Lecture 5

Ciprian Crainiceanu

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health
Johns Hopkins University

September 8, 2020

Table of contents

Outline

1. Define conditional probabilities
2. Define conditional mass functions and densities
3. Motivate the conditional density
4. Bayes' rule
5. Applications of Bayes' rule to diagnostic testing

Conditional probability, examples

- What is the probability for a 30 year old woman to develop breast cancer within 10 years?
- X is "develop cancer within the next 10 years"
- We would like to calculate probabilities of the type

$$P(X = 1 | \mathrm{sex} = 1, \mathrm{age} = 30)$$

- What happens if $\mathrm{age} = 50$?
- What happens if the person is a man $\mathrm{sex} = 0$?
- What one conditions on is crucial

Conditional probability, examples

- What is the probability of surviving more than 1 year for a man who is 50 years old and has an estimated glomerular filtration rate (eGFR) equal to 15?
- X is surviving time
- We would like to calculate probabilities of the type

$$P(X > 1 | \mathrm{sex} = 0, \mathrm{age} = 50, \mathrm{eGFR} = 15)$$

Conditional probability, motivation

- The probability of getting a one when rolling a (standard) die is usually assumed to be one sixth
- Suppose you were given the extra information that the die roll was an odd number (hence 1, 3 or 5)
- *conditional on this new information*, the probability of a one is now one third

Conditional probability, definition

- Let B be an event so that $P(B) > 0$
- Then the conditional probability of an event A given that B has occurred is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

- Notice that if A and B are independent, then

$$P(A \mid B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

Example

- Consider our die roll example
- $B = \{1, 3, 5\}$
- $A = \{1\}$

$$
\begin{aligned}
P(\text{one given that roll is odd}) &= P(A \mid B) \\
&= \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(A)}{P(B)} \\
&= \frac{1/6}{3/6} = \frac{1}{3}
\end{aligned}
$$

Conditional densities and mass functions

- Conditional densities or mass functions of one variable conditional on the value of another
- Let $f(x, y)$ be a bivariate density or mass function for random variables X and Y
- Let $f(x)$ and $f(y)$ be the associated marginal mass function or densities disregarding the other variables

$$f(y) = \int f(x, y)dx \quad \text{or} \quad f(y) = \sum_x f(x, y)dx.$$

- Then the **conditional** density or mass function *given that* $Y = y$ is given by

$$f(x \mid y) = f(x, y)/f(y)$$

Notes

- It is easy to see that, in the discrete case, the definition of conditional probability is exactly as in the definition for conditional events where $A =$ the event that $X = x_0$ and $B =$ the event that $Y = y_0$
- The continuous definition is harder to motivate, since the events $X = x_0$ and $Y = y_0$ each have probability 0
- However, a useful motivation can be performed by taking the appropriate limits as follows
- Define $A = \{X \leq x_0\}$ while $B = \{Y \in [y_0, y_0 + \epsilon]\}$

Continued

$$
\begin{aligned}
P(X \leq x_0 \mid Y \in [y_0, y_0+\epsilon]) &= P(A \mid B) = \frac{P(A \cap B)}{P(B)} \\
&= \frac{P(X \leq x_0, Y \in [y_0, y_0+\epsilon])}{P(Y \in [y_0, y_0+\epsilon])} \\
&= \frac{\int_{y_0}^{y_0+\epsilon} \int_{-\infty}^{x_0} f(x,y)dxdy}{\int_{y_0}^{y_0+\epsilon} f(y)dy} \\
&= \frac{\epsilon \int_{y_0}^{y_0+\epsilon} \int_{-\infty}^{x_0} f(x,y)dxdy}{\epsilon \int_{y_0}^{y_0+\epsilon} f(y)dy}
\end{aligned}
$$

