Measuring Success in Prediction

BST 226

Statistical Methods for Bioinformatics

David M. Rocke

Binary Classification

- Suppose we have two groups for which each case is a member of one or the other, and that we know the correct classification ("truth"). We will call the two groups Disease and Healthy
- Suppose we have a prediction method that produces a single numerical value, and that small values of that number suggest membership in the Healthy group and large values suggest membership in the Disease group.
- How can we measure the success of the prediction method?
- First, consider the case when we have a cutoff that defines which group is predicted.

	Disease	Healthy	Total
Predict Disease	A (True Positive)	B (False Positive)	A+B
Predict Healthy	C (False Negative)	D (True Negative)	C+D
Total	A+C	B+D	A+B+C+D

- A: True Positive (TP), hit
- D: True negative (TN), correct rejection
- B: False positive (FP), false alarm, Type I error
- C: False negative (FN), miss, Type II error

	Disease	Healthy	Total
Predict Disease	A (True Positive)	B (False Positive)	A+B
Predict Healthy	C (False Negative)	D (True Negative)	C+D
Total	A+C (Positive)	B+D (Negative)	A+B+C+D

- Sensitivity, True Positive Rate (TPR), recall
  - TPR = TP/P = TP/(TP+FN) = A/(A+C)
  - Fraction of those with the Disease that are correctly predicted
- Specificity (SPC), True Negative Rate
  - SPC = TN/N = TN/(TN+FP) = D/(B+D)
  - Fraction of those Healthy who are correctly predicted
- Precision, Positive Predictive Value (PPV)
  - PPV = TP/(TP+FP) = A/(A+B)
  - Fraction of those predicted to have the Disease who do have it
- Negative Predictive value (NPV)
  - NPV = TN/(TN+FN) = D/(C+D)
  - Fraction of those predicted to be healthy who are healthy
- Fall-out or False Positive Rate (FPR)
  - FPR = FP/N = FP/(FP+TN) = 1 – SPC
  - Fraction of those healthy who are predicted to have the disease
- False Discovery Rate (FDR)
  - FDR = FP/(TP+FP) = 1 – PPV
  - Fraction of those predicted to have the disease who are healthy
- Accuracy (ACC)
  - ACC = (TP+TN)/(P+N)

