Measuring Success in Prediction

SPH 247

Statistical Analysis of Laboratory Data

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

	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

- 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)

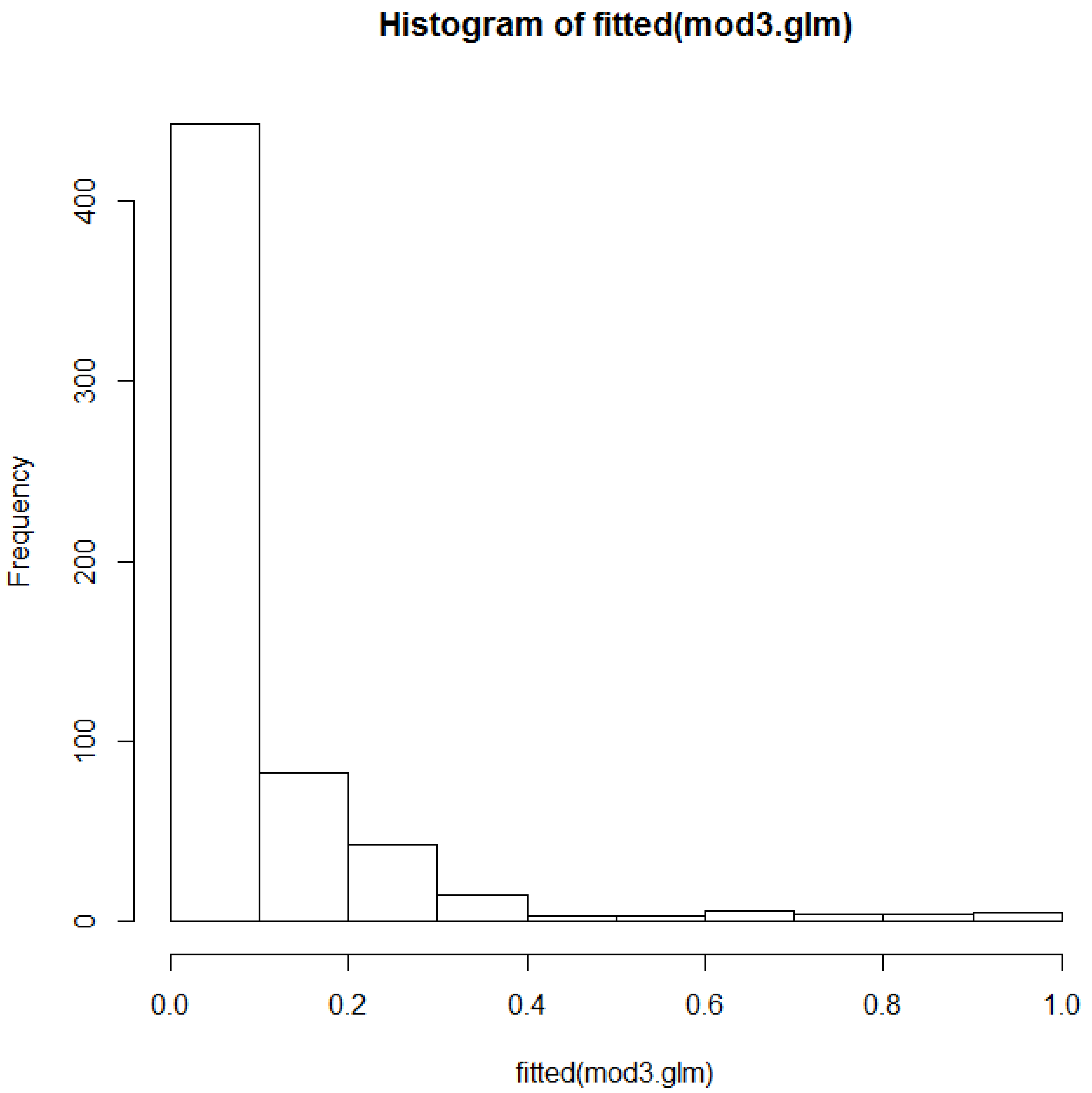
Dependence on Population

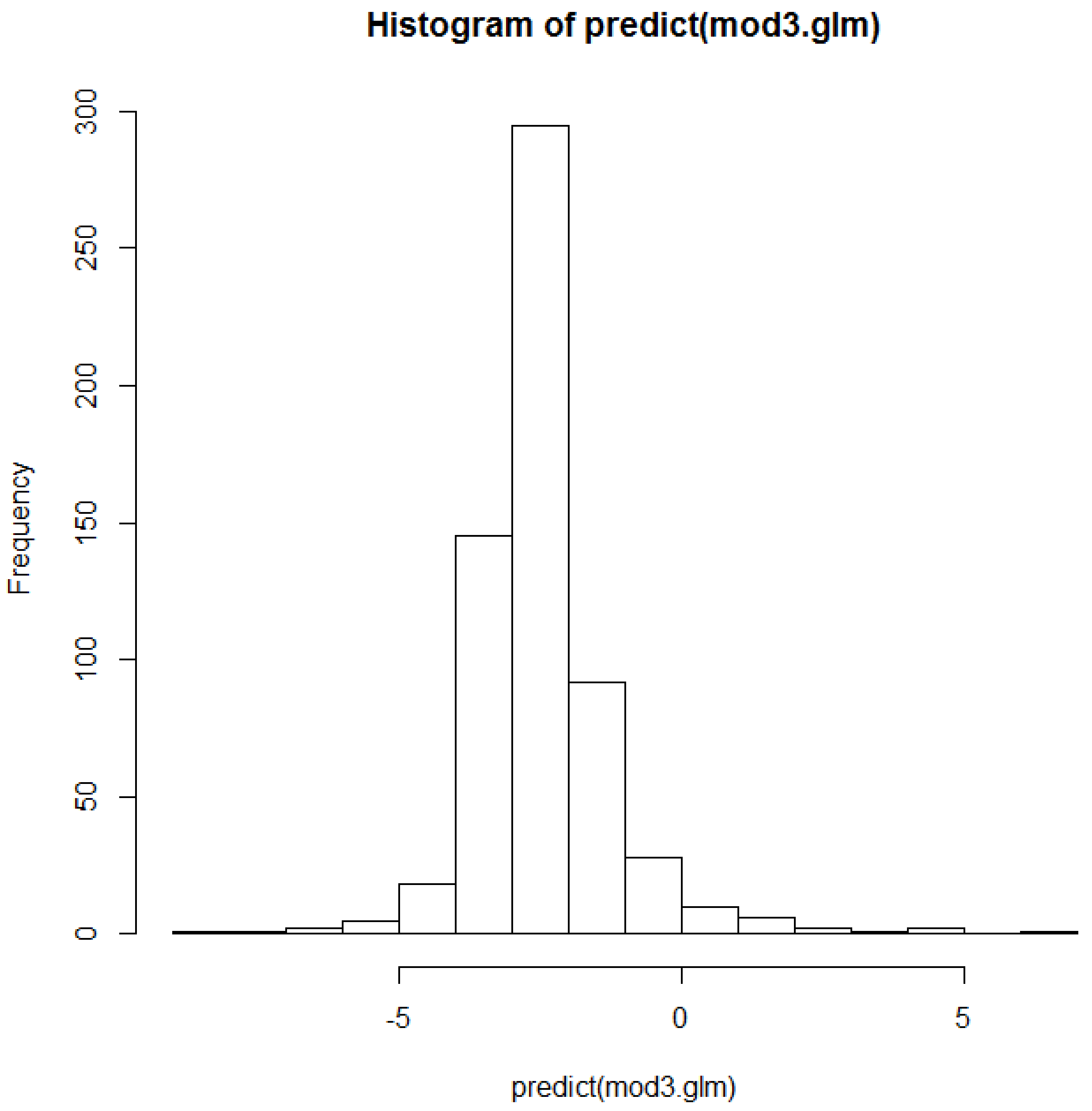
- Sensitivity and Specificity depend only on the test, not on the composition of the population, other figures are dependent
- Sensitivity = fraction of patients with the disease who are predicted to have the disease (p = 0.98).
- Specificity = fraction of patients who are healthy that are classified 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**

