

# Simulated ROC

jdt

2/15/2021

## Contents

<b>Preface and Assignment</b>	<b>1</b>
<b>Theory</b>	<b>1</b>
Receiver Operating Characteristics Curve (ROC) . . . . .	1
<b>R</b>	<b>6</b>
<b>SAS</b>	<b>9</b>
Code . . . . .	9
Output . . . . .	11

## Preface and Assignment

In this document I have used Rmarkdown and SAS to LaTeX to create a pdf document file containing a description of the theory used in this problem with the code and output for the analysis using Rmarkdown with RStudio and SAS. I do not expect you to be able to reproduce a document of this type but I do want you to be able to preform the analysis with simulated data. The R and SAS code are found in the pdf document. You should be able to copy this material for use in R or RStudio and SAS. Generate you own simulation by controlling the separation between the population means,  $c_0$  and  $c_1$ . What conclusion can you reach?

## Theory

### Receiver Operating Characteristics Curve (ROC)

In this section, assume that the random variable  $Y$  is continuous and that the test is said to be positive if  $Y \geq c$ , for some  $c$ . For example, let  $Y$  denote the PSA levels that is commonly used to indicate potential problems with the prostate gland when  $Y$  is “large”. The binary test given in the previous section can be constructed for any value of  $c$ . That is, the test is positive if  $Y \geq c$  and is negative if  $Y < c$ , from which we have

$$\begin{aligned}\text{FPF}(c) &= \Pr[Y \geq c \mid D = 0] \\ \text{TPF}(c) &= \Pr[Y \geq c \mid D = 1].\end{aligned}$$

The receiver operating characteristic curve (ROC) for a test using the random variable  $Y$  is defined as

$$\text{ROC}(\cdot) = \{(\text{FPF}(c), \text{TPF}(c)), c \in (-\infty, \infty)\} \quad (1)$$

or

$$\text{ROC}(\cdot) = \{(t, \text{ROC}(t)), t \in (0, 1)\}. \quad (2)$$

Some of the properties for the ROC include:

1. The ROC curve is invariant to strictly (monotone) increasing transformations of  $Y$
2. Let  $S_D = 1 - F_Y(y \mid D = 1)$  and  $S_{\bar{D}} = 1 - F_Y(y \mid D = 0)$  denote the survivor functions of  $Y$  for the diseased and non-diseased populations given by

$$S_D(y) = \Pr[Y \geq y \mid D = 1]$$

$$S_{\bar{D}}(y) = \Pr[Y \geq y \mid D = 0]$$

then

$$ROC(t) = S_D(S_{\bar{D}}^{-1}(t)), t \in (0, 1).$$

- 3.

$$\frac{\partial ROC(t)}{\partial t} = \frac{f_D(S_{\bar{D}}^{-1}(t))}{f_{\bar{D}}(S_{\bar{D}}^{-1}(t))}$$

where  $f_D$  denotes the probability density for  $Y$  in the diseased population ( $D = 1$ ) and  $f_{\bar{D}}$  denotes the probability density for  $Y$  in the healthy population ( $D = 0$ ).

4. The area under the ROC curve (AUC) is

$$AUC = \Pr[Y_D > Y_{\bar{D}}] \tag{3}$$

$$= \int ROC(t) dt. \tag{4}$$

5. (Special Case - Parametric Binormal Form) Suppose that  $Y_D \sim N(\mu_D, \sigma_D^2)$  and  $Y_{\bar{D}} \sim N(\mu_{\bar{D}}, \sigma_{\bar{D}}^2)$  then

$$ROC(t) = \Phi(a + b\Phi^{-1}(t))$$

and

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$$

where  $a = \frac{\mu_D - \mu_{\bar{D}}}{\sigma_D}$ ,  $b = \frac{\sigma_{\bar{D}}}{\sigma_D}$  and  $\Phi$  is the standard normal c.d.f.