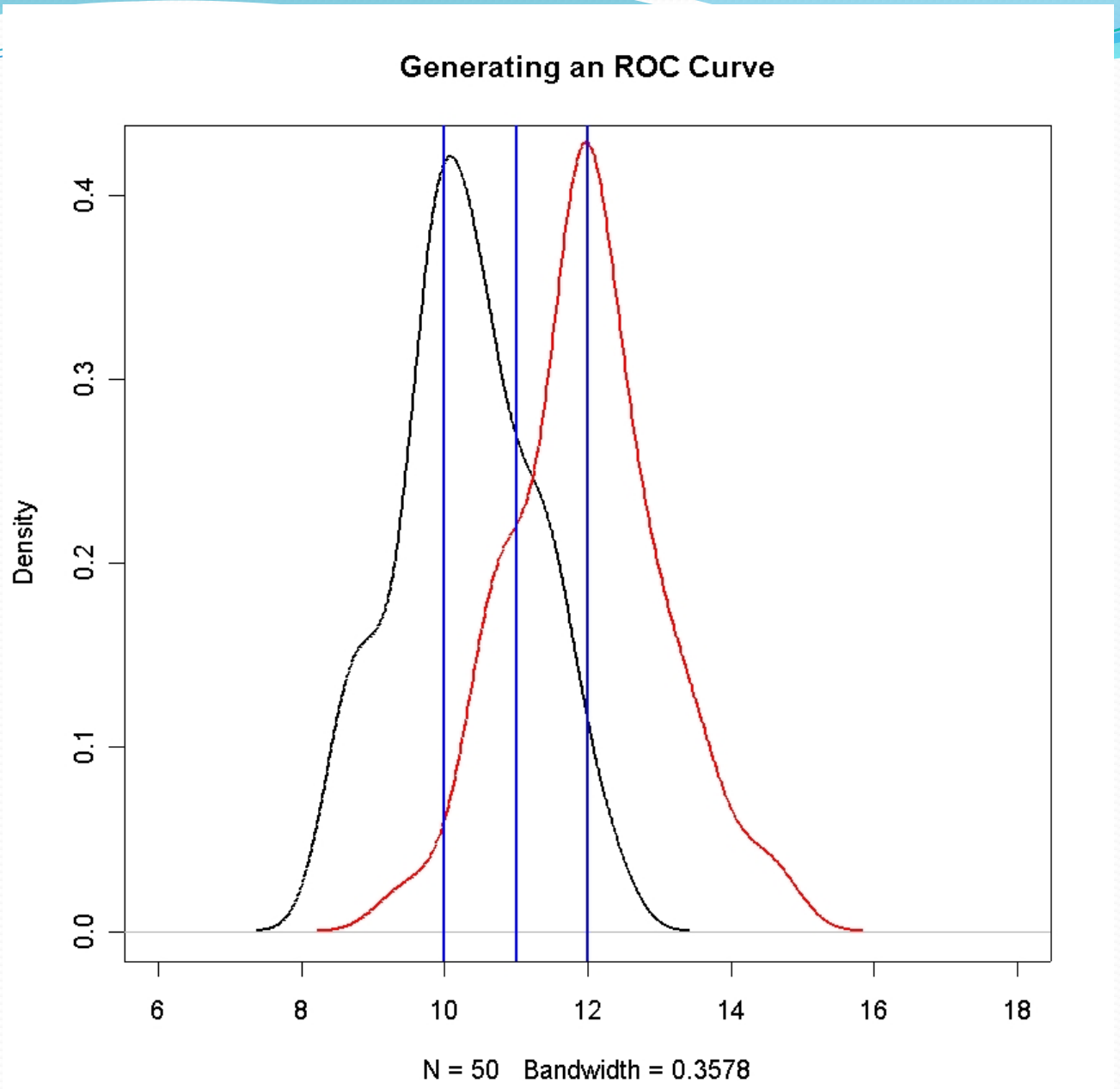
Dependence on Population

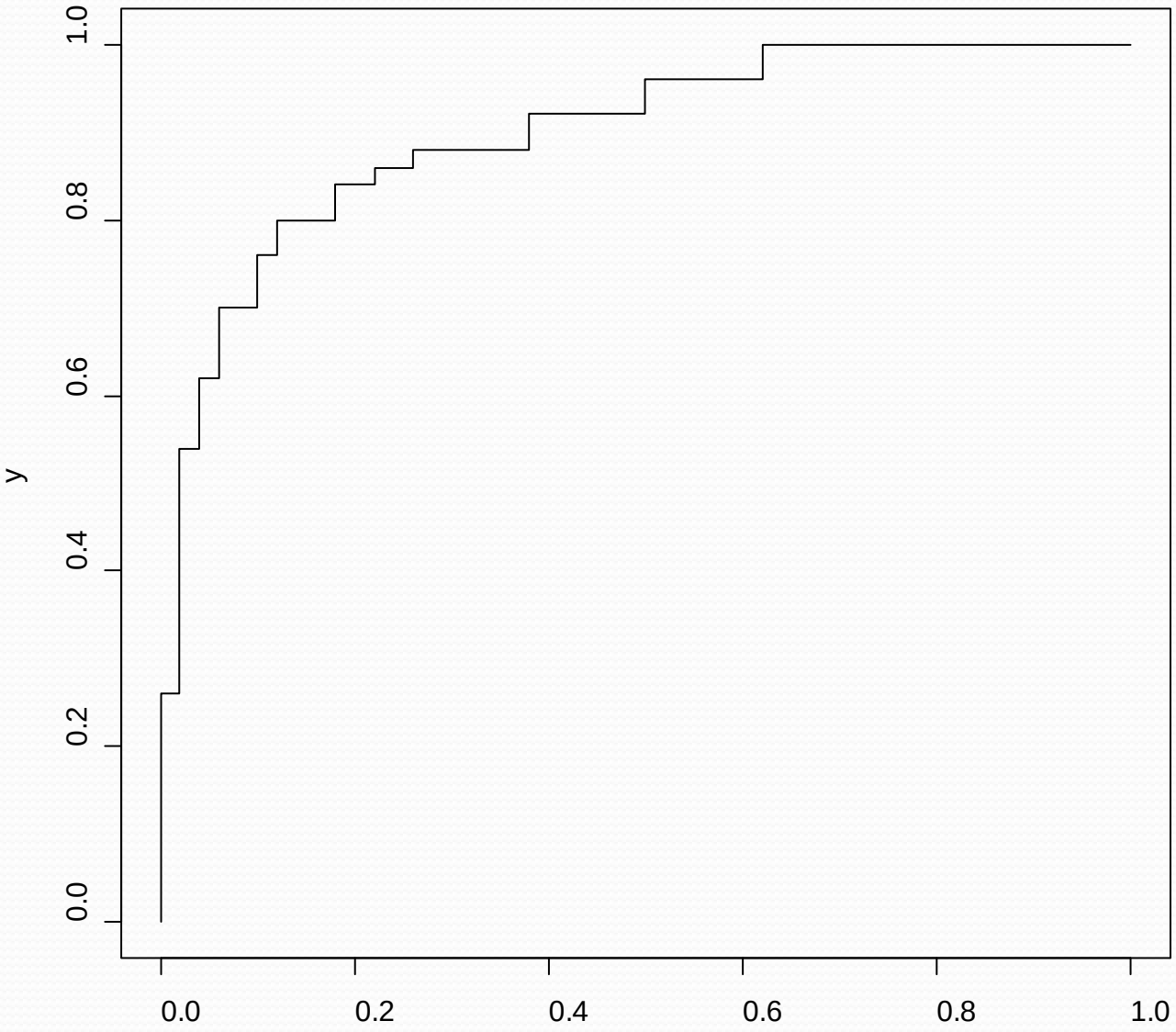
- Sensitivity and Specificity depend only on the test, not on the composition of the population, other figures are dependent
- Sensitivity = fraction of patients with the disease who are predicted to have the disease (p = 0.98).
- Specificity = fraction of patients who are healthy that are classified as healthy (q = 0.99).
- If the population is 500 Disease and 500 healthy, then TP = 490, FN = 10, TN = 495, FP = 5 and
  PPV = 490/(490 + 5) = **0.9899**
- If the population is 100 Disease and 1000 healthy, then TP = 98, FN = 2, TN = 990, FP = 10 and
  PPV = 98/(98 + 10) = **0.9074**
- If the population is 100 Disease and 10,000 healthy, then TP = 98, FN = 2, TN = 9900, FP = 100 and
  PPV = 98/(98 + 100) = **0.4949**

