band is actually below the 50-50 line. The confidence bands represent shifts in location between the two data sets. The upper band, for example, represents the ROC curve that would occur should the location of the B data shift to correspond to the upper confidence limit on AUC.

A Bayesian curve is shown in Figure (4). Such curves have many names: one of the more popular is "Precision-Recall." I will say more about this later, but for the moment just that this curve shows population probability estimates which are pretty essential in making any sort of prediction from the data.

The software prints out the AUC value, it's p-value, and a 95% confidence interval. It also prints out the optimum value, its location on the plot, and a confidence interval in

[2]The threshold numbers are grouped by two in this table, which means that one must look at the raw data to double check the points on the curve.

the vertical (sensitivity in this case) direction. The output looks like the following.

```
$AUC
               AUC p-value 0.025% 0.975%
Mann-Whitney 0.75  0.0315   0.46  0.905

$parity
             value 1-Specificity Sensitivity 0.025% 0.975%
Optimum Value    15         0.321       0.778   0.28   0.98

$TIES
[1] "0.0%"
```

Figure 1: Illustrative ROC curve

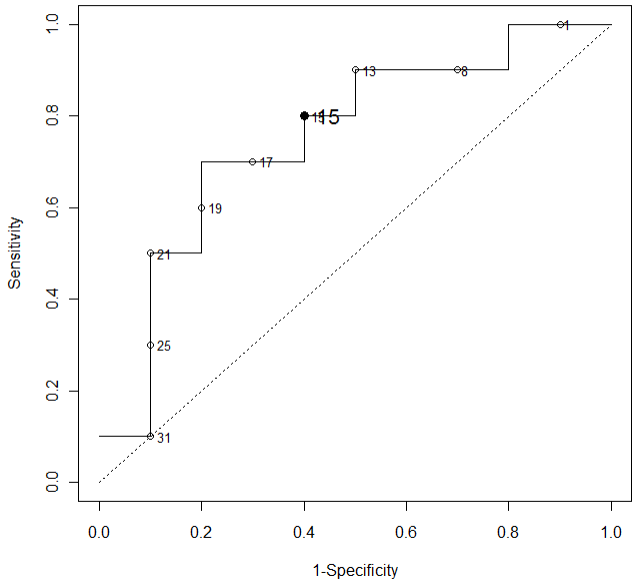
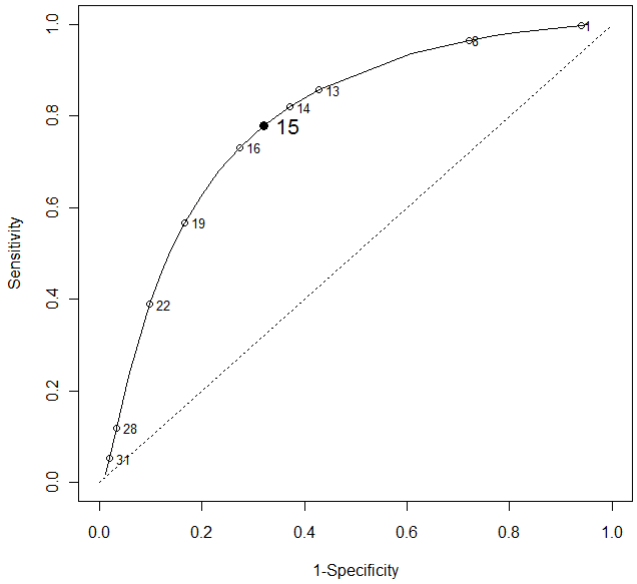


Figure 2: Illustrative smooth ROC curve



3 ROC in general

The basic problem addressed by ROC has a long history. It was at the heart of the first statistical investigations, when [Fechner, 1860] performed the first psychophysical investigations, and [Peirce and Jastrow, 1884] followed up with the first randomized statistical experiments. Later [Pearson, 1904] devised the contingency table, which is the basic two by two table used to describe the four possible classifications into which a decision process may be parceled. Very briefly, the problem in modern terms is to "detect a signal in a noisy environment." The problem achieved its modern form and its name from a research activity set up to study wartime problems, which of course did not solve the problems

Figure 3: Illustrative smooth ROC curve

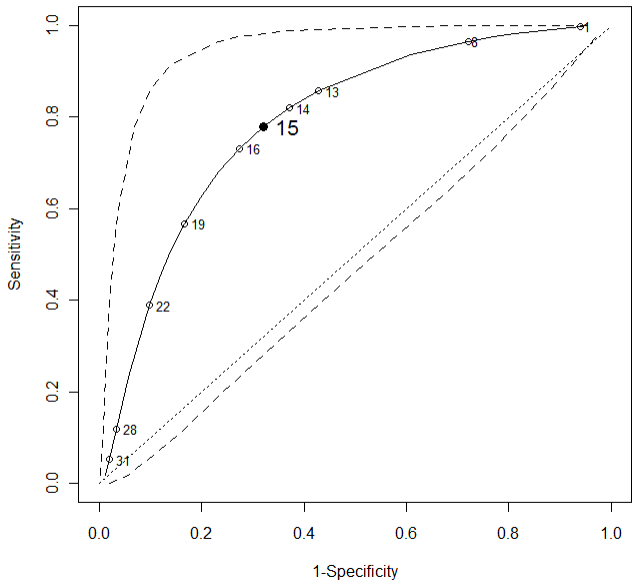
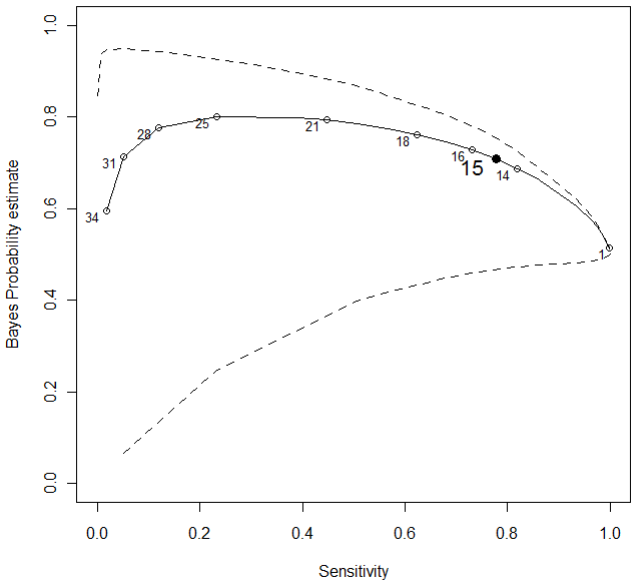


Figure 4: Illustrative smooth Bayes curve



in a timely fashion, but did persist and the long gestation finally produced the present solution some ten years after the end of WWII.

The problem that inspired the effort was that of deciding whether or not a radar signal represented German aircraft or a flock of birds. [Marcum, 1947] writing just after the end of the war states the problem in terms of simply detecting an aircraft, but the birds problem was the inspiration for the effort. The Electronics Defense Group, which was set up to engage scientists in the war effort, at the University of Michigan, cast the problem in its present form and [Peterson et al., 1953] coined the name ROC from "receiver operating characteristic." [Tanner and Swets, 1954] working in the same group, applied these ideas to the visual system, and of course later, [Green and Swets, 1966] wrote the definitive book, the latest printing contains an enormous bibliography of papers and books using the methodology published during the years 1967 to 1988. I have no idea of the vast list that might be published nowadays.