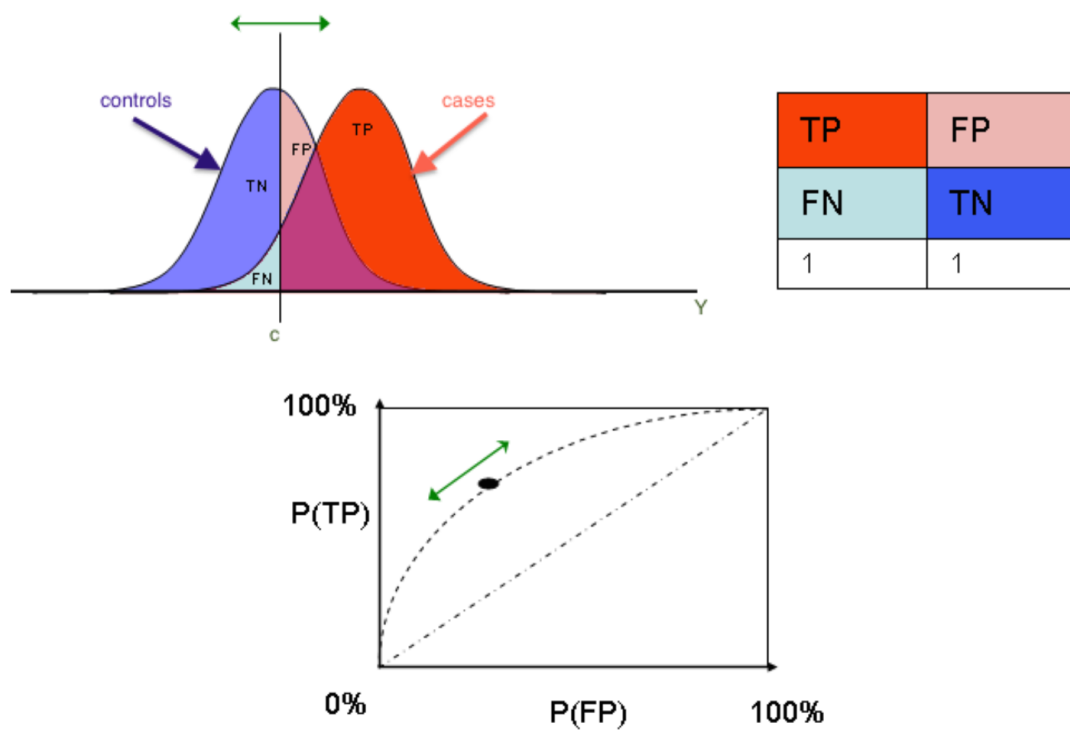


Figure 1: ROC Basics

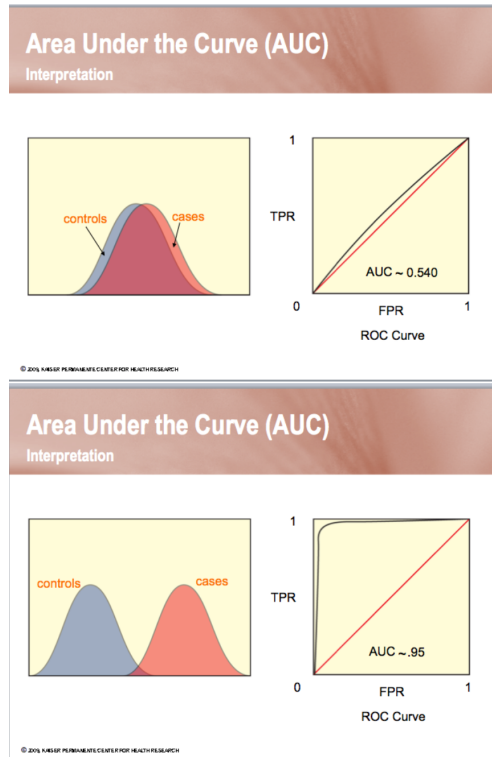


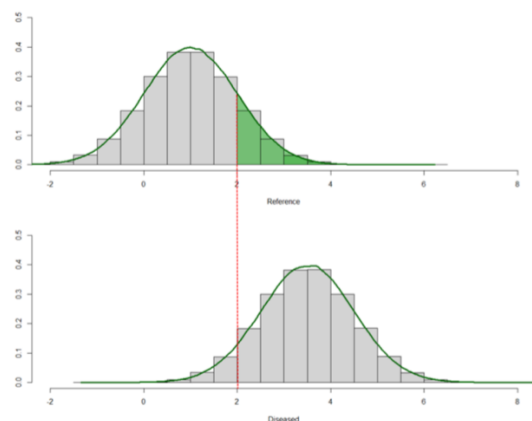
Figure 2: Area Under the Curve - AUC

## Placement Values

A useful concept that is related to the ROC and AUC is a value called the Placement score.