ROC Curve (Receiver Operating Characteristic)

- If we pick a cutpoint t, we can assign any case with a predicted value $\leq t$ to Healthy and the others to Disease.
- For that value of t, we can compute the number correctly assigned to Disease and the number incorrectly assigned to Disease (true positives and false positives).
- For t small enough, all will be assigned to Disease and for t large enough all will be assigned to Healthy.
- The ROC curve is a plot of true positive rate vs. false positive rate.
- If everyone is classified positive (t = 0), then
  $TPR = TP/(TP+FN) = FP/(FP + 0) = 1$
  $FPR = FP/(FP + TN) = FP/(FP + 0) = 1$
- If everyone is classified negative (t = 1), then
  $TPR = TP/(TP+FN) = 0/(0 + FN) = 0$
  $FPR = FP/(FP + TN) = 0/(0 + TN) = 0$

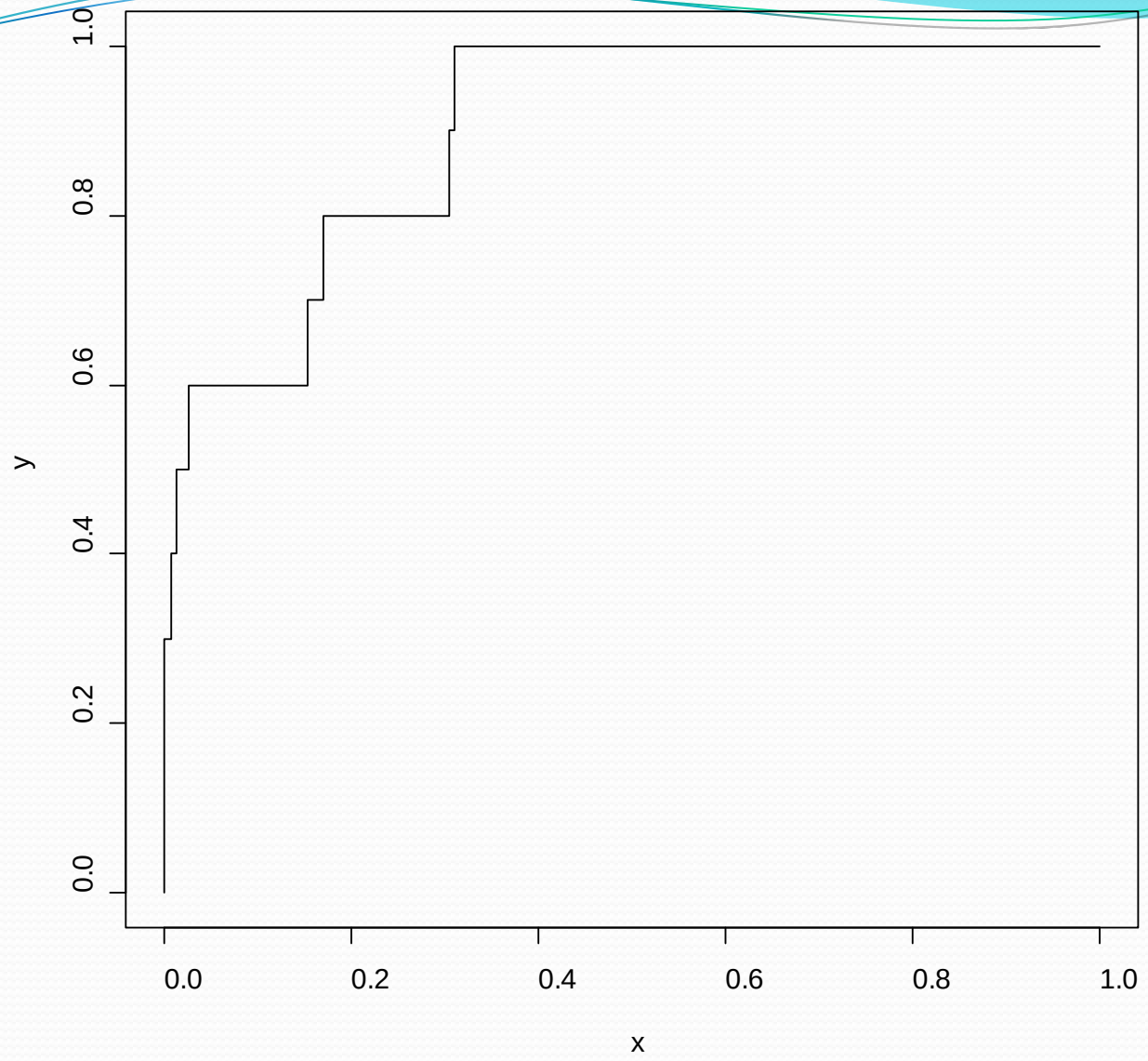
```
truth <- rep(0:1,each=50)
pred <- c(rnorm(50,10,1),rnorm(50,12,1))
library(ROC)
roc.data <- rocdemo.sca(truth,pred)

plot1 <- function()
{
  nz <- sum(truth==0)
  n <- length(truth)
  plot(density(pred[1:nz]),lwd=2,xlim=c(6,18),
    main="Generating an ROC Curve")
  lines(density(pred[(nz+1):n]),col=2,lwd=2)
  abline(v=10,col=4,lwd=2)
  abline(v=11,col=4,lwd=2)
  abline(v=12,col=4,lwd=2)
}
> plot1()
> plot(roc.data)
> AUC(roc.data)
[1] 0.8988
```