As I noted in section (2), ROC is a graphical method for comparing two sets of data. In many cases this is fairly evident from the problem at hand. For example, a medical researcher attempting to choose a threshold for a test can hardly help but be aware of a comparison between the two sets of data. In other cases, the data paradigm is not as obvious. In a machine learning application, for example, the data is really nothing but a "training" set.

A common rubric is to describe the classification rule in terms of a "confusion matrix," which is nothing more than a "contingency table," such as in Table (4), where I have taken some liberty with the usual names for the elements in such a table – the usual names are really confusing. A confusion matrix contains the counts of the data allocated to the

four possible categories. If one assumes that the columns sum to the margins, so that $T = H + M$, then "sensitivity" is H/T and "specificity" is P/C, hence a ROC plot has H/T on the vertical axis and F/C on the horizontal – in other words it plots the two "Signal" categories against each other.

Table 4: Confusion Matrix

Response	(T) Treatment	(C) Control
Signal	(H) Hit	(F) Fail
Noise	(M) Miss	(P) Pass

The confusion matrix is used to tabulate data from a "classifier" of some sort. It might be the "rights" and "wrongs" from a psychophysics test to determine a threshold for some sensation; it might be the "successes" and "failures" of a forecast of icy roads by a weather bureau; it might be a pathologist's classification of a tumor; etc. As a contingency table, one can calculate a probability relevant to the effect of the treatment. This sort of thing is quite useful, and has been so for a hundred years; however, more things are possible.

If one can collect the data in stages, then one can investigate the changes in the decision process as the stages progress. If these stages can be associated with a numerical value, then one can plot the data as a function of this value; which of course is what ROC does. One may view the horizontal line between Signal and Noise in Table (4) as a threshold that can be raised or lowered. As it is raised the proportions (H/T and F/C) in the Signal row which start out at zero, increase and become larger tending toward unity. The trace of these two proportions is the ROC curve. In general the Positives increase faster than the Fails, leading to the hope of a trade-off. [Fawcett, 2006] describes the confusion matrix and ROC in general, so I will not dwell on it here.

A final word, however, seems appropriate. One can imagine continuous random variables as progenitors of signals which are chopped up with a threshold producing a binary response (yes-no perhaps), and the corresponding imaginary graph for these random variables of which the ROC plot is a pale imitation. There may be theoretical reasons for dealing with such ghostly variables, but in the end, one has data drawn from two populations to deal with. It would be well not to get too far away from this fundamental fact. See [Bamber, 1975] for a discussion of the theory involving what I call the "ghost" variables.

4 Confidence intervals and bands

Confidence statements are an essential adjunct to any sort of statistical calculation. In the present case, such statements are needed for AUC and for the curves on the plots in order to avoid the errors that naturally occur when individuals are presented with figures for interpretation. In a sense, however, confidence statements which are developed with great precision and a deal of concentrated thought can be misleading in practice because all too often the random sampling that is essential for their validity is lacking. What good is a precisely derived confidence interval if the data bounces from sample to sample in an

arbitrary manner?

The probabilities cited in this software are probabilities and not estimates, since they are based on permutations of the data at hand, in the same way that the probability of a double six is a probability for a pair of fair dice and not an estimate. (A count of the number of times double sixes are thrown on a casino table, on the other hand, may be used as an estimate of the fairness of the casino's dice.) For a given data set, there are just so many possibilities and all of them are enumerated in order to provide the probabilities cited in the software. These probabilities become estimates when one considers the data as samples from populations. The optimum point, for example, is an estimate because it is based on an assumption about the relative sizes of the two populations. In other words probabilities become estimates when one generalizes the results; and in order for those generalizations to be meaningful, the data must have been randomly chosen from the populations. If it is not, as is often the case, then the estimates will lack validity – they will have only "face validity."

Several methods for placing confidence intervals about AUC estimates are available. They all use a normal approximation $\hat{\mu} \pm k\hat{\sigma}$ to produce confidence limits; the methods differ in how $\hat{\sigma}$ is estimated. They trace back to equation (6.5) in [Noether, 1967], and although [DeLong et al., 1988] derives things differently, her estimate is equivalent. [Cortes and Mohri, 2005] examined and compared the available candidates. They all rely on the assumption of random sampling from two populations.

The distribution of AUC under the null hypothesis of "no difference between the control and treatment" is well established, but there are many alternative hypotheses to consider when the null hypothesis does not hold. In other words, if the distributions are really different then the probabilities applicable to the null case are no longer valid. Wilcox-Mann-Whitney provides a significance test of the hypothesis that the samples in the control and treatment data come from the same distribution, but it makes no assertation about how the distributions differ when the test is rejected – the alternatives can be anything, skewed, multmodal, non-homoschedastic, etc. The methods cited above concern themselves with estimating the standard deviation of the AUC distribution in the non-null case by estimating some characteristics of the two distributions. It should be noted, however, that the Wilcox-Mann-Whitney test has very little power against anything but a shift alternative, see Section (10).

Although the AUC is interesting, it is not the only item of importance. The curve is the *piece de resistance*, and it should be accompanied by confidence bands. [Macskassy and Provost, 2004] survey several methods for producing confidence bands. They all seem quite *ad hoc*.

Although contemplating of the gamut of non-null distributions is interesting, the most important non-null hypothesis is a simple shift, Δ, in location between the two distributions: this is fortunate because of the limited power of the Wilcox-Mann-Whitner test against non-shift alternatives. The confidence problem becomes quite tame if one concentrates on a location shift, since the quantiles under the alternative hypothesis are those under the null hypothesis shifted by a constant, and confidence intervals are easily constructed.

[Bauer, 1972] gives a method for dealing with this problem. In essence one lists all of the nm differences in the values of n controls and m treatments. The median[3] $\hat{\Delta}$ of this list is an estimate of Δ, and the confidence interval for $\hat{\Delta}$ is given by entering the list with the appropriate p-values of the Mann-Whitney statistic: e.g. suppose $m = 25$ and $n = 35$, then the (2.5%, 97.5%) points for the Mann-Whitney statistic are (307, 568); thus the confidence limits for $\hat{\Delta}$ are the (307, 586)th values in the increasingly sorted list of the nm differences (treatment-control).

If m is the median and the confidence limits for $\hat{\Delta}$ are (l,u), then one may obtain confidence limits on AUC, by calculating AUC with treatment+(u-m), and treatment + (l-m). Similarly, confidence bands on the ROC curve may be obtained by drawing curves with the treatments adjusted in the same way.

The methodology for obtaining confidence limits on $\hat{\Delta}$ is built into R for the wilcox.test() function: I merely call it from rocplus(). Simulations indicate that it is on a par with the methods cited above for estimating the AUC standard deviation unless the two distributions are wildly different.

5 Smoothing curves

There is some literature on smoothing the ROC curve, [Zou et al., 1997]. Some of it is quite complicated, but for the most part, attempts are made to best fit a collection of standard distributions, such as the normal, Poisson, beta, etc. This is always problematical. The fact that there exist systems of distributions that are very flexible does not seem to have been explored. Either the Pearson or Johnson system seem appropriate (see [Johnson et al., 1994]), and since I am familiar with the Johnson system and since it is implemented in R[4], I have used it. A description of the system and details of how one fits it from sample quantiles are given in [Wheeler, 1980].