Continued

$$= \frac{\frac{\int_{-\infty}^{y_0+\epsilon}\int_{\infty}^{x_0} f(x,y)dxdy - \int_{-\infty}^{y_0}\int_{-\infty}^{x_0} f(x,y)dxdy}{\epsilon}}{\frac{\int_{-\infty}^{y_0+\epsilon} f(y)dy - \int_{-\infty}^{y_0} f(y)dy}{\epsilon}}$$

$$= \frac{\frac{g_1(y_0+\epsilon)-g_1(y_0)}{\epsilon}}{\frac{g_2(y_0+\epsilon)-g_2(y_0)}{\epsilon}}$$

where

$$g_1(y_0) = \int_{-\infty}^{y_0}\int_{-\infty}^{x_0} f(x,y)dxdy \quad \text{and} \quad g_2(y_0) = \int_{-\infty}^{y_0} f(y)dy.$$

- Notice that the limit of the numerator and denominator tends to g_1' and g_2' as ϵ gets smaller and smaller
- Hence we have that the conditional distribution function is

$$P(X \leq x_0 \mid Y = y_0) = \frac{\int_{-\infty}^{x_0} f(x, y_0) dx}{f(y_0)}.$$

- Now, taking the derivative with respect to x yields the conditional density

$$f(x_0 \mid y_0) = \frac{f(x_0, y_0)}{f(y_0)}$$

for every x_0 and y_0 and subscript can now be dropped

Geometry

- Geometrically, the conditional density is obtained by taking the relevant slice of the joint density and appropriately renormalizing it
- This idea extends to any other linear or non-linear function

Example

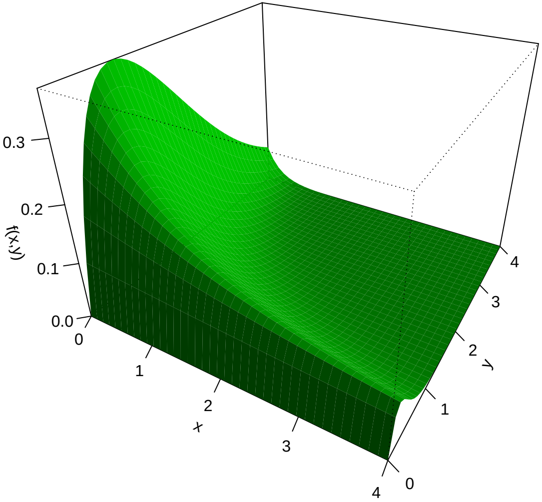
- Let $f(x, y) = ye^{-xy-y}$ for $0 \leq x$ and $0 \leq y$
- Then note

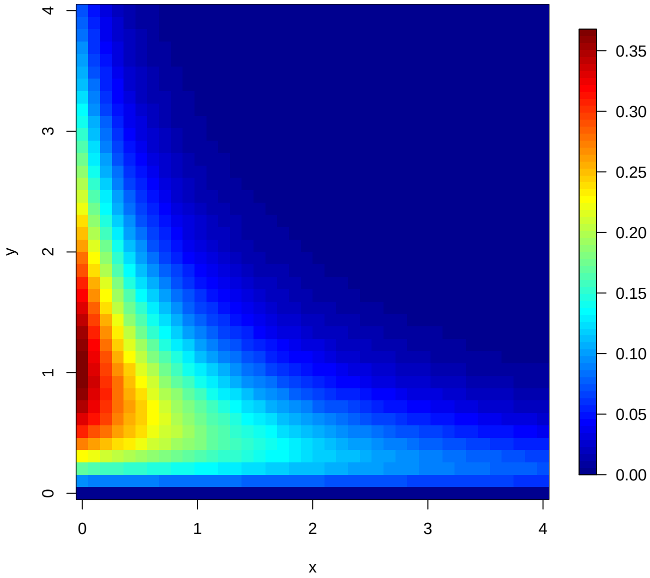
$$f(y) = \int_0^\infty f(x, y)dx = e^{-y} \int_0^\infty ye^{-xy}dx = e^{-y}$$

