9. Conditional Probability

Readings: Rosner: 3.6-3.9
OpenIntro Statistics: 2.2

R packages: ggplot2, Epi, ROC

Review: Lecture 3 Section F, Lecture SA3 Intro Probability Sections C and D

Homework: Homework 3 due by 11:59 pm on September 24
Homework 4 due by 11:59 pm on October 1
Homework 5 due by 11:59 pm on October 8

Overview

A) Definitions
B) Sensitivity and Specificity
C) Bayes' Theorem
D) Positive and Negative Predictive Values
E) Receiver Operating Characteristic (ROC) Curves

A) Definitions

Conditional Probability: determined by fixing one variable.

Probability of B, given A = P (B|A) = $\frac{P(B \cap A)}{P(A)} = \frac{P(B)P(A|B)}{P(A)}$

First we limit our sample space to A, then find the fraction of P(B) that is contained in P(A) – see smoking and respiratory problems example, Lecture 3

In-class example:

Total Probability Rule (Lecture SA3, Section D):
A very useful result for computing marginal (overall or unconditional probabilities). The total (unconditional probability) of event B is:

$$P(B) = \sum_{i=1}^{k} P(B \cap A_i) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

In-class example:

B) Sensitivity and Specificity

Sensitivity and specificity are attributes of screening and diagnostic tests. These do not depend on the prevalence of the disease in the population and are usually estimated from studies with a large number of cases with and without disease.

Motivating example: Assume there are two ways of diagnosing coronary heart disease (CHD): [1] angiogram (gold standard) and [2] treadmill test (new test). How good is the treadmill test as an approximation (it's cheaper and easier to administer) to the gold standard angiogram?

For notation, let D = CHD, $\overline{D}$ = no CHD,
T = treadmill test positive for CHD, $\overline{T}$ = treadmill test negative for CHD

For these calculations, assume we designed our study to enroll 50 with D and 50 with $\overline{D}$. In other words, we've fixed the number with CHD based on the angiogram and those without CHD based on the angiogram, so any estimate of the prevalence of CHD from our sample will not reflect the population (unless the prevalence is actually 0.5).

Sensitivity (True Positive Rate, TPR): the probability that a test will indicate 'disease' among those with the disease (i.e., how good is test at detecting (ruling in) disease when disease is there?):

$$\text{Sensitivity} = P(T|D) = \frac{P(T \cap D)}{P(D)} = \frac{P(\text{treadmill test} + \cap \text{CHD})}{P(\text{CHD})}$$

$P(D)$ is the prior or pretest probability of disease
1 – Sensitivity = 1 – True Positive Rate = False Negative Rate (FNR)

Specificity (True Negative Rate, TNR): the probability that a test will *not* indicate 'disease' among those without the disease (i.e., how good is test at ruling out disease when disease is not there?):

$$\text{Specificity} = P(\bar{T}|\bar{D}) = \frac{P(\bar{T} \cap \bar{D})}{P(\bar{D})} = \frac{P(\text{treadmill test} - \cap \text{no CHD})}{P(\text{no CHD})}$$

1 – Specificity = 1 – True Negative Rate = False Positive Rate (FPR)

The 2x2 table, confusion matrix, test by gold standard, etc.

Test	Gold Standard/ Disease Status		Total
	Positive (D)	**Negative ($\bar{D}$)**	
Positive (T)	a	b	a+b
Negative ($\bar{T}$)	c	d	c+d
Total	a+c	b+d	N (a+b+c+d)

$$\text{Sensitivity} = P(T|D) = \frac{P(T \cap D)}{P(D)} = \frac{a}{a+c} \times 100$$

$$\text{Specificity} = P(\bar{T}|\bar{D}) = \frac{P(\bar{T} \cap \bar{D})}{P(\bar{D})} = \frac{d}{b+d} \times 100$$

$$\text{FNR} = P(\bar{T}|D) = \frac{P(\bar{T} \cap D)}{P(D)} = \frac{c}{a+c} \times 100$$

$$\text{FPR} = P(T|\bar{D}) = \frac{P(T \cap \bar{D})}{P(\bar{D})} = \frac{b}{b+d} \times 100$$

Assume 100 participants were observed and their results for the treadmill test and the angiogram were recorded such that we observe:

Treadmill Test	Angiogram (GS)	
	Positive (H)	**Negative ($\bar{H}$)**
Positive (T)	40	5
Negative ($\bar{T}$)	10	45

What is the sensitivity (TPR) and how do we interpret it?