Although the distributions of the data used for a ROC curve will usually follow some standard form, sampling variations can produce anomalies when used to smooth a ROC curve. The Johnson system will always describe the data well[5], and a smooth curve drawn using the system will always trace through the points of a rough stair step plot drawn from the data. This may be checked with the software by drawing both rough and smooth on the same plot. It is a mindless procedure, which is what is wanted: one wants the curve to smoothly traverse the data, and how that curve came to be is seldom of much interest. In fact, data drawn from the unbounded normal distribution is almost always fit by a bounded curve; it thus seems fairly meaningless to attempt to wring some meaning out of the fit – in any case, rocplus() does not report the Johnson form used.

[3]The location estimate is not exactly the median, but I will use median here to avoid unnecessary chatter.

[4]The SuppDists package.

[5]It can on rare occasions fail to fit data, but rocplus() will report that when it happens

6 Likelihood and optimum points

The idea of likelihood is due to [Fisher, 1921], who defined likelihood as a probability density written as a function of a parameter and a variable, but viewed as a function of the parameter instead of the variable. Likelihood has meaning only in connection with likelihood ratios. For ROC, the parameter has only two values designating the two distributions. Figure (5) shows two probability densities, labeled $f_p(x)$ and $f_q(x)$ giving the likelihoods for each of the parameter values, p and q. The likelihood ratio at the center of the plot, $L(x) = f_p(x)/f_q(x)$, has a value of 1: it of course has different values for different x's The likelihood ratio, $L(x)$, may be interpreted as the relative strength of evidence for the two distributions given x. If one observes a large x, then the likelihood ratio suggests a preference for the distribution on the right. The likelihood ratio, which in a sense expresses the relative merits of two distributions, has turned out to be a very important criterion in statistics. For ROC, most criteria for deciding between the two distributions turn out to be functions of the likelihood ratio, as will now be illustrated.

The driving force behind ROC technology is a desire to choose a decision criterion to use in classifying two streams of information, assigning each datum to one or the other source, which we will call "control" and "treatment." This is done by taking a measurement from each datum and classifying it as it does or does not exceed a threshold. A ROC curve, displays a gamut of thresholds, x, in terms of two probabilities; the sensitivity of the threshold to classifying a datum correctly as a treatment, and how it errs in specifying a control datum.

Refer to Table (4) and let $F_p(x)$ be the probability of incorrectly classifying a treatment datum (a miss) and $F_q(x)$ the probability of correctly classifying a control datum (pass), then, the vertical axis plots the sensitivity, $1 - F_p(x)$ (a hit), and the horizontal $1 - F_q(x)$, or 1-specificity (fail). These are genuine probabilities for the data set at hand, but become estimates when generalized to populations.

For any threshold, x, the probability of classification error, miss plus fail, is $e(x)$:

$$e(x) = F_p(x)\lambda + (1 - F_q(x))(1 - \lambda), \tag{1}$$

where λ is the prevalence of the treatment condition in the population. The minimum of $e(x)$ is

$$min(e(x)) = L(x) = \frac{f_p(x)}{fq(x)} \geq \frac{(1-\lambda)}{\lambda} = w, \tag{2}$$

where the little f's represent the derivatives of the big F's, and $L(x)$ is the likelihood ratio function. The optimum threshold is the minimum threshold for which $L(x) \geq w$.

The above result was stated by [Peterson et al., 1953], who show that the likelihood criterion is equivalent to a number of other "optimality" criteria. In addition, one may incorporate costs or any sort of value into equation (2) by replacing w with the relative values for the two conditions. For example, if one does not care for the idea of minimizing

the sum of errors as it is done in equation (1), one can weight the λ in some way. For example if $\lambda = 0.03$ is the prevalence of a disease in a population, but that one believes each hit is worth 10 times that of a fail, then one can use equation (2) with $\lambda = 0.3$.

The use of a likelihood ratio is illustrated in Figure (5), where the (optimum point) likelihood ratio for a λ of 0.5 is shown by the line labeled $L(x) = 1$. As may be seen, this optimum point likelihood ratio balances the mass of the distributions, making both errors and correct decisions equally likely at this threshold. For no other threshold are the errors smaller.

When the prevalence λ is different from 0.5, one of the distributions will be sampled more frequently than the other. The likelihood for $\lambda = 0.2$ is shown in Figure (5) by the line labeled $L(x) = 4$, which is obtained from $\frac{1-\lambda}{\lambda} = \frac{1-0.02}{0.02} = 4$. This may be represented as shown in Figure (5) by the line labeled $L(x) = 4$. Figure (6) shows an alternative representation, where the two distributions are weighted according to $\lambda = 0.2$. It is worth noting that although the sum of the errors is minimized when $L(x) = 4$, the two errors are no longer equal as they were when $L(x) = 1$.

As mentioned in the Introduction, the slope of the line through the optimum point is 45° when the prevalence is 0.5, which corresponds to a likelihood ratio of 1. For other prevalence values the slope changes. It is 8/2 for Figure (6), and this is illustrated in Figure (7) where the slope at the optimum point is 8/2. In fact, the slope at any point on the curve is the ratio of the likelihoods; which means that points on the left hand side of the curve represent high likelihood ratios, while those on the right small ones. If one chooses a threshold on the left hand side of the curve, then one is more likely to classify a datum as a treatment than a control; but of course, this is good only if one can endure a greater error rate for treatments.

7 Bayes estimates, precision-recall and, predicted values

Optimum points are defined in terms of the prevalence of the treatment condition in the population, which makes them estimates because they generalize to the population and are not confined to the samples at hand. As such, they represent points different from the other points on the ROC, curve which are independent of prevalence. The points on the ROC curve are conditional on the two categories, control and treatment, and the coordinates of a point on the curve indicate what will happen for a given threshold with respect to either a genuine control or treatment datum. The vertical axis gives the proportion of true treatment data that exceed various thresholds, while the horizontal axis gives the proportion of true control data that exceed various thresholds. This is grand, but it does not really tell the user the consequence of using a particular threshold when one is in ignorance of the category from which the data comes; for that, one needs the relative proportions of the two sorts of data; in other words one needs to know the prevalence of the treatment in the general population.

Figure 5: Equal Areas

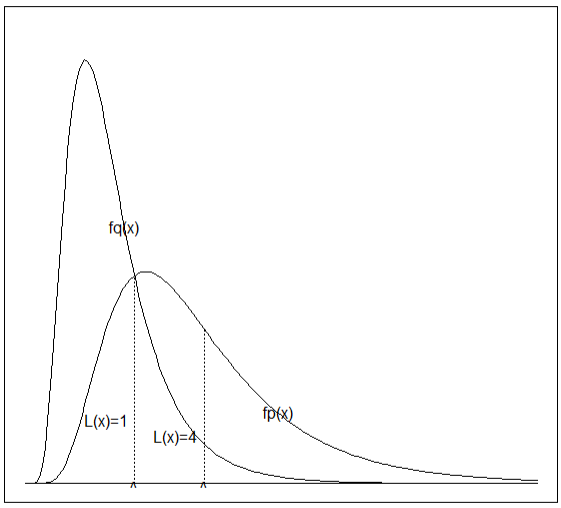
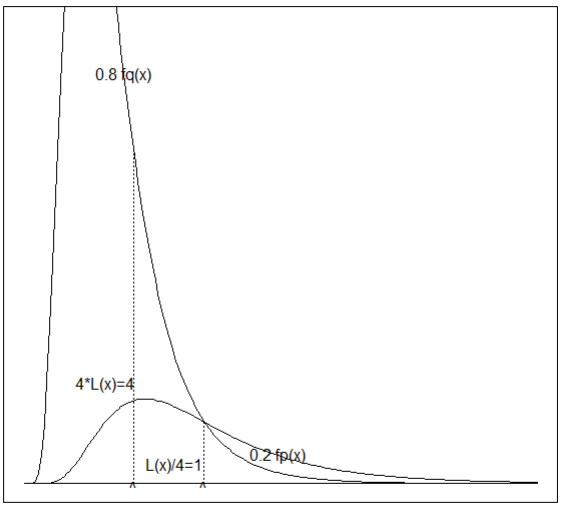