Dependence on Population

- 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**

```
> mod3.glm <- glm(CHD~CHL*CAT+SMK+HPT+HPT:CHL+HPT:CAT,binomial,evans)
> summary(mod3.glm)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.981678   1.307727  -3.045  0.00233 **
CHL           0.003506   0.005848   0.599  0.54887
CAT         -13.723211   3.213895  -4.270 1.96e-05 ***
SMK           0.712280   0.326897   2.179  0.02934 *
HPT           4.603360   1.769643   2.601  0.00929 **
CHL:CAT       0.075636   0.014704   5.144 2.69e-07 ***
CHL:HPT      -0.016542   0.008186  -2.021  0.04330 *
CAT:HPT      -2.158014   0.746246  -2.892  0.00383 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 438.56  on 608  degrees of freedom
Residual deviance: 348.80  on 601  degrees of freedom
AIC: 364.8

Number of Fisher Scoring iterations: 6
```


Histogram of fitted(mod3.glm)
Frequency
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
1.0
fitted(mod3.glm)


Histogram of predict(mod3.glm)
Frequency
0
50
100
150
200
250
300
-5
0
5
predict(mod3.glm)

```
> table(fitted(mod3.glm)>0.5,evans$CHD)

          0   1
  FALSE 533  54
  TRUE    5  17


Sensitivity = 17/71 = 23.9%
Specificity = 533/538 = 99.1%
Accuracy = (533+17)/609 = 90.3%
```

```
Predict all are non-CHD


Sensitivity = 0/71 = 0%
Specificity = 538/538 = 100%
Accuracy = (538)/609 = 88.3%%
```

```
> table(fitted(mod3.glm)>0.1,evans$CHD)

          0   1
  FALSE 421  22
  TRUE  117  49

Sensitivity = 49/71 = 69.0%
Specificity = 421/538 = 78.3%%
Accuracy = (421+49)/609 = 77.2%

> 71/609
[1] 0.1165846
```

```
> median(predict(mod3.glm))
[1] -2.554262
> median(fitted(mod3.glm))
[1] 0.0721407
> table(fitted(mod3.glm)>0.0721,evans$CHD)

          0   1
  FALSE 290  13
  TRUE  248  58
```