Generating an ROC Curve
Density
0.0
0.1
0.2
0.3
0.4
6
8
10
12
14
16
18
N = 50 Bandwidth = 0.3578


y
x
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

We now show the ROC curve for a rare outcome:

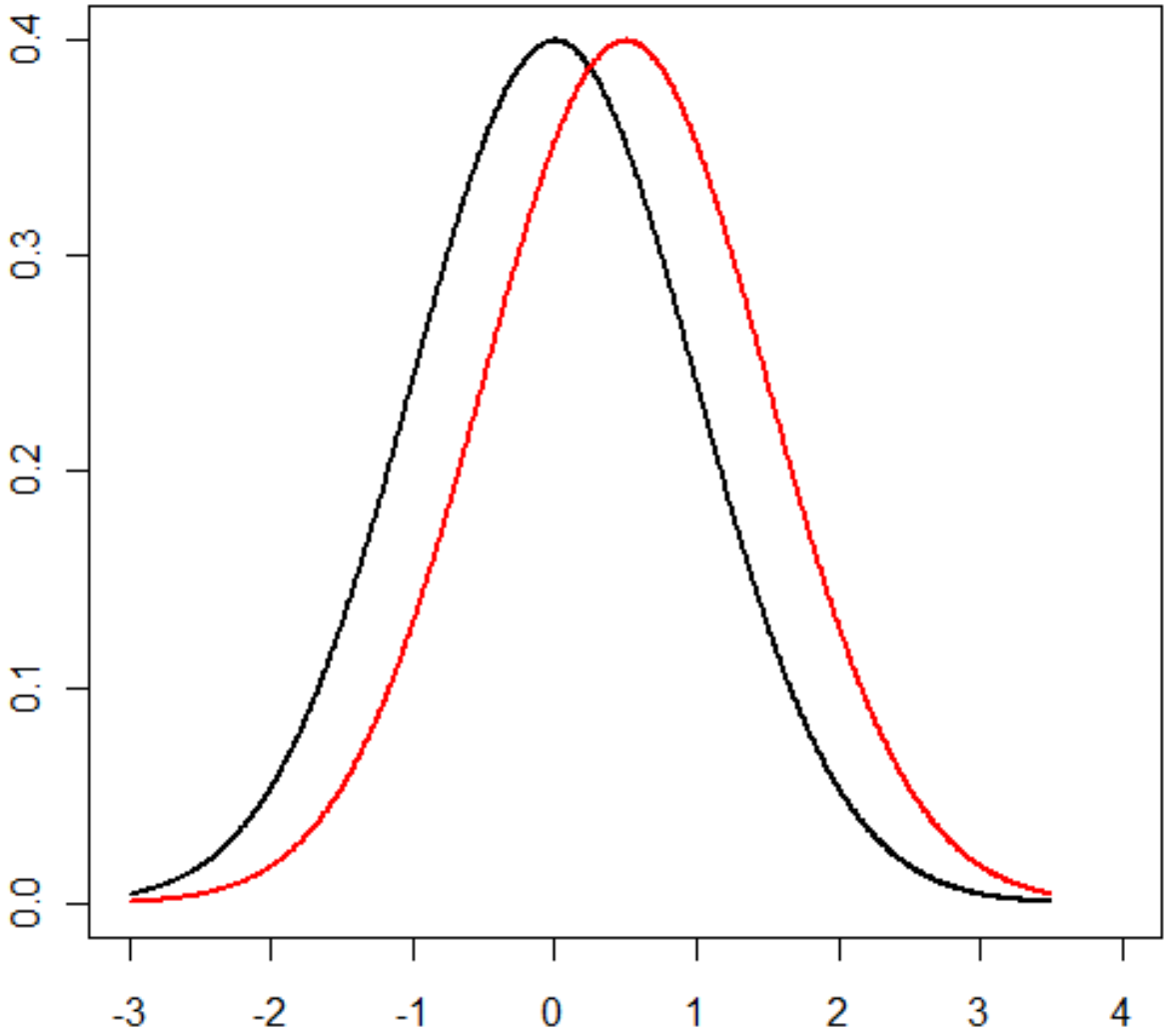
```
> truth <- rep(0:1,c(990,10))
> pred <- c(rnorm(990,10,1),rnorm(10,12,1))
> plot(rocdemo.sca(truth,pred))
> AUC(rocdemo.sca(truth,pred))
[1] 0.9011111
```