Figure 6: Unequal Areas



If one knows the prevalence[6] of the treatment, the calculation is a simple one. Assume prevalence is about 0.3, and that the data set has 1000 items. The number of items that will correctly be classed as treatment (hits) is $A = 300$ x sensitivity, and the number erroneously classed as treatment (fails) will be $B = 600$ x (1 - specificity). If one assumes that the threshold is 20 as in Figure (7), with sensitivity = 0.57 and 1-specificity = 0.14, then $A = 171$ and $B = 84$ are the number of items that will be classed as treatment. Hence, the probability that a randomly selected item classed as a treatment will actually be a treatment is $A/(A + B) = 171/(171 + 84) = 0.67$. The software marks the axis showing this probability as the "Bayesian Probability estimate."

This is the Bayesian argument for estimating a prior probability. It is quite proper in this case, and provides a valuable estimate of how the classifier threshold will perform in practice. Of course, with anything statistical, one needs a confidence statement. The Bayesian plot in Figure(8) shows how this is done. The confidence interval on the optimum point ranges from 0.22 to 0.74, which is pretty wide, but then this is just illustrative data.

The Bayesian plot shows the result of the above Bayesian calculation for all thresholds. There are many names for the plots associated with ROC: one of the alternate names for a Bayesian plot is "precision-recall."

Precision and recall are terms used in information retrieval. Recall is defined as the number of relevant documents retrieved by a search divided by the total number of retrieved documents, while precision is defined as the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

[6]The value of prevalence does not have to be very precise, since the calculations are often robust to substantial error.

Figure 7: Optimum point, prevalence = 20%

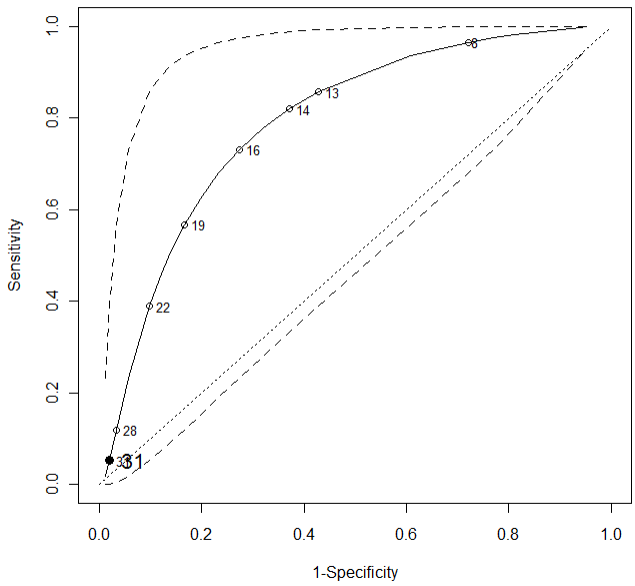
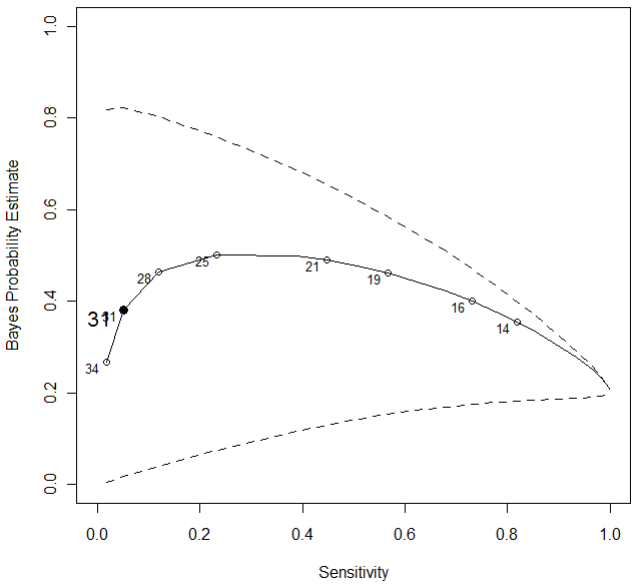


Figure 8: A Bayesian Plot, prevalence = 20%



In other words "recall" is sensitivity and "precision" is the Bayesian estimate. The only difference is that precision does not break out the prevalence probability as an explicit factor, dealing as it does with the total numbers of documents. Precision is nonetheless a probability. It is the probability that a retrieved document will be "relevant."

Another name for the Bayesian estimate is "positive predicted value," which is sometimes referenced in the Medial literature. It is, alas, not used as frequently as it should be. One reason for this is perhaps the difficulty of estimating prevalence; whereas an investigator can derive statistics from the data that they collect for their study and from this draw a ROC curve, the prevalence is an external item that must be sought out and incorporated.

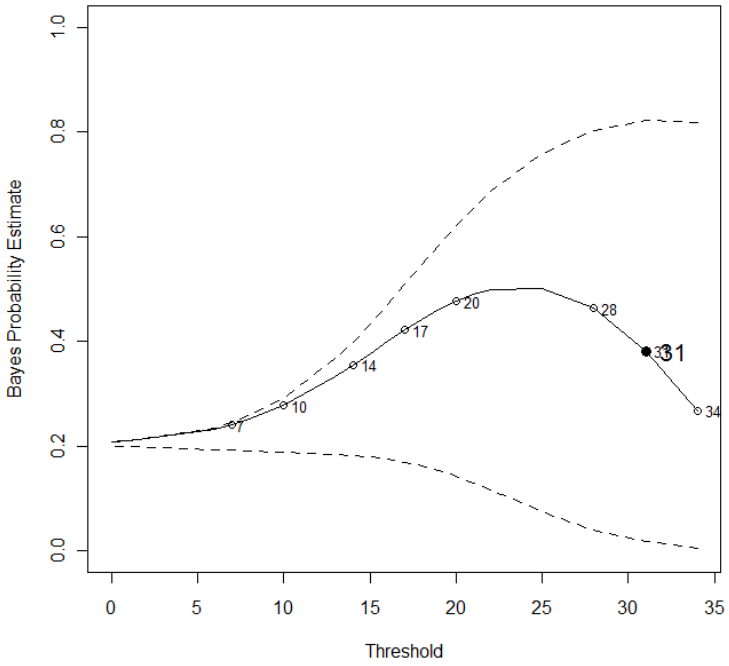
It is useful to plot the Bayesian probability estimate against the thresholds, as is done in Figure (9). Each point on the curve gives an estimate of the probability that a threshold value will be exceeded. If the data represents a disease, then the estimate is an estimate of the occurrence of the disease at or above the threshold.

8 Multiple plots and the convex hull.

It is often useful to view multiple classifiers on the same plot. Figure (10) shows a plot with two classifiers labeled "One" and "Two." Such plots are interesting for all plot types, not just ROC.