- Therefore

$$f(x \mid y) = f(x, y)/f(y) = \frac{ye^{-xy-y}}{e^{-y}} = ye^{-xy}$$

- Calculate $P(X \geq 5|Y = 3)$


0.3
0.2
0.1
0.0
f(x,y)
0
1
2
3
4
x
0
1
2
3
4
y


y
x
0
1
2
3
4
0
1
2
3
4
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00

- Check out the R functions `persp`, `image.plot`, `plot3D`, `surface3d`
- Useful packages: `rgl`, `fields`

Example

- Let $f(x,y) = 1/\pi r^2$ for $x^2 + y^2 \leq r^2$
- X and Y are uniform on a disk with radius r
- What is the conditional density of X given that $Y = 0$?
- Probably easiest to think geometrically

$$f(x \mid y = 0) \propto 1 \text{ for } -r \leq x \leq r$$

- Therefore

$$f(x \mid y = 0) = \frac{1}{2r} \text{ for } -r \leq x \leq r$$

Bayes' rule

- Let $f(x \mid y)$ be the conditional density or mass function for X given that $Y = y$
- Let $f(y)$ be the marginal distribution for y
- Then if y is continuous

$$f(y \mid x) = \frac{f(x \mid y)f(y)}{\int f(x \mid t)f(t)dt}$$

- If y is discrete

$$f(y \mid x) = \frac{f(x \mid y)f(y)}{\sum_t f(x \mid t)f(t)}$$

Notes

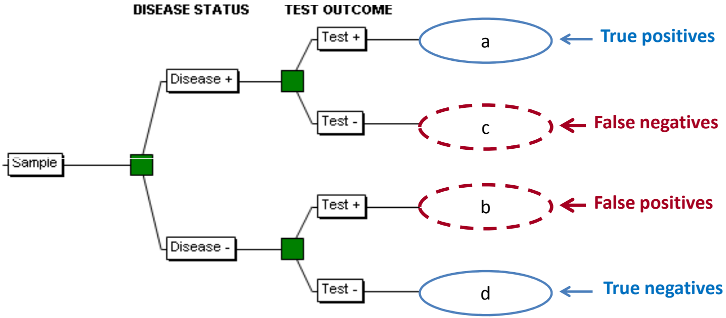
- Bayes' rule relates the conditional density of $f(y \mid x)$ to the conditional density $f(x \mid y)$ and the marginal density $f(y)$
- A special case of this kind relationship is for two sets A and B, which yields that

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)}.$$

Proof:

- Let X be an indicator that event A has occurred
- Let Y be an indicator that event B has occurred
- Plug into the discrete version of Bayes' rule

Example: diagnostic tests



Example: diagnostic tests

- Let $+$ and $-$ be the events that the result of a diagnostic test is positive or negative, respectively
- Let D and D^c be the event that the subject of the test has or does not have the disease respectively
- The **sensitivity** is the probability that the test is positive given that the subject actually has the disease, $P(+ \mid D)$
- The **specificity** is the probability that the test is negative given that the subject does not have the disease, $P(- \mid D^c)$

More definitions

- The **positive predictive value** is the probability that the subject has the disease given that the test is positive, $P(D \mid +)$
- The **negative predictive value** is the probability that the subject does not have the disease given that the test is negative, $P(D^c \mid -)$
- The **prevalence of the disease** is the marginal probability of disease, $P(D)$