ROC Curve for Rare Outcome

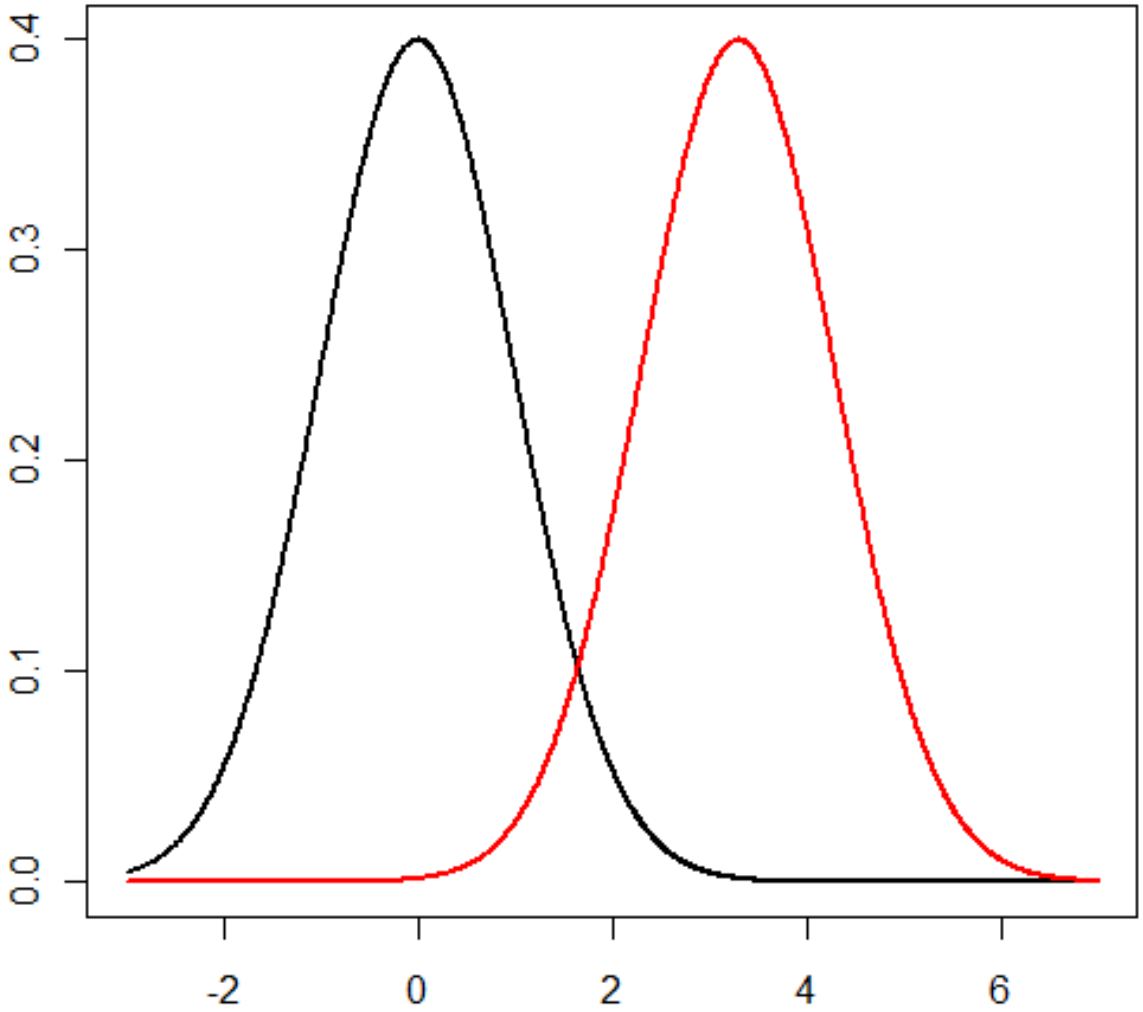
Statistical Significance and Classification Success

- It is easier for a variable to be statistically significant than for the classification using that variable to be highly accurate, measured, for example, by the ROC curve.
- Suppose we have 100 patients, 50 in each group (say disease and control).
- If the groups are separated by 0.5 times the within group standard deviation, then the p-value for the test of significance will be around 0.01 but the classification will only be 60% correct.