If one traces the outline of the two curves in Figure (10), it may be seen that the

Figure 9: Probabiity Threshold plot, prevalence = 20%



trace along the upper boundary dips in two places. If one considers the convex closure of these curves, with is the set of all points obtainable by convexly blending points on the curves, one sees that these dips vanish. Figure (11) shows the result, and a few marked points on the curve to boot. A point on this convex closure can actually be achieved by using a randomizing device with the classifiers involved. For example, the point at (0.6, 0.95) marked on the curve can be obtained by randomly choosing the classifier One, with frequency 0.22 and the classifier Two with frequency 0.78. The threshold values for these two classifiers are different. For One, the threshold is 90, for Two the threshold is 147. See [Fawcett, 2006] for a discussion of the use of the Convex Hull: note, such blending makes sense only for ROC curves.

The program outputs the values cited above in the following form. In this case, the optimum point is also output. The optimum point is always a 50-50 blend of two neighboring points because the points on the convex hull are discrete, and the best that can be done is to find two points on the convex hull that bracket the optimum.

```
$optimum
            alpha 1-alpha 1-Specificity Sensitivity Left Threshold Right Threshold
{Two | One}   0.5     0.5          0.15        0.53             67              69

$MarkedPts
            alpha 1-alpha 1-Specificity Sensitivity Left Threshold Right Threshold
{One | Two}  0.22    0.78           0.6        0.95             90             147
```

Figure 10: Multiple Plot

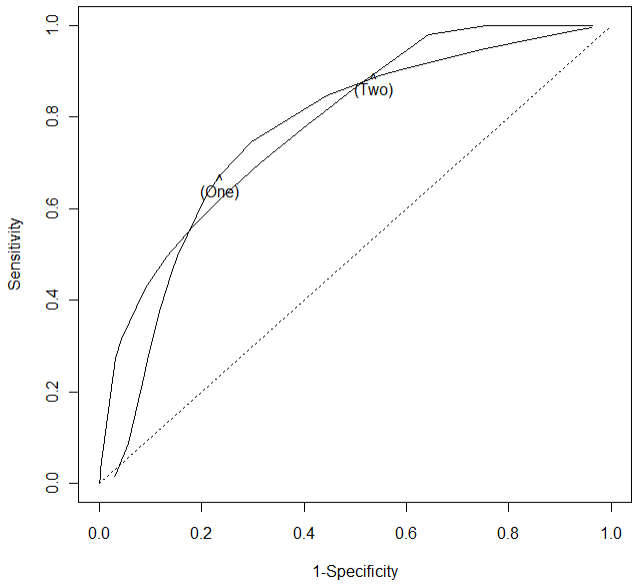
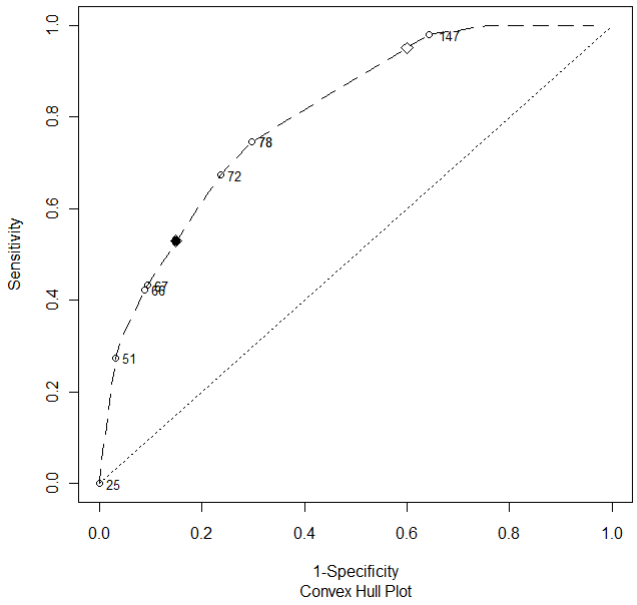


Figure 11: Convex hull plot with points and optimum point for prevalence 0.5



9 Case Study

Increased levels of prostrate specific antigen (PSA) can indicate the presence of cancer. Such an increase can also be due to other things such as benign prostatic hyperplasia (BPH) or prostatis. PSA tests are inexpensive and easily obtained during a routine blood workup. The problem is to make a judgment about the presence or absence of cancer from such a test. Barry [2001] gives a summary of the testing situation as of 2001, which does not seem to have changed much in the last few years.

Since prostate specific antigen has come to the notice of the medical community in the late 1970's, there has grown up a considerable literature discussing its applicability to the detection of prostate cancer. Much of this literature relies on statistical methods which have been of great help in clarifying a complex problem involving research with humans and their variabilities. ROC, rightly or wrongly, has played a prominent role in this literature.

The PSA test is a vexed subject, because it produces far too many failed tests – the "false positive" rate is too high. Each failed test can require an expensive, painful, and worrisome follow up, and surgeons being what there are, too many unnecessary operations. To illustrate the problem, consider Figure (12), which shows the ROC curve for a study by [Gann et al., 1995] when the prevalence is 30%. The usual threshold for the PSA test is 4 gm/mL, which was adopted as a standard by the Hypritech Tandem-R PSA test in February 1986[7], and has remained so ever since.

[7]The reference range of less than 4 ng/mL was based on a study that found 99% of 472 apparently

As may be seen from this figure, the choice is not unreasonable – not perfect, but something that can be lived with. Unfortunately, the actual prevalence is not 30% but something close to 3%, which means that the ROC curve is describing a situation in which detecting an individual with prostrate cancer is weighed ten times more than falsely classifying a cancer free individual as one who has cancer. There is nothing wrong with this, but nowhere in the literature does this weighing appear. In fact most of the ROC graphs in the literature use a prevalence of 50%.

Figure (13) shows a Bayes plot using the correct prevalence, and from this it is immediately clear that the false positive rate, which is one minus the Bayesian estimate is in the neighborhood of 80% for a PSA of 4 mg/mL. The PSA test has receive a bad press because of the obvious fact, born out by many unhappy false positive individuals, that the test is much more likely to be wrong then right.

The difficulties are further compounded by the fact that the optimum point is about 30 gm/mL, which produces reduced sensitivity. What good is a test that is usually right, but which misses 80% of the cancer cases? There is no apparent solution to this quandary.

In these figures, the optimum points are shown, 3.85 and 30.1, as well as two reference values, 4 and 10, which are of special interest. The 4 (the units are gm/mL), as mentioned, is the standard upper limit of the reference range. An individual who's PSA exceed this can expect a recommendation from his physician for more work and probably a biopsy. The 10 is another oft-cited limit beyond which serious problems are forecast.

The [Gann et al., 1995] data was taken from the Physicians Health Study, which is an ongoing double-blind trial of beta carotene among 22071 US male physicians aged 40 to 82 years in 1982. The sampling was random and is representative of the population. The treatment sample were those who had received a diagnosis of cancer, and the controls were selected to match them as closely as possible.

To illustrate the difficulties inherent in the studies of the PSA test, I will cite another data set from [Morgan et al., 1996]. Figure (14), and Figure (15) show results from this study. The data was obtained from admissions at the Walter Reed Army Medical Center. The sample does not appear to be a random sample from a specified population, and the diagnosis of cancer was by digital rectal examination and biopsy. I certainly will not say that this data is erroneous, but it clearly does not apply to the same population as the [Gann et al., 1995] data, which is in general a problem with all the studies in this field.