More definitions

- The **diagnostic likelihood ratio of a positive test**, labeled DLR_+, is $P(+ \mid D)/P(+ \mid D^c)$, which is the

$$\mathit{sensitivity}/(1 - \mathit{specificity})$$

- The **diagnostic likelihood ratio of a negative test**, labeled DLR_-, is $P(- \mid D)/P(- \mid D^c)$, which is the

$$(1 - \mathit{sensitivity})/\mathit{specificity}$$

Example

- A study comparing the efficacy of HIV tests, reports on an experiment which concluded that HIV antibody tests have a sensitivity of 99.7% and a specificity of 98.5%
- Suppose that a subject, from a population with a .1% prevalence of HIV, receives a positive test result. What is the probability that this subject has HIV?
- Mathematically, we want $P(D \mid +)$ given the sensitivity, $P(+ \mid D) = .997$, the specificity, $P(- \mid D^c) = .985$, and the prevalence $P(D) = .001$

Using Bayes' formula

$$
\begin{aligned}
P(D \mid +) &= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} \\
&= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + \{1 - P(- \mid D^c)\}\{1 - P(D)\}} \\
&= \frac{.997 \times .001}{.997 \times .001 + .015 \times .999} \\
&= .062
\end{aligned}
$$

- In this population a positive test result only suggests a 6% probability that the subject has the disease
- (The positive predictive value is 6% for this test)

More on this example

- The low positive predictive value is due to low prevalence of disease and the somewhat modest specificity
- Suppose it was known that the subject was an intravenous drug user and routinely had intercourse with an HIV infected partner
- Notice that the evidence implied by a positive test result does not change because of the prevalence of disease in the subject's population, only our interpretation of that evidence changes

Likelihood ratios

- Using Bayes rule, we have

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)}$$

and

$$P(D^c \mid +) = \frac{P(+ \mid D^c)P(D^c)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)}.$$

- Therefore

$$\frac{P(D \mid +)}{P(D^c \mid +)} = \frac{P(+ \mid D)}{P(+ \mid D^c)} \times \frac{P(D)}{P(D^c)}$$

ie

$$\text{post-test odds of } D = DLR_+ \times \text{pre-test odds of } D$$

- Similarly, DLR_- relates the decrease in the odds of the disease after a negative test result to the odds of disease prior to the test.

HIV example revisited

- Suppose a subject has a positive HIV test
- $DLR_{+} = .997/(1 - .985) \approx 66$
- The result of the positive test is that the odds of disease is now 66 times the pretest odds
- Or, equivalently, the hypothesis of disease is 66 times more supported by the data than the hypothesis of no disease

HIV example revisited

- Suppose that a subject has a negative test result
- $DLR_{-} = (1 - .997)/.985 \approx .003$
- Therefore, the post-test odds of disease is now .3% of the pretest odds given the negative test.
- Or, the hypothesis of disease is supported .003 times that of the hypothesis of absence of disease given the negative test result

Comparing two tests

- Test 1: $DLR_+ = a$, Test 2: $DLR_+ = b$
- Test 1: a is the factor that multiplies the pre-test odds to obtain the post-test odds

$$\frac{P(D|T_1 = +)}{P(D_C|T_1 = +)} = a \times \frac{P(D)}{P(D_C)}$$

- Test 2: b is the factor that multiplies the pre-test odds to obtain the post-test odds

$$\begin{aligned} O(D|T_1 = +, T_2 = +) &= b \times O(D|T_1 = +) \\ &= a \times b \times O(D) \end{aligned}$$

Tests and 2×2 tables

A particularly interesting and important question today is that of testing for drugs. Suppose it is assumed that about 5% of the general population uses drugs. You employ a test that is 95% accurate, which we will say means that if the individual is a user, the test will be positive 95% of the time, and if the individual is a nonuser, the test will be negative 95% of the time. A person is selected at random and is given the test. It's positive. What does such a result suggest? Would you conclude that the individual is a drug user? What is the probability that the person is a drug user?