0.0
0.1
0.2
0.3
0.4
-3
-2
-1
0
1
2
3
4

Statistical Significance and Classification Success

- If the classification is to be correct 95% of the time, then the groups need to be separated by 3.3 times the within group standard deviation, and then the p-value for the test of significance will be around essentially 0.

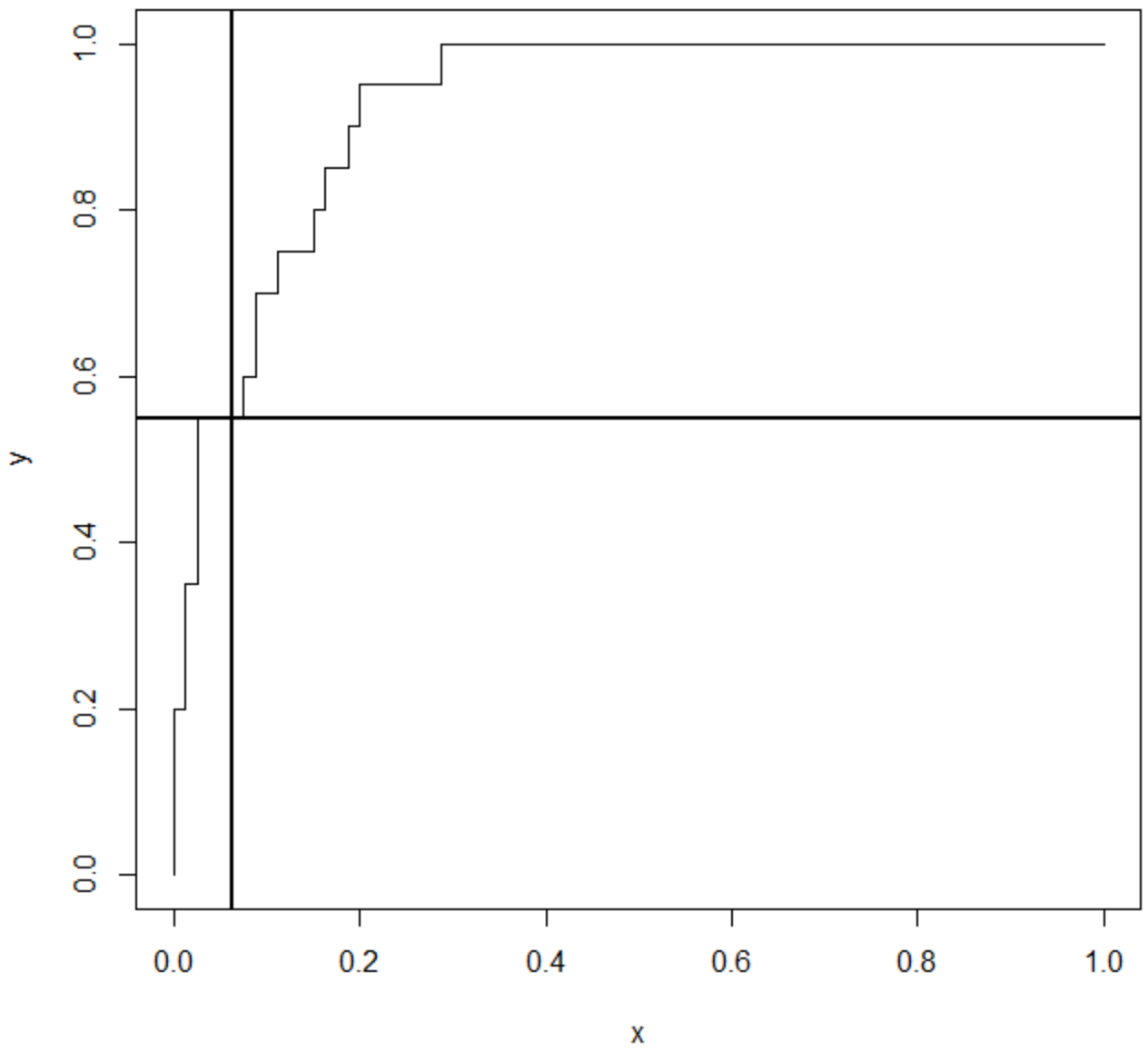

0.0
0.1
0.2
0.3
0.4
-2
0
2
4
6

```
> truth <- rep(0:1,c(80,20))
> summary(glm(truth~var1,family=binomial))

Call:
glm(formula = truth ~ var1, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.04601  -0.45586  -0.21127  -0.05413   2.11889

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.4727     0.6775  -5.125 2.97e-07 ***
var1          1.8202     0.4038   4.508 6.55e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 100.080  on 99  degrees of freedom
Residual deviance:  56.222  on 98  degrees of freedom
AIC: 60.222

Number of Fisher Scoring iterations: 6
```

```
> pred2 <- predict(glm(truth~var1,family=binomial),type="response")
> table(truth,pred2 > .5)

truth FALSE TRUE
    0    75    5
    1     9   11
```