The fact that the two studies produce different results for individuals with cancer is not too surprising since the methods use in making the cancer determination are different; however, one would expect that the cancer free individuals would be similar in the two studies. This is not the case as may be seen in Figure (16), and the following output from the program, where there is a statistically significant difference for the Mann-Whitney test. It is clear that the two samples were drawn from different populations.

healthy men had a total PSA level below 4 ng/mL. This has since been substantially verified by [Oesterling et al., 1993], although it does vary with age.

```
$AUC
                AUC p-value 0.025% 0.975%
Mann-Whitney 0.645   2e-04  0.569  0.703
```

That this is not always the case, may be observed by comparing the cancer free individuals from the [Gann et al., 1995] study with those from a study by [Oesterling et al., 1995]. This study drew a random sample from the population of Olmsted County Minnesota. Figure (17) shows a ROC curve and the non-significant Mann-Whitney test is as follows:

```
$AUC
                AUC p-value 0.025% 0.975%
Mann-Whitney 0.527   0.252  0.444  0.599
```

For the most part, the data in the various studies are drawn from unstated populations and the methods by which the cancer is determined differ from study to study; and yet all purport to offer conclusions for the general population. Needless to say, confidence intervals and other statistics based on probability have little meaning in such situations.

Figure 12: Gann PSA ROC for prevalence=30%

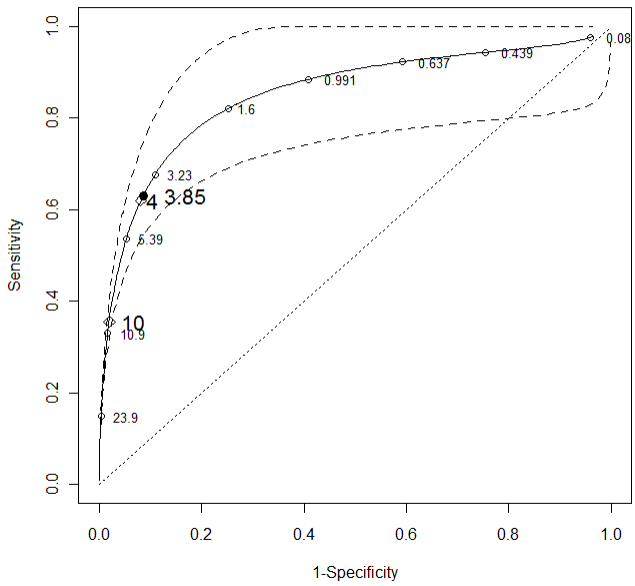
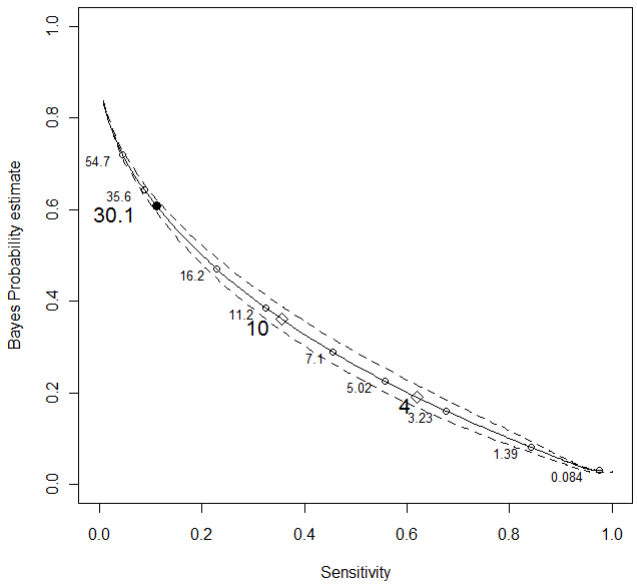


Figure 13: Gann PSA Bayesian for prevalence=3%



Because of the unsatisfactory nature of the PSA test as it is currently used, there has been a fervid search for alternatives and ways to improve it's performance. Most such work reports minor improvements in sensitivity and specificity, which may or may not be the result of sampling and measurement fluctuations. However, some work has been done on alternative forms of the antigen. In particular, the "free" PSA. Most PSA is

Figure 14: Morgan PSA ROC for prevalence=30%

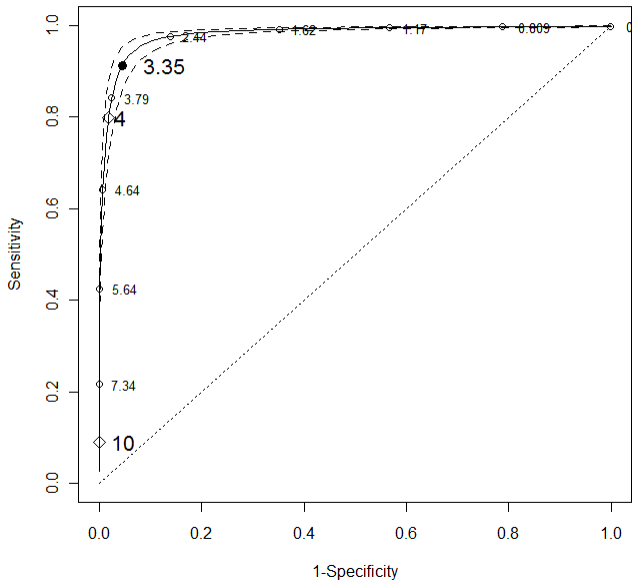
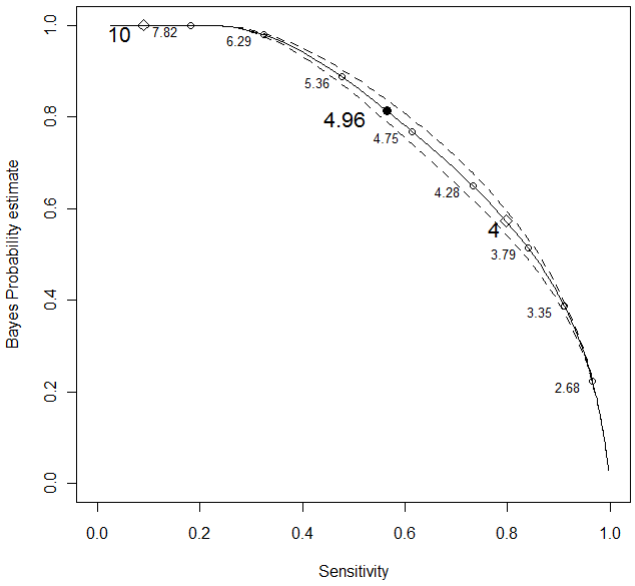


Figure 15: Morgan PSA Bayesian for prevalence=3%



complexed with other things, but a small fraction is not, and it has been observed that the ratio of this free form to the total PSA may be an independent predictor of cancer.

A study by [Catalona et al., 1998], the principal investigative group for this viewpoint seems to provide evidence for this viewpoint. This study selected individuals who for one reason or another had had biopsies and sought to determine whether or not free PSA could be used to winnow out those with cancer from those who were cancer free. The sample was not obviously a random sample, as described in the paper, but the implication is that it should be considered representative of those who were otherwise candidates for biopsy. The statement was made that of those who were candidates, some 25% would be found to have cancer in this group. This 0.25 prevalence was obtained from other data not included in the paper. The principal claim of the paper is that by using a recommended cut-off for free PSA, some 20% of unnecessary biopsies would be avoided. The implication being that the "false positive" rate would be reduced. Let us examine these claims.

I do not have access to the original data, but there is sufficient information in the paper to construct an equivalent set, which agrees well enough with the charts and graphs in the paper to support my conclusions. Figure (18) shows the ROC graph, and the 25% free PSA cut-off recommended in this paper. As may be seen the 1-Specificity value is at approximately 0.80, which means as the paper claims that by using this cut-off some 20% of those individuals who are cancer free will be able to avoid a biopsy. Figure (reffig:CatalonaBayes) shows the Bayes plot taking into account the 0.25 prevalence that is claimed by the authors. This figure shows that the probability of detecting cancer is in the neighborhood of 0.25, or that the false positive rate is about 75%.