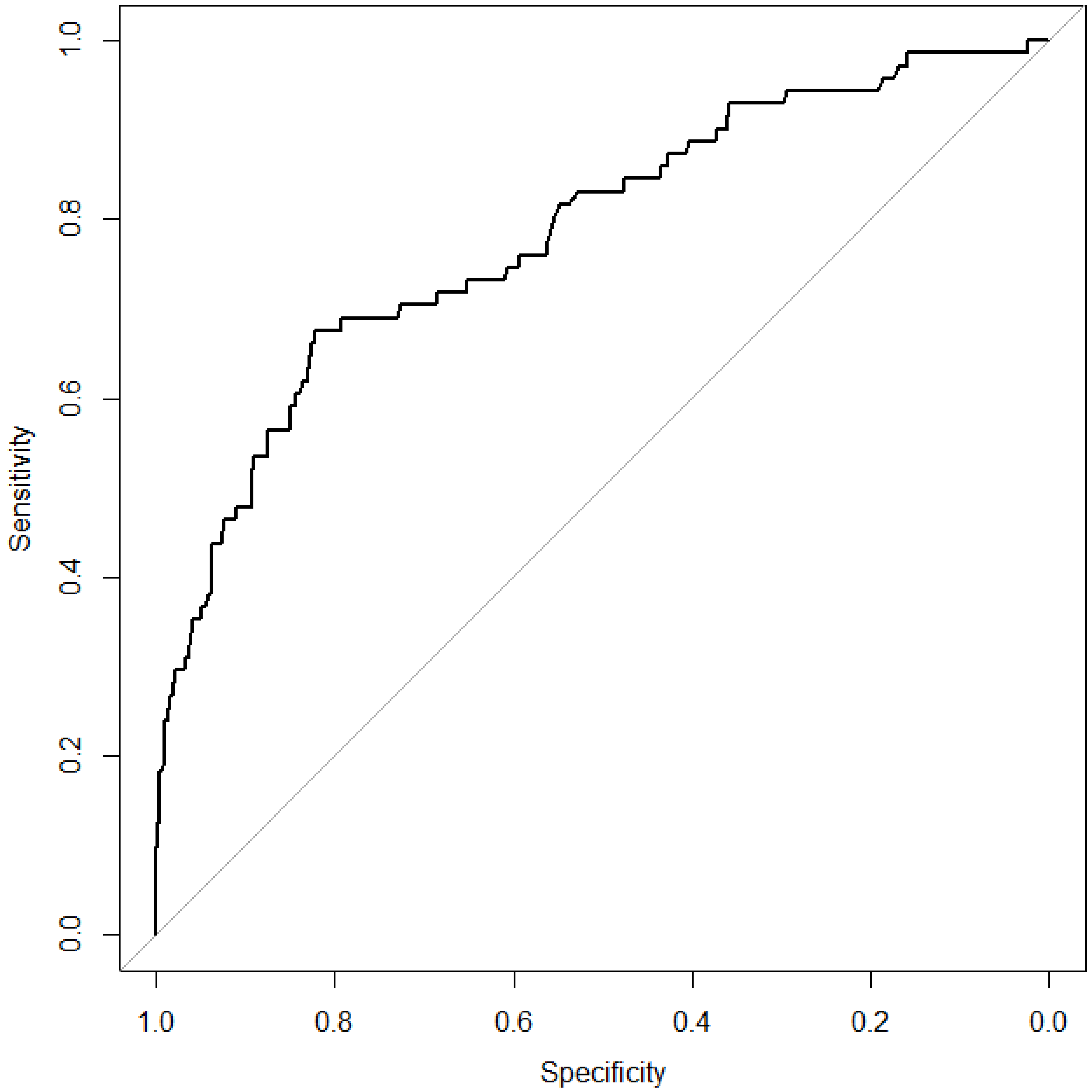
ROC Curve (Receiver Operating Characteristic)