## Placement Values

- Cai(2002) defines  $PV_D = S_{\bar{D}}(Y_D)$ .



- The ROC is equivalent to the cdf of  $PV_D$ .

$$\begin{aligned}
 P[PV_D \leq t | \mathbf{X}] &= P[S_{\bar{D}\mathbf{X}}(Y_D) \leq t | \mathbf{X}] \\
 &= P[Y_D \geq [S_{\bar{D}\mathbf{X}}^{-1}(t) | \mathbf{X}]] \\
 &= ROC_{\mathbf{X}}(t).
 \end{aligned}$$

Figure 3:

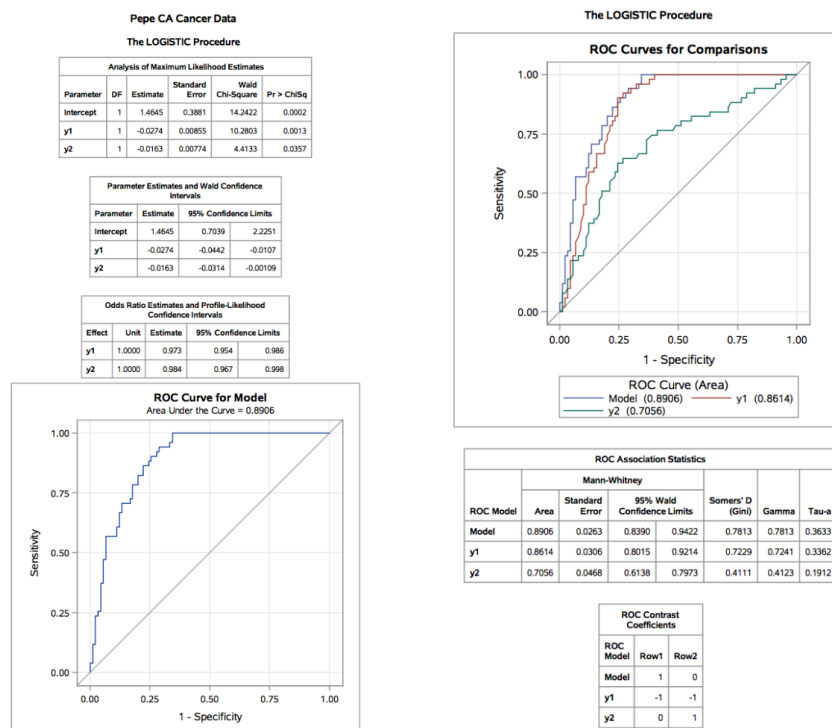
## SAS – Example

The SAS code is,

```
proc logistic data=temp plots(only)=roc;;
  model d=y1 y2 / scale=none
          clparm=wald
          clodds=pl
          rsquare;

  roc 'y1' y1;
  roc 'y2' y2;
  roccontrast reference('y1') / estimate e;
run;
```

Figure contains the SAS output.