TPR = 11/20 = 0.55
SPC = TNR = 75/80 = 0.9375
PPV = 11/16 = 0.6875
NPV = 75/84 = 0.8929
FPR = 5/80 = 0.0625
FDR = 5/16 = 0.3125

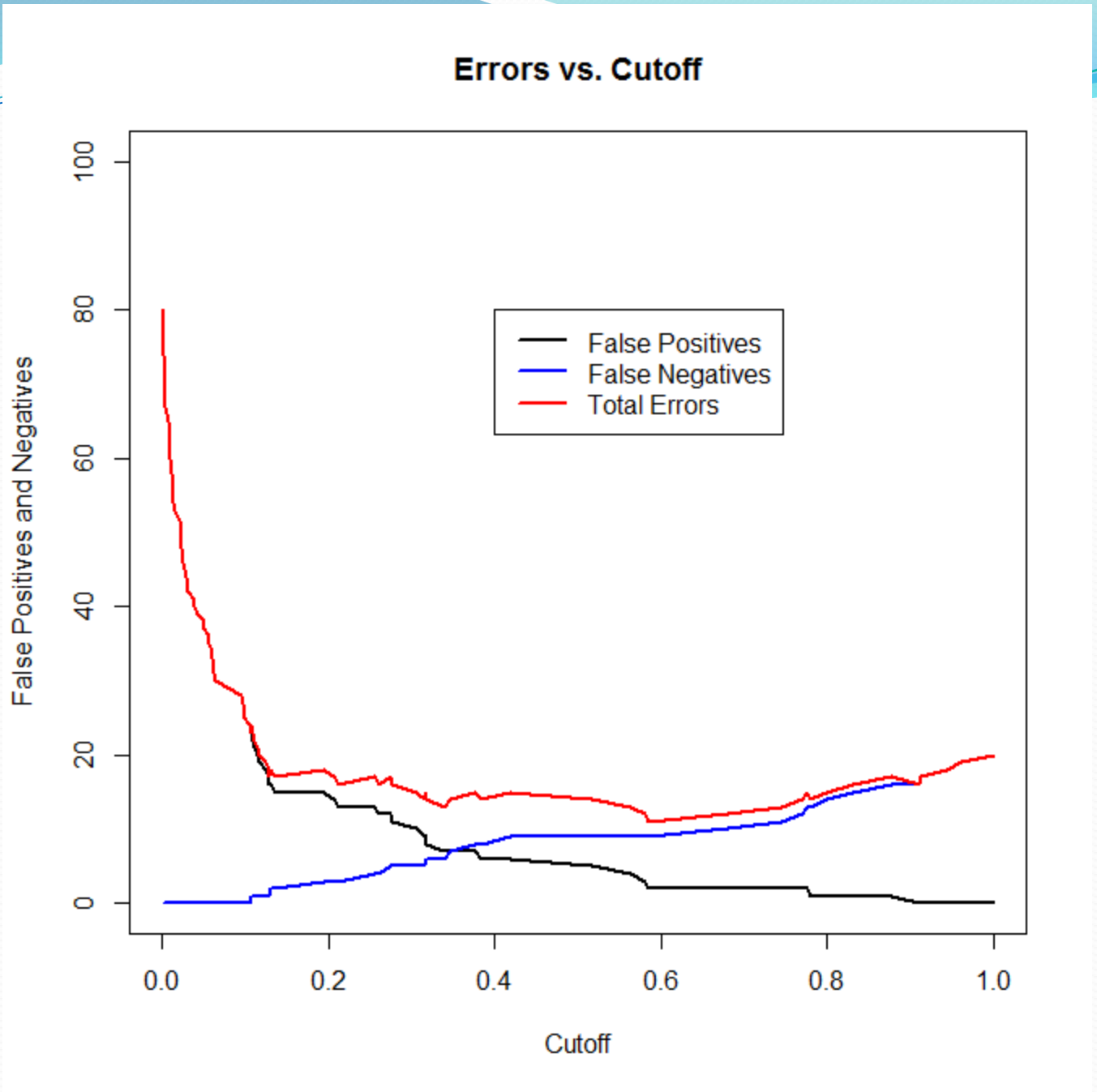
```
> source("http://bioconductor.org/biocLite.R")
> biocLite("ROC")
> library(ROC)
> plot(rocdemo.sca(truth,pred2))
> abline(v=0.0625,lwd=2)
> abline(h=0.55,lwd=2)
```