Figure 16: Gann vs Morgan cancer free

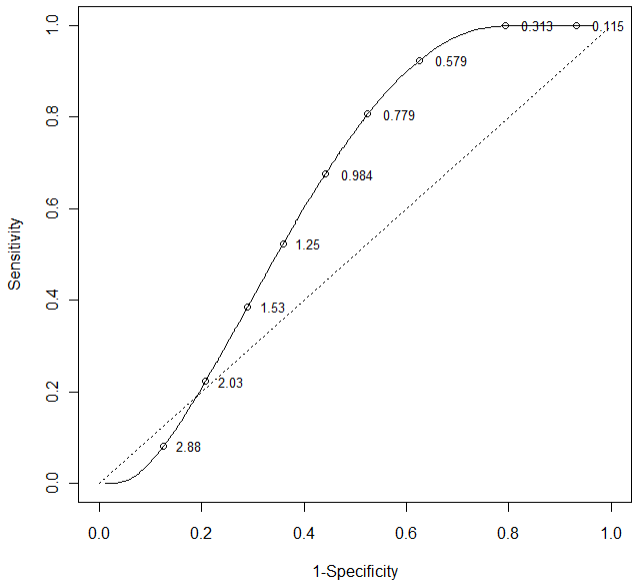
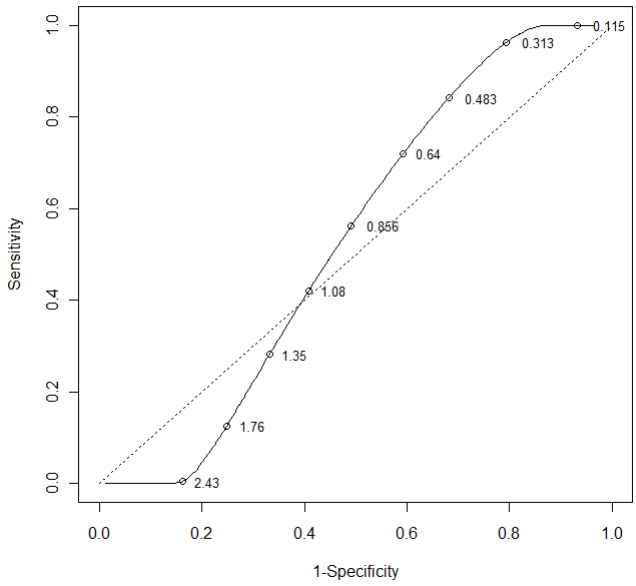


Figure 17: Gann vs Oesterling cancer free



The false positive rate has not changed. It was 75% before the test, an it remains at 75% after the test. However, some of the less likely individuals have been spared a biopsy. These individuals still have a measurable probability of cancer.

A more rational approach to recognize that missing a cancer is more important than misclassifying a cancer free individual. In this case, the value of detecting a cancer is apparently some seven times that of misclassification. Figures (20) and (21) illustrate this point, and produce an optimum point of about 24.

10 ROC limitations

There is a tendency to use ROC when other forms of analyses are more appropriate. ROC does not work well when the variances of the two distributions are different. It is after all nothing but a graphical representation of the Wilcox-Mann-Whitney test which is most sensitive to a shift in location. [Morgan et al., 1996] studied PSA levels in black and white men, but failed to observe that the variances of the two distributions are different: ROC curves for cancer verses no-cancer were drawn within each group, but not between the groups. Figure (22) shows a ROC plot contrasting the two groups for cancer free individuals: the Mann-Whitney test produced the non-significant result below. This is highly misleading, since the two distributions are different.

Figure 18: Catalona ROC with prevalence = 25%

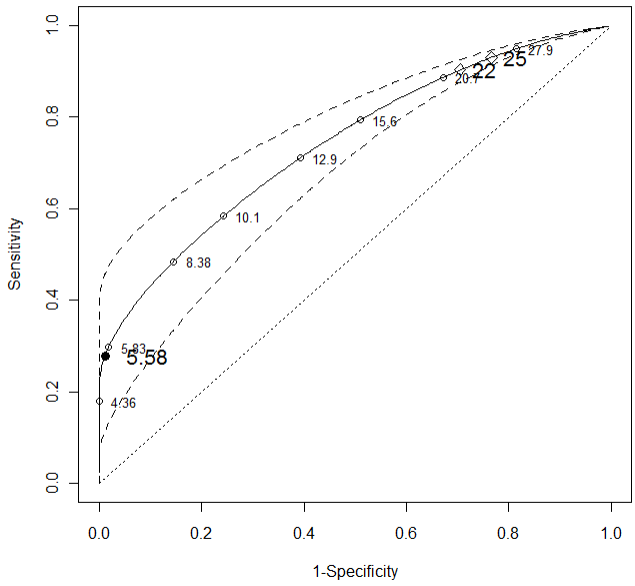
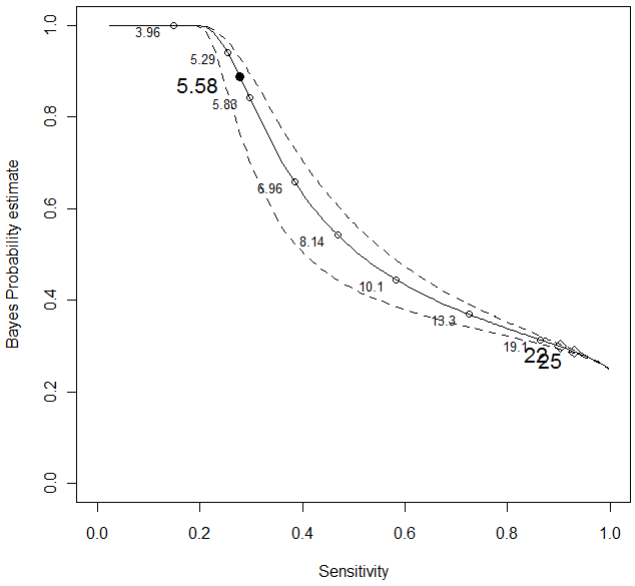


Figure 19: Catalona Bayes with prevalence = 25%




```
$AUC
                AUC p-value 0.025% 0.975%
Mann-Whitney 0.527   0.252  0.444  0.599
```

This sort of "bent" plot is typical of data in which the variances are different. The data is in fact well described by a log-normal distribution, and the variances on a log scale are 0.335 for white and 1.06 for black. With 410 and 1672 degrees of freedom, the p-value for the F-test is zero to several decimal places. In other words, the ROC AUC analysis "ain't tellin the truth."

One of the problems in the PSA literature is the failure to make use of the well known fact that PSA values follow a log-normal distribution. Much confusion would be avoided, were investigators to concentrate on the log-normal parameters instead of rushing hell-bent to a ROC analysis.

References

D. Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graphl. *Jour. Math. Psychol.*, 12:387–415, 1975.

M.J Barry. Prostate-specific-antigen testing for early diagnosis of prostate cancer. *NE J. Med*, 344(18):1273–1377, 2001.

D.F. Bauer. Constructing confidence sets using rank statistics. *J. Am.Stat.Assoc.*, 67 (339):687–690, 1972.