## R

Set seed for the simulation

```
# clear the environment and set seed
rm(list = ls())
set.seed(12345)
```

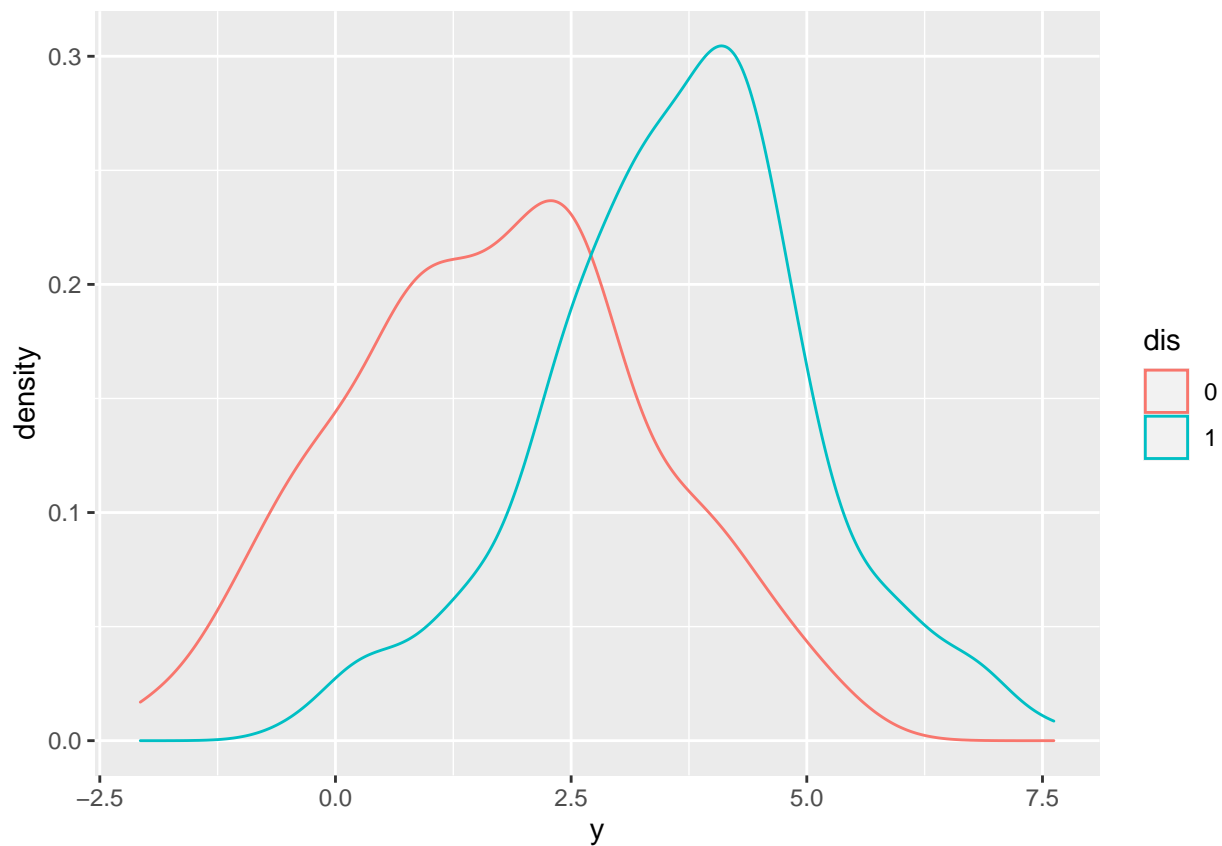
Function to generate normal data. For dis=0 (“Control”) and dis=1 (“Disease”). The separation between the two groups is controlled by one’s choice of c0, c1, sd\_e0, and sd\_e1. n0 and n1 are the sample sizes for the groups.

```
library(ggplot2)

gen_Norm_data = function(c0,sd_e0,n0,c1,sd_e1,n1){
  eps0 = rnorm(n0,0,sd_e0)
  eps1 = rnorm(n1,0,sd_e1)
  y0 = c0 + eps0
  y1 = c1 + eps1

  data.frame(y = c(y0, y1),
             dis = as.factor(c(rep(0,n0), rep(1,n1))))
}
dat1 = gen_Norm_data(1.5,1.5,200,3.5,1.5,200)

#plot density functions
ggplot(dat1, aes(color = dis, y)) + geom_density()
```



Create discrete variables for Y

```
y=dat1[,1]
summary(y)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.071  1.395   2.737   2.672  4.024   7.621
```

```
case=dat1[,2]
```