The 2 × 2 table

	Disease +	Disease -	Total
Test +	a	b	a + b
Test -	c	d	c + d
Total	a + c	b + d	a + b + c + d

$$PPV = P(D \mid +) = \frac{a}{a+b}$$

$$NPV = P(\overline{D} \mid -) = \frac{d}{c+d}$$

$$Sens = P(+ \mid D) = \frac{a}{a+c}$$

$$Spec = P(- \mid \overline{D}) = \frac{d}{b+d}$$

The 2 × 2 table: example

	Disease +	Disease -	Total
Test +	48	47	95
Test -	2	903	905
Total	50	950	1000

PPV = 51%

NPV = 99%

P(D) = 0.05

The 2 × 2 table: example

	Disease +	Disease -	Total
Test +	190	40	230
Test -	10	760	770
Total	200	800	1000

PPV = 83%

NPV = 99%

P(D) = 0.20

Point: PPV depends on **prior probability** of disease in the population

Prediction with binary outcomes

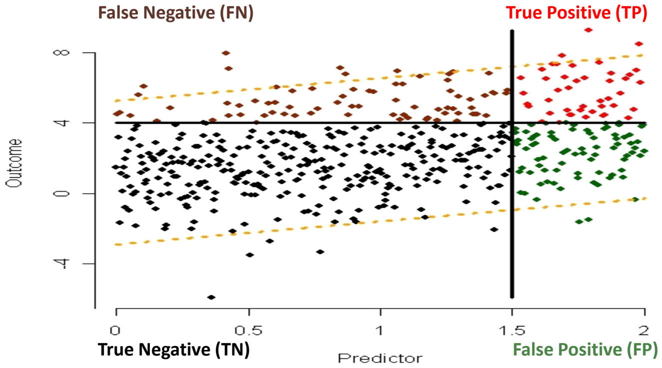
- Outcome is 0/1
- Examples
  - Non-diseased/diseased
  - Alive/Dead
  - Failure/Success (procedure)
- Continuous predictor
- Examples
  - Outcome of a diagnostic test
  - Prediction score (based on multiple characteristics)
  - Clinical score (e.g. SOFA score in ICU)

The ROC

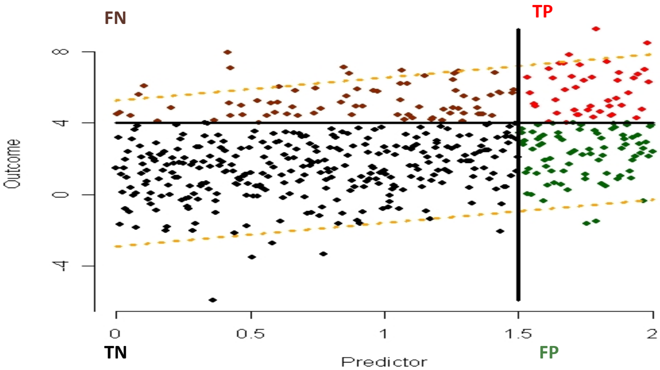
- Outcome $D \in \{0, 1\}$, X scalar predictor
- For every threshold t predict $\widehat{D} = 1$ if $X > t$
- $\text{Sens}(t) = P(X > t|D = 1)$, $\text{Spec}(t) = P(X \leq t|D = 0)$
- The receiver operatic characteristic (ROC) function is