Figure 20: Catalona ROC with prevalence = 25% and weight 7

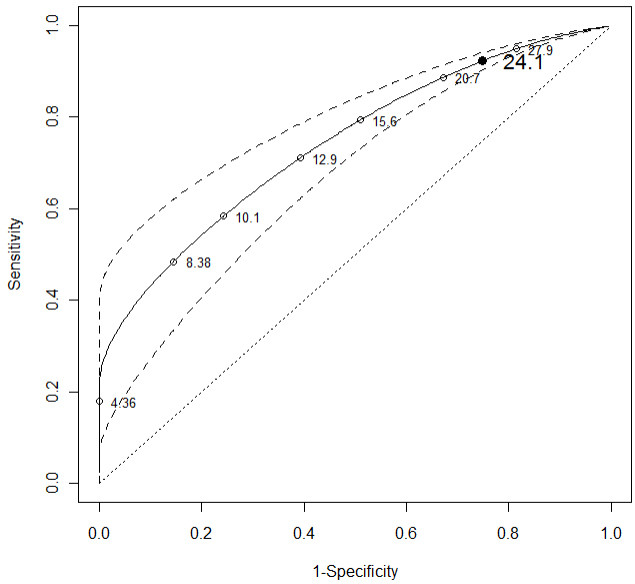
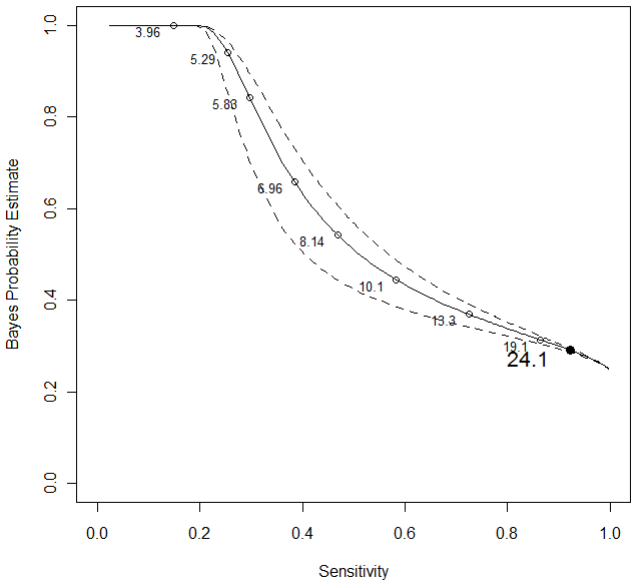


Figure 21: Catalona Bayes with prevalence = 25% and weight 7



W.J Catalona, A.W. Partin, K.M. Slawin, M.K. Brawer, et al. Use of the percentage of free prostate-specific antigen to enhance differentiation of prostate cancer from benign prostatic disease. *JAMA*, 279(19):1542–1547, 1998.

Corinna Cortes and Mehryar Mohri. Confidence intervals for the area under the roc curve. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 305–312. MIT Press, Cambridge, MA, 2005.

E. DeLong, D. DeLong, and D. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometircs*, 44:837–845, 1988.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

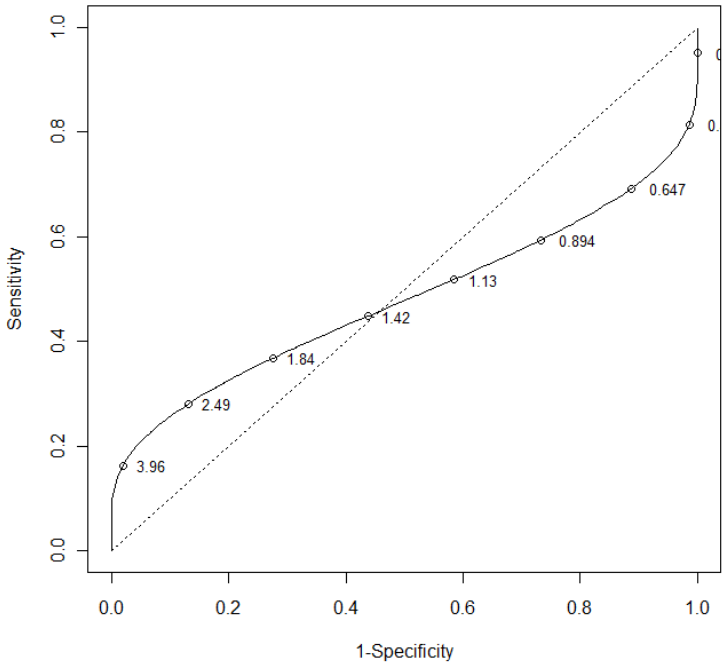
Gustav Theodor Fechner. *Elements of Psychophysics*. Holt, Rinehart and Winston, Howes, D.H. and Boring, E.G editors, New York, N.Y., 1860. Translation of Elemente de Psychophysic, Vol1 by Helmut E. Adler.

R.A. Fisher. On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1(4):3–32, 1921.

P.H. Gann, H. Hennekens, and M.J. Stamfer. A prospective evaluation of plasma prostate-specific antigen for detection of prosatic cancer. *JAMA*, 273(4):289–294, 1995.

D.M. Green and J.A. Swets. *Signal detection theory and psychophysics*. Wiley, New York, 1966. Reprint with corrections 1974.

Figure 22: Morgan black vs white – different variances



N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions, Vol I.* Wiley, New York, second edition, 1994.

A. Macskassy and F. Provost. Confidence bands for roc curves: Methods and an empirical study. *Proceedings of the First Workshop on ROC Analysis in AI. August 2004*, CeDER-PP-200407, 2004.

J. I. Marcum. *A Statistical Theory of Target Detection by Pulsed Radar.* RAND Corporation, Santa Monica, CA, 1947.

T.O. Morgan, S.T. Jacobsen, W.F. McCarthy, et al. Age-specific reference ranges for serum prostate-specific antigen in black men. *N. Engl. J. Med.*, 335(5):304–310, 1996.

G.E. Noether. *Elements of nonparametric statistics.* Wiley, New York, 1967.

J.E. Oesterling, S.J. Jacobsen, C.G. Chute, et al. Serum prostate-specific antigen in a community-based population of healthy men. *JAMA*, 270(7):860–864, 1993.

J.E. Oesterling, S.J. Jacobsen, G.G. Klee, et al. Free, complexed and total serum prostate specific antigen: the establishment of appropriate reference ranges for their concentrations and ratios. *J.Urol.*, 154(Sept.):1090–1095, 1995.

K. Pearson. On a novel method of regarding the association of. two variates classed solely in alternate categories. *Biom.*, Ser. I:1–35, 1904. Draper's Co. Res. Mem.

Charles Peirce and Joseph Jastrow. On small differences of sensation. *Memoirs of the National Academy of Sciences for 1884*, 3:75–83, 1884.

W.W. Peterson, T.G. Birdsall, and W.C. Fox. *The theory of signal detection.* Electronics Defense Group, University of Michigan, 1953. Technical Report no.13, also pub. Trans IRE Prof. Grp. 1954, 4(4). p171-212.

W.P. Jr. Tanner and J. Swets. *A new theory of visual detection.* Electronics Defense Group, University of Michigan, 1954. Technical Report no.19.

R.E. Wheeler. Quantile estimators of johnson curve parameters. *Biometrika*, 67(3):725–728, 1980.

K. Zou, W.J. Hall, and D.E. Shapiro. Smooth non parametric receiver operatingcharacteristic (roc) curves for continuous diagnostic tests. *Statistics in Medicine*, 16:2143–2156, 1997.