Create binary table with cutoff  $y > 2.5$

```
high_y = y > 2.5
tex1 = table(case,high_y)
addmargins(tex1)
```

```
##      high_y
## case FALSE TRUE Sum
##  0      137   63 200
##  1       40  160 200
##  Sum    177  223 400
```

```
prop.out = prop.table(tex1,1)
specificity = prop.out[1,1]
sensitivity = prop.out[2,2]
prop.out
```

```
##      high_y
## case FALSE  TRUE
##  0 0.685 0.315
##  1 0.200 0.800
```

```
sensitivity
```

```
## [1] 0.8
```

```
specificity
```

```
## [1] 0.685
```

```
TPR = sensitivity
FPR = 1 - specificity
TPR
```

```
## [1] 0.8
```

```
FPR
```

```
## [1] 0.315
```

Create binary table with cutoff  $y > 3.0$

```
high_y = y > 3.0
tex1 = table(case,high_y)
addmargins(tex1)
```

```
##      high_y
## case FALSE TRUE Sum
##  0      161   39 200
##  1       60  140 200
##  Sum    221  179 400
```

```
prop.out = prop.table(tex1,1)
specificity = prop.out[1,1]
sensitivity = prop.out[2,2]
prop.out
```

```
##      high_y
## case FALSE  TRUE
##  0 0.805 0.195
##  1 0.300 0.700
```



```
sensitivity
```

```
## [1] 0.7
```

```
specificity
```

```
## [1] 0.805
```

```
TPR = sensitivity
```

```
FPR = 1 - specificity
```

```
TPR
```

```
## [1] 0.7
```

```
FPR
```

```
## [1] 0.195
```

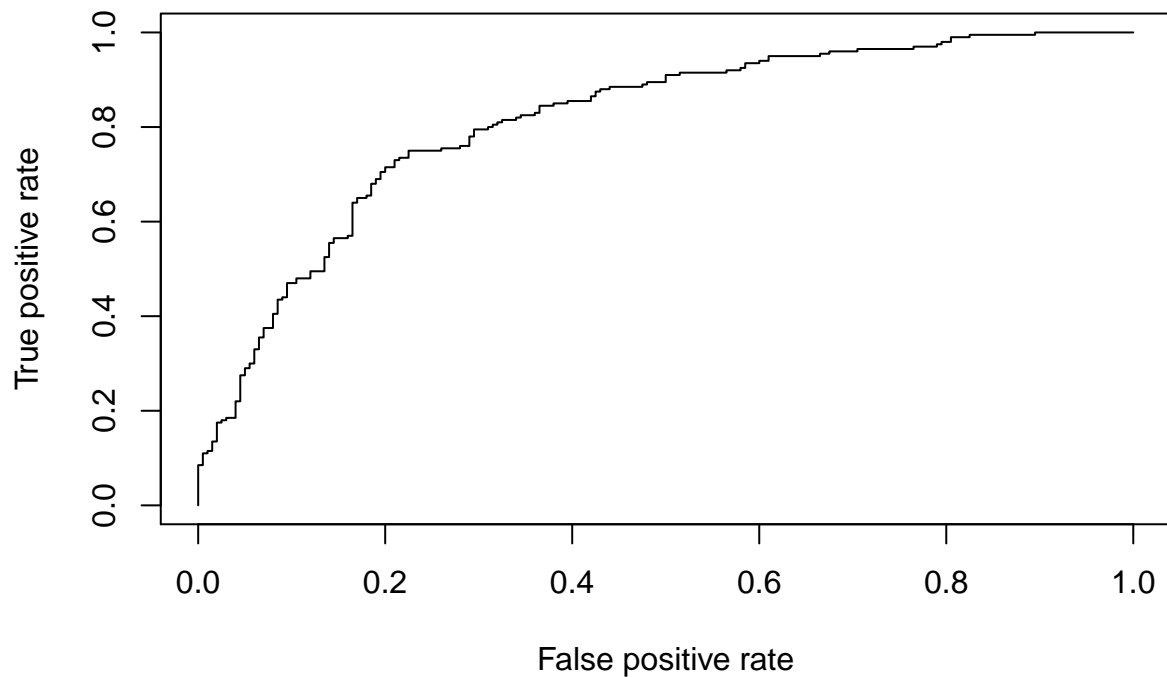
Instead of creating a table at each cutoff point one can construct a ROC plot for a continuous variable Y. Which is constructing using the pairs (TPR, FPR) at each cutoff point (in R). Sometimes (as in SAS) this curve is smoothed.

```
library(ROCR)
```

```
pred = prediction(y,case)
```

```
perf=performance(pred, "tpr", "fpr")
```

```
plot(perf)
```