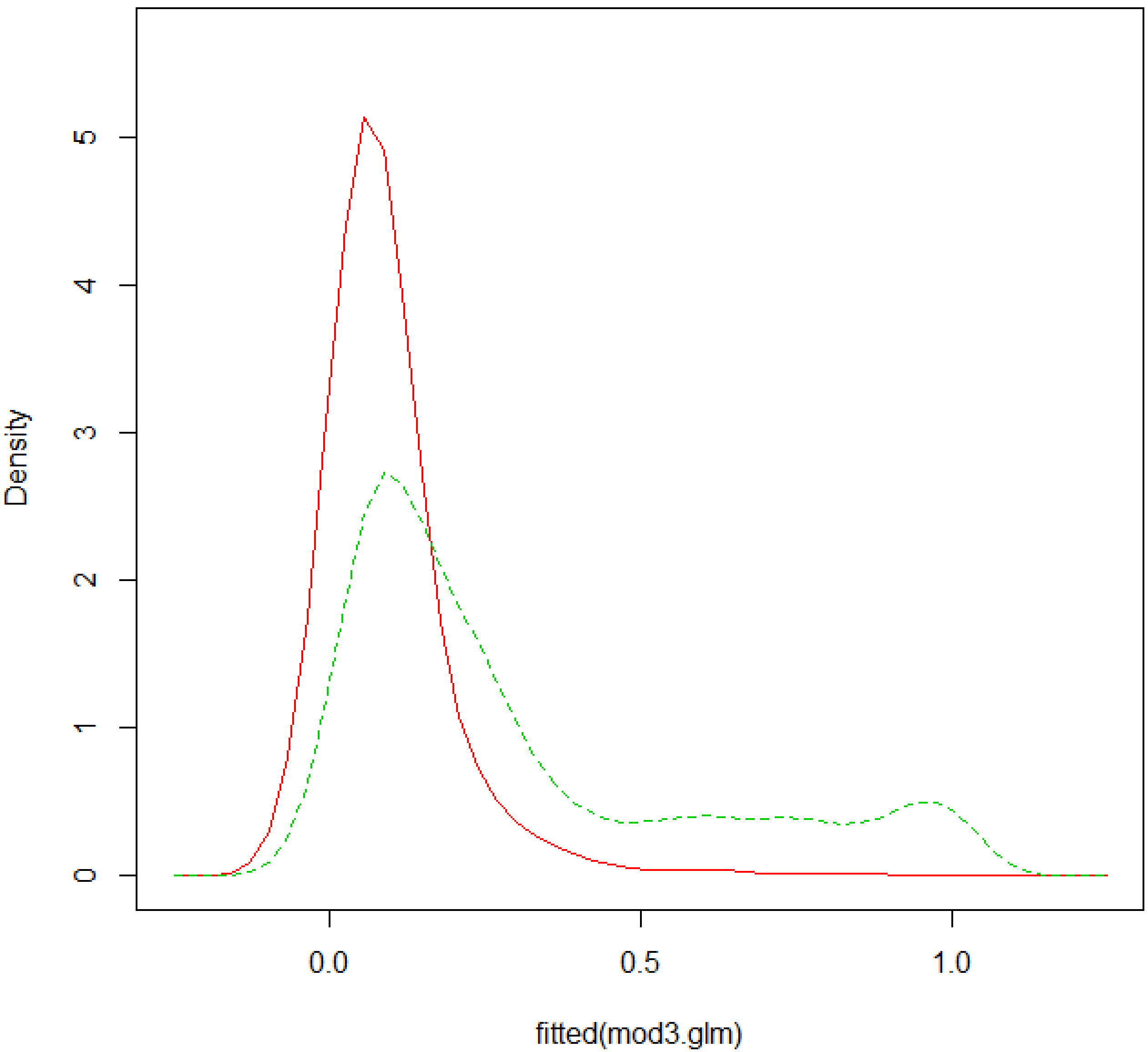
- If we pick a cutpoint *t*, we can assign any case with a predicted value ≤ *t* to Healthy and the others to Disease.
- For that value of *t*, we can compute the number correctly assigned to Disease and the number incorrectly assigned to Disease (true positives and false positives).
- For *t* small 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

R Packages for ROC Curves

- There seem to be many such packages.
- ROCR is the most comprehensive, but a simple ROC plot requires several steps.
- pROC seems easy to use.
- The package sm allows comparison of densities.