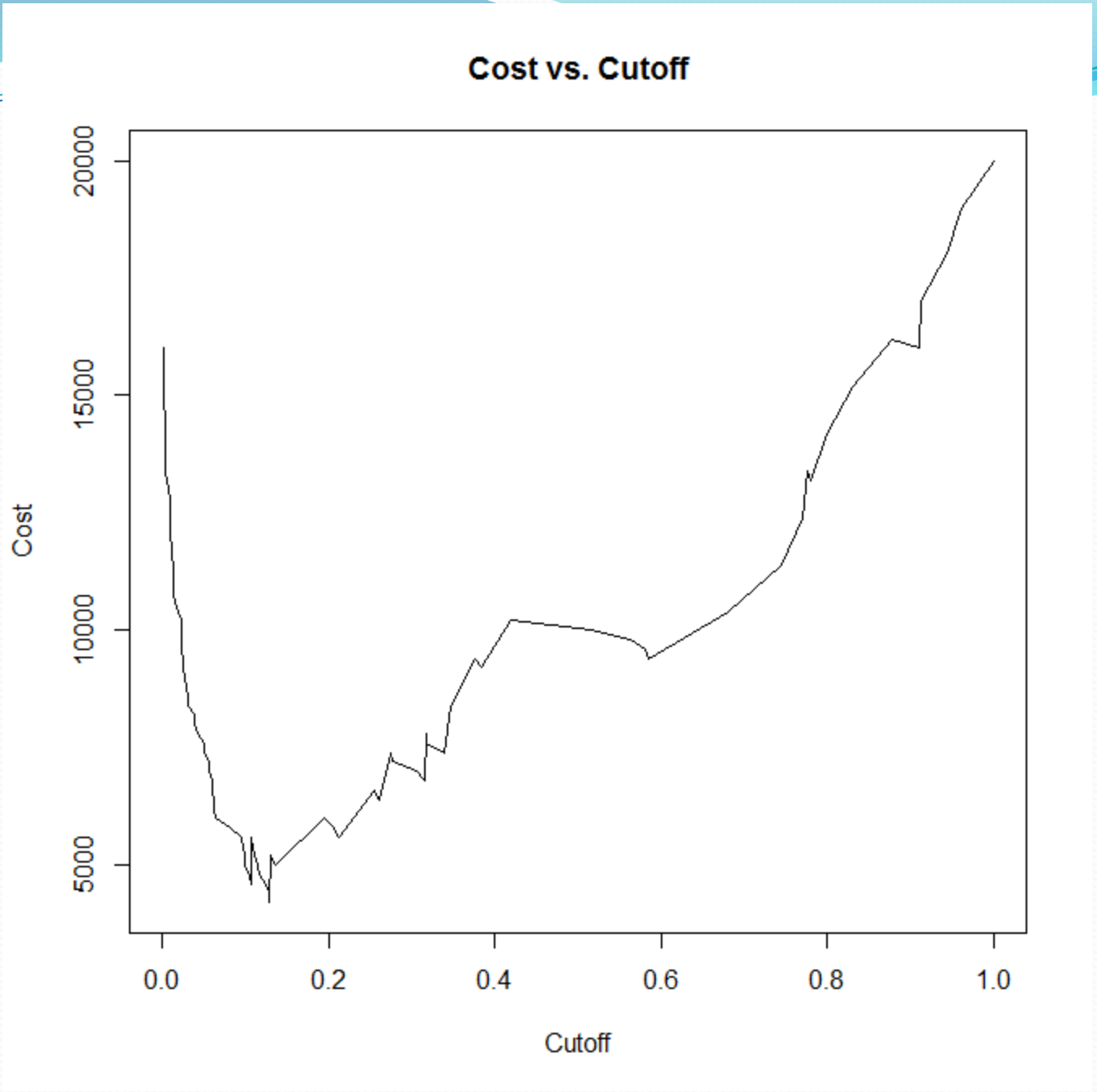

y
0.0
0.2
0.4
0.6
0.8
1.0
x
0.0
0.2
0.4
0.6
0.8
1.0

Choosing a Cutoff

- Suppose that missing a disease case has an implicit cost of $1000 and a false diagnosis of disease has an implicit cost of $200.
- Then the cost of the procedure is 1000×FN+200×FP.
- With a cut-off of 0.5, the estimated cost would be (1000)(9) + (200)(5) = $10,000 per 100 patients or $100 per patient.
- Let's compute the cost for different cutoffs.

```
diagcost <- function(truth,predq,costp,costn)
{
  n <- length(predq)
  names(predq) <- ""
  cutoffs <- c(sort(predq),1)
  fpvec <- rep(0,n+1)
  fnvec <- rep(0,n+1)
  costvec <- rep(0,n+1)
  for (i in 1:(n+1))
  {
    predb <- predq >= cutoffs[i]
    fp <- sum(predb &  !truth)
    fn <- sum(!predb & truth)
    cost <- fp*costp+fn*costn
    fpvec[i] <- fp
    fnvec[i] <- fn
    costvec[i] <- cost
  }
  return(data.frame(1:(n+1),cutoffs,fpvec,fnvec,costvec))
}
The least cost of $4200 (vs. $10,000) is at cutoff = 0.1286845279
with 1 false negative and 16 false positives

The cutoff of 0.5853639 minimizes the total errors with 2 false positives
and 9 false negatives (cost $9400)
```


Errors vs. Cutoff
False Positives and Negatives
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
Cutoff
False Positives
False Negatives
Total Errors


Cost vs. Cutoff
Cost
5000
10000
15000
20000
0.0
0.2
0.4
0.6
0.8
1.0
Cutoff