## SAS

### Code

```
libname LDATA '/home/jacktubbs/my_shared_file_links/jacktubbs/myfolders/Titanic/';
options center nodate pagesize=100 ls=80;
/* Simplified LaTeX output that uses plain LaTeX tables */
ods tagsets.simplelatex file="/home/jacktubbs/my_shared_file_links
/jacktubbs/LaTeX/ROC_sim.tex"
```

```

stylesheet="/home/jacktubbs/my_shared_file_links
/jacktubbs/LaTeX/sas.sty"(url="sas");

/*
The above will create a new file that can be inputted into LaTeX
(simple.tex) and the new style needed by LaTeX (sas.sty).

The following example can be found at

http://support.sas.com/rnd/base/ods/odsmarkup/latex.html

*/
ods graphics on;

title1 'Simulated Data for ROC';
* Run Macro;
%macro binorm(dsn, title);
title2 &title;
data a;set &dsn;
seed=12345;
do i = 1 to n0;
group='control';
y = c0 + rand("Normal", 0, sd_e0);
output;
end;
do i = 1 to n1;
group='disease';
y = c1 + rand("Normal", 0, sd_e1);
output;
end;
run;

data a; set a; y_cut = (y > cutoff); run;

proc sgplot data=a;
density y /type=kernel group=group;
run;

proc freq; table group*y_cut/ nopercnt nocol outpct out=b ; run;
data c; set b;
if group = 'disease' and y_cut = '1' then sensitivity = pct_row;
if group = 'control' and y_cut = '0' then specificity = pct_row;
keep sensitivity specificity;
run;
proc print data=c; var sensitivity specificity; run;

proc logistic data=a plots(only)=roc ;
class group ;
model group(event='disease')=y;
run;
%mend binorm;

*Run data generation;
data parms1; c0=1.5; sd_e0=1.5; n0=100; c1=2.0; sd_e1=1.5; n1=100; cutoff = 3.0;
run;
data parms2; c0=1.5; sd_e0=1.5; n0=100; c1=2.5; sd_e1=1.5; n1=100; cutoff = 3.0;
run;
data parms3; c0=1.5; sd_e0=1.5; n0=100; c1=3.0; sd_e1=1.5; n1=100; cutoff = 3.0;
run;