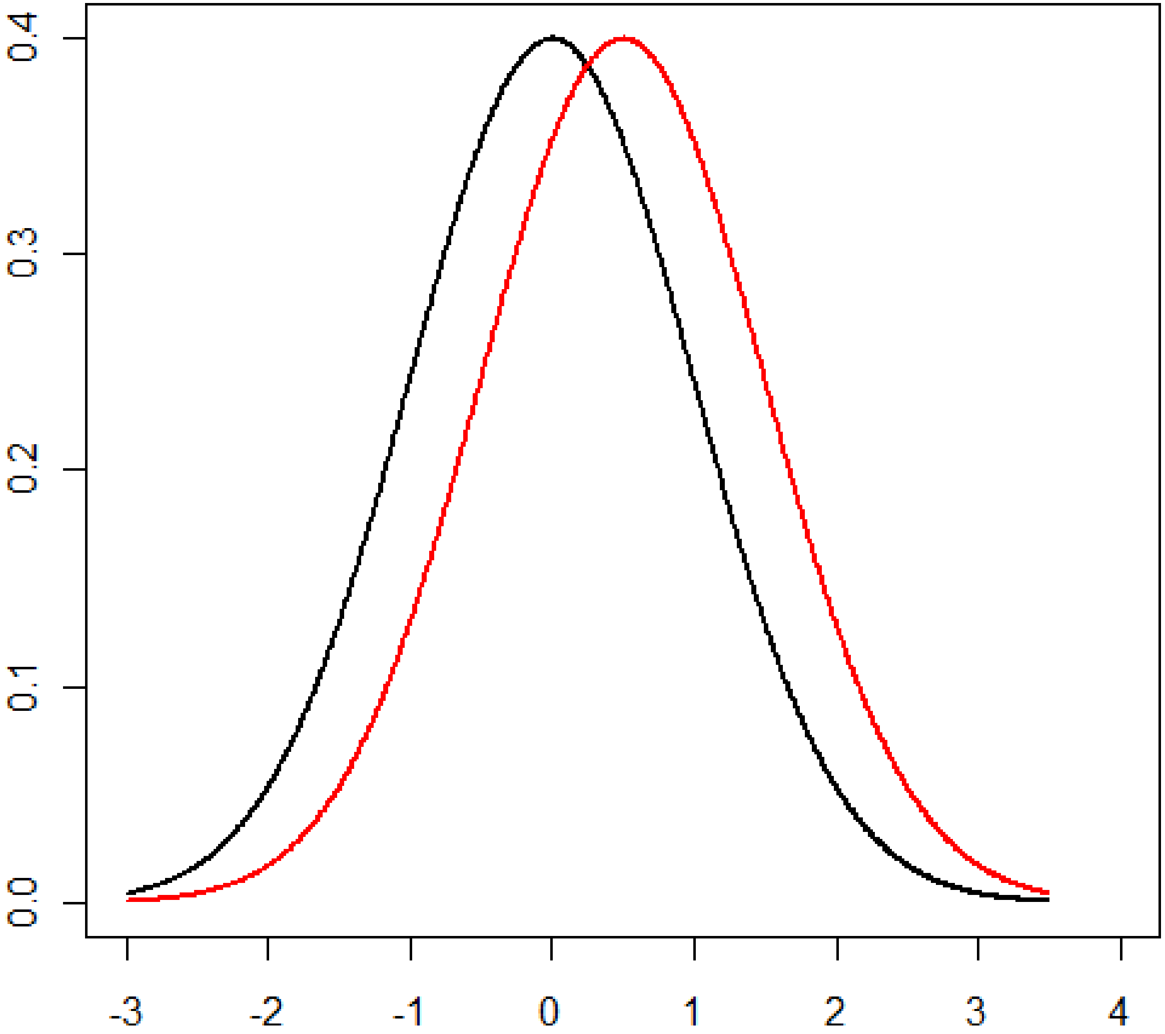
```
> library(pROC)
> mod3.roc <- roc(evans$CHD,fitted(mod3.glm))
> plot(mod3.roc)
Data: fitted(mod3.glm) in 538 controls (evans$CHD 0) <
71 cases (evans$CHD 1).
Area under the curve: 0.7839
> library(sm)
> sm.density.compare(fitted(mod3.glm),evans$CHD)
```


Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0


Density
0
1
2
3
4
5
0.0
0.5
1.0
fitted(mod3.glm)