$$\{1 - \text{Spec}(t), \text{Sens}(t)\} \text{ for all } t$$

Luck, error and randomness



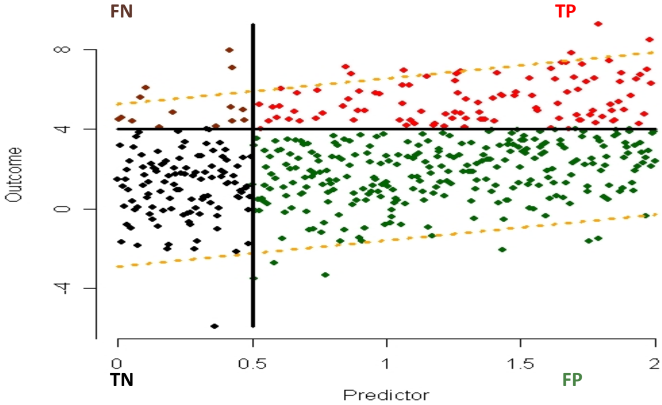
Dependence on the threshold



$$\text{Spec} = \frac{\text{TN}}{\text{TN+FP}} = \frac{301}{301+81} = 0.79$$

$$\text{Sens} = \frac{\text{TP}}{\text{TP+FN}} = 0.37$$

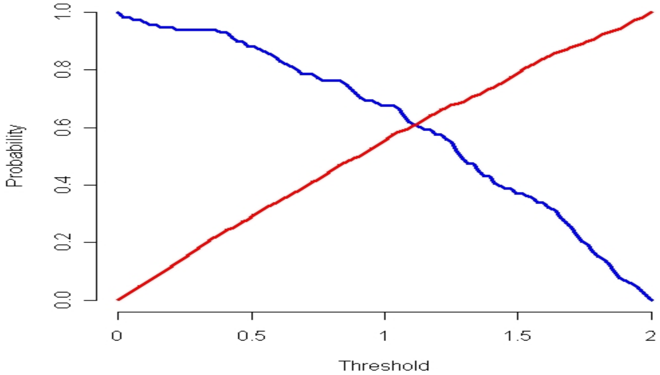
Dependence on the threshold



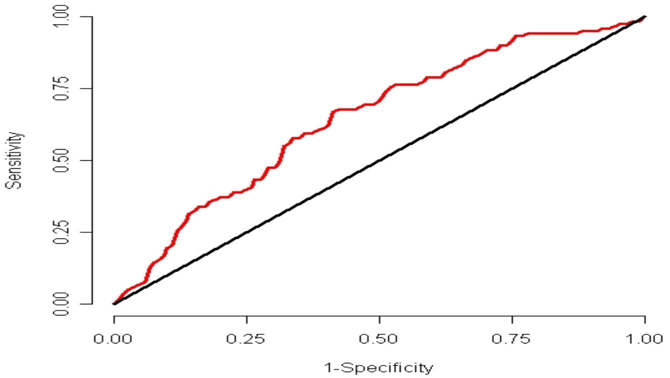
$$\text{Spec} = \frac{\text{TN}}{\text{TN+FP}} = \frac{111}{111+271} = 0.29$$

$$\text{Sens} = \frac{\text{TP}}{\text{TP+FN}} = 0.88$$

Sensitivity and Specificity curves



ROC


Sensitivity
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
1-Specificity

AUC

- Area under the ROC curve is denoted by AUC
- Probability that the model will assign a higher probability of an event to the subject who will experience the event than to the one who will not experience the event
- AUC is one of the main criteria for assessing discrimination accuracy
- AUC=0.68 in the example

AUC interpretation proof

$$
\begin{aligned}
\text{Sens}(t) &= S(t) = P(X > t|D = 1) = \int_t^1 f(x|D = 1)dx \\
1 - \text{Spec}(t) &= P(t) = P(X > t|D = 0) = \int_t^1 f(x|D = 0)dx \\
\text{AUC} &= \int_1^0 S(t)\frac{d}{dt}P(t) = \int_0^1 S(t)f(t|D = 0)dt \\
&= P(X_i > X_j|D_i = 1, D_i = 0)
\end{aligned}
$$