```

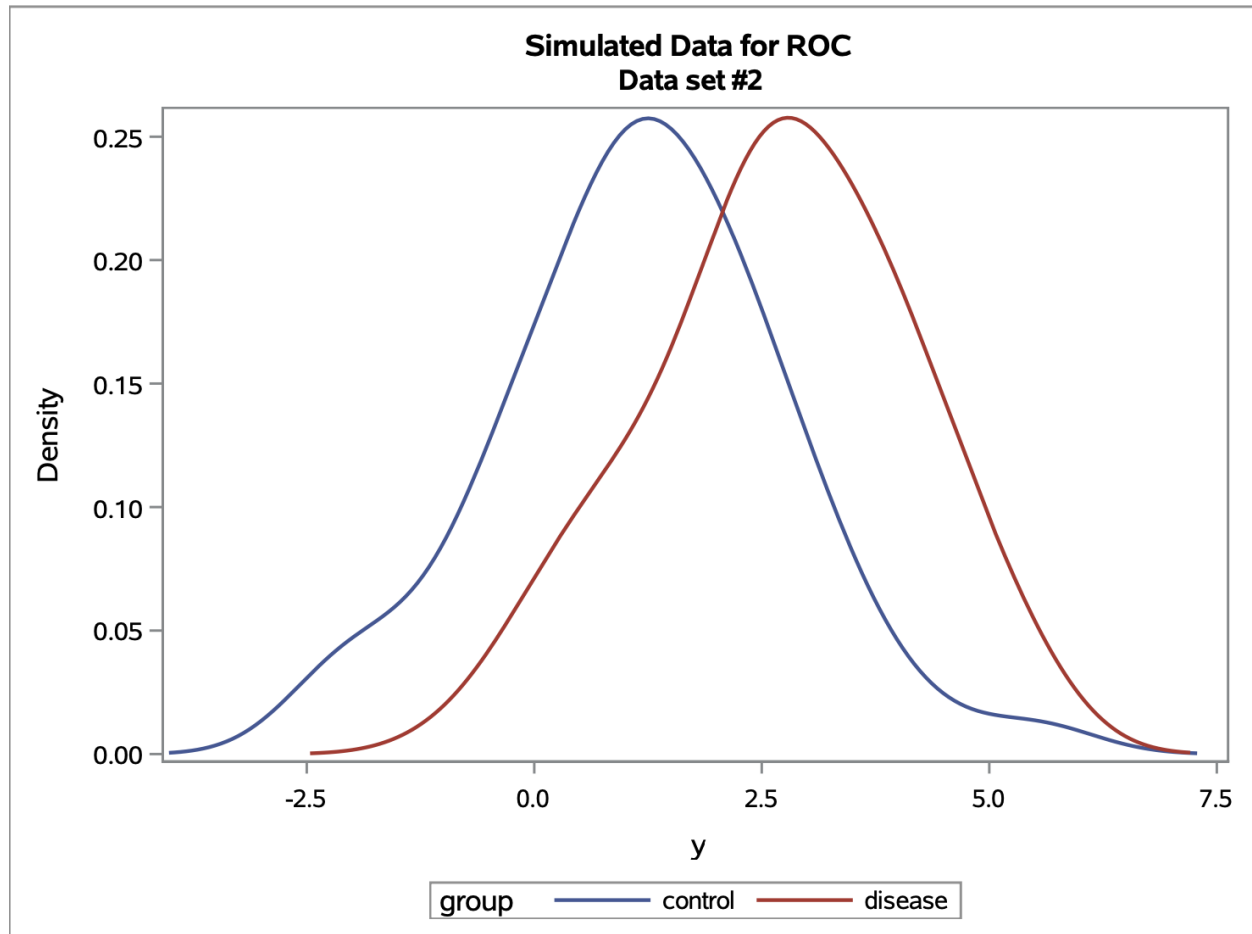
```

data parms4; c0=1.5; sd_e0=1.5; n0=100; c1=3.5; sd_e1=1.5; n1=100; cutoff = 3.0;
run;

%binorm(parms1,'Data set #1');
%binorm(parms2,'Data set #2');
%binorm(parms3,'Data set #3');
%binorm(parms4,'Data set #4');
quit;

```

## Output



***Simulated Data for ROC***  
***Data set #2***  
***The FREQ Procedure***

Table of group by y_cut			
group	y_cut		
	0	1	Total
control	86	14	100
	86.00	14.00	
disease	60	40	100
	60.00	40.00	
Total	146	54	200

***Simulated Data for ROC***  
***Data set #2***

Obs	sensitivity	specificity
1	.	86
2	.	.
3	.	.
4	40	.

***Simulated Data for ROC***  
***Data set #2***  
***The LOGISTIC Procedure***

Model Information	
Data Set	WORK.A
Response Variable	group
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	200
Number of Observations Used	200

Response Profile		
Ordered Value	group	Total Frequency
1	control	100
2	disease	100

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	279.259	255.599
SC	282.557	262.196
-2 Log L	277.259	251.599

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	25.6598	1	<.0001
Score	24.2216	1	<.0001
Wald	21.7853	1	<.0001

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.9345	0.2504	13.9281	0.0002
y	1	0.4804	0.1029	21.7853	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
y	1.617	1.321	1.978

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	70.2	Somers' D	0.405
Percent Discordant	29.8	Gamma	0.405
Percent Tied	0.0	Tau-a	0.203
Pairs	10000	c	0.702