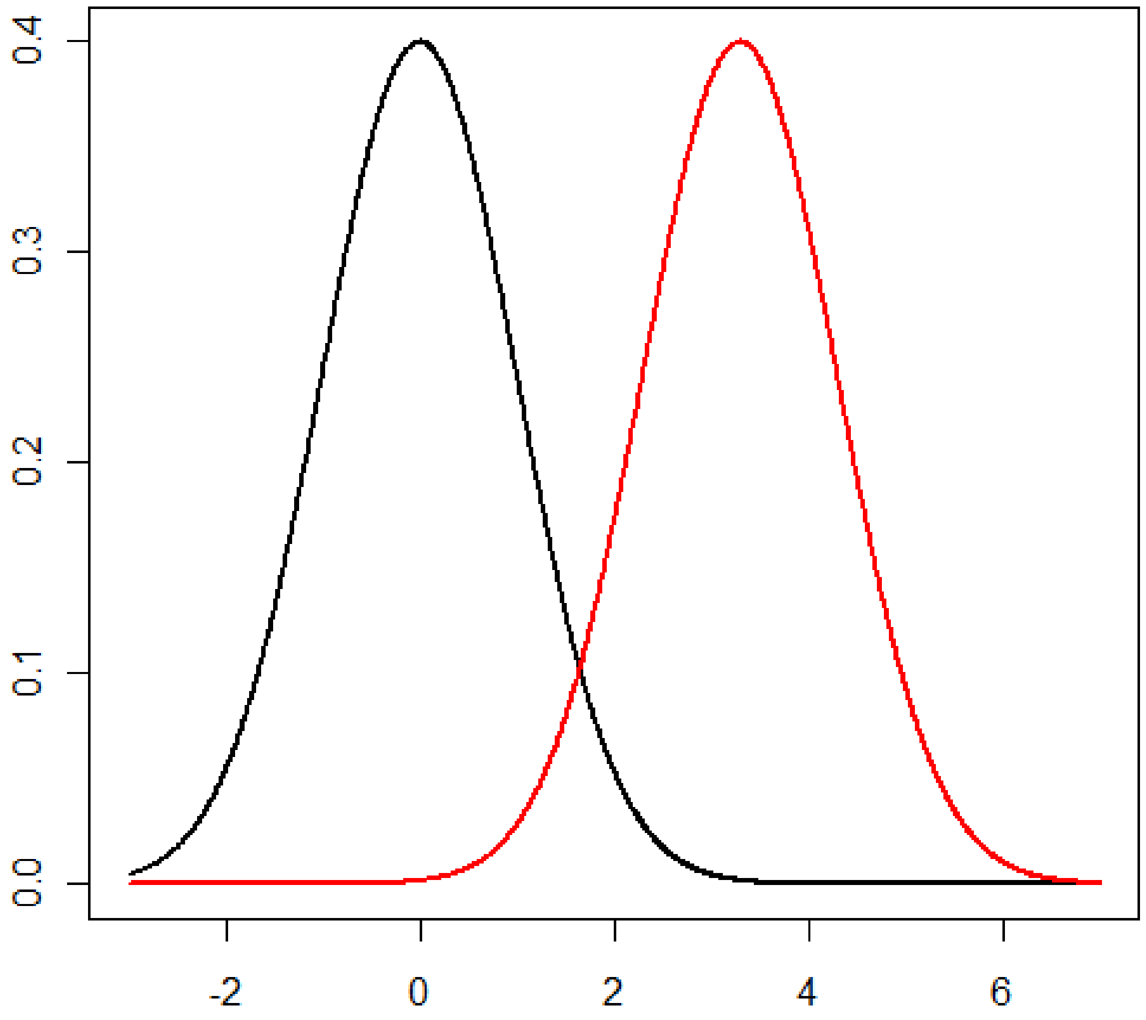
Statistical Significance and Classification Success

- 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.


0.4
0.3
0.2
0.1
0.0
-2
0
2
4
6