Note that $f(x_i, x_j|D_i = 1, D_j = 0) = f(x_i|D_i = 1)f(x_j|D_j = 0)$

Some comments

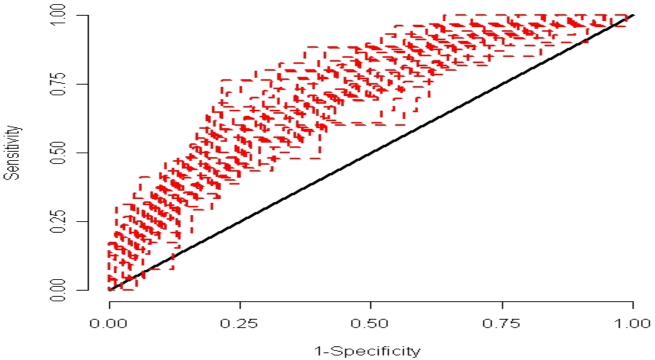
- ROC, AUC are never observed
- They are estimated based on a data set
- They have statistical variability
- Variability is controlled by the amount of data
- Important fact: more data improves the precision of the ROC and AUC estimators. It does not improve prediction!

Bootstrapping ROCs and AUCs

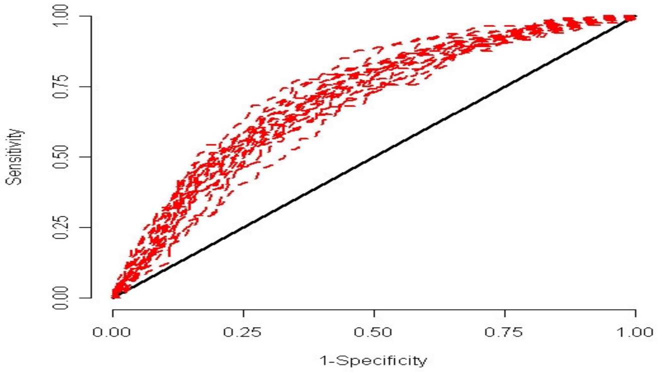
- Have a method for estimating ROC, AUC from data
- Bootstrap subjects nonparametrically (say 10,000 times)
- Repeat the estimation procedure for each data set
- Report the bootstrap distribution of ROCs and AUCs

```
for (i in 1:10000)
   {boot<-sample(n,replace=TRUE)}
```

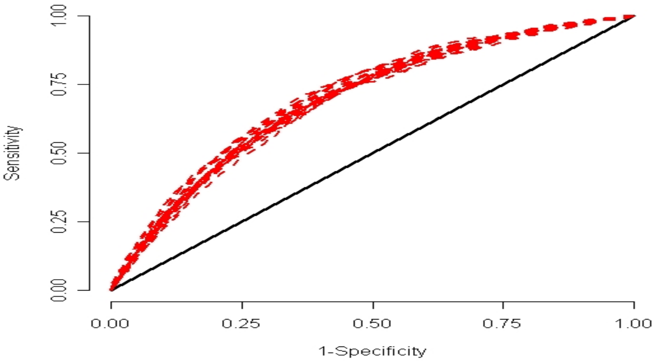
ROC


Sensitivity
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
1-Specificity

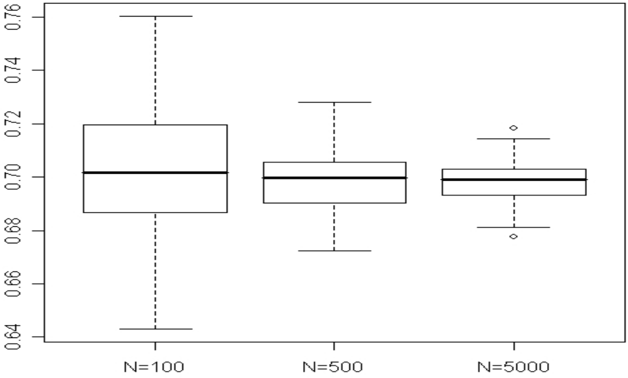
ROC


Sensitivity
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
1-Specificity

ROC


Sensitivity
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
1-Specificity

ROC



Lessons

- Variability can be very large even for large data sets
- Variability can be mistaken for signal
- This can lead to spurious, irreproducible results

"As reviewer of grants dedicated to discovery of novel biomarkers, I cannot believe how often the emphasis is on p-values (statistical significance) and not on predictive measures (predictive performance)"