Interpretation: If someone has CHD, there is an ____% probability that the treadmill test will be positive.

What is the specificity (TNR) and how do we interpret it?

Interpretation: If someone does not have CHD, there is a ____% probability that the treadmill test will be negative.

What are our false positive and false negative rates?

C) Bayes' Theorem (or Bayes' Rule or Bayes' Law)

Bayes' Theorem calculates the posterior probability of an event based on some prior probability by utilizing conditional probabilities.

The theorem shows how to take prior probabilities (e.g., assumed prevalence of disease), incorporate new information (e.g., diagnostic test results), and obtain revised (posterior) probabilities (e.g., predictive values).

D_i represents *i* mutually exclusive and exhaustive disease states where *i* ranges from 1,…, k
T represents positive test/symptom

$$P(D_i|T) = \frac{P(T \cap D_i)}{P(T)} = \frac{P(T|D_i)P(D_i)}{\sum_{i=1}^{k} P(T|D_i)P(D_i)},$$

where the first equality is by the definition of conditional probability and the second is by the definitions of joint probability and total probability.

Suppose a patient arrives to have a diagnostic test administered. Our prior probability that the patient has disease is the prevalence of the disease $P(D)$. The result of the test should alter this to $P(D|T)$ if the test is positive or to $P(D|\overline{T})$ if the test is negative.

The effectiveness of diagnostic tests is described by sensitivity, $P(T|D)$, and specificity, $P(\overline{T}|\overline{D})$. ***Recall, these do not depend on the prevalence of the disease*** and are usually estimated from studies with a large number of persons with and without the disease.

Bayes' theorem shows how to relate them to *predictive values* using the actual population prevalence.

D) Positive Predictive Value (PPV or PV^+) and Negative Predictive Value (NPV or PV^-): an application of Bayes' theorem in clinical epidemiology

The *predictive values* of a test depend on sensitivity, specificity, and the underlying prevalence or prior probability of having the disease.

Through Bayes' rule we can relate sensitivity, specificity, and the underlying prevalence or prior probability of disease to predictive values using the actual population prevalence.

The usual scenario is we have sensitivity, specificity, and assume an underlying prevalence (*prior probability*) to obtain positive and negative predictive values, PPV and NPV.

Note: PPV and NPV are ***very*** dependent on prevalence. Using study prevalence is **not** appropriate when calculating their values, unless it's from a large prospective study from which prevalence can be well estimated.

Positive predictive values are calculated using:

$$PPV = P(D|T) \overset{\text{def1}}{=} \frac{P(T \cap D)}{P(T)} \overset{\text{def2}}{=} \frac{P(T|D)P(D)}{P(T \cap D) + P(T \cap \bar{D})} = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})}$$

Where def1 is by the definition of conditional probability and def2 is by the total probability rule. Note that our final statement has terms we've already defined:

$$\frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})} = \frac{(\text{sensitivity})(\text{prior probability})}{(\text{sensitivity})(\text{prior prob}) + (1 - \text{specificity})(1 - \text{prior prob})}$$

Negative predictive values are calculated using:

$$NPV = P(\bar{D}|\bar{T}) \overset{\text{def1}}{=} \frac{P(\bar{T} \cap \bar{D})}{P(\bar{T})} \overset{\text{def2}}{=} \frac{P(\bar{T}|\bar{D})P(\bar{D})}{P(\bar{T} \cap D) + P(\bar{T} \cap \bar{D})} = \frac{P(\bar{T}|\bar{D})P(\bar{D})}{P(\bar{T}|D)P(D) + P(\bar{T}|\bar{D})P(\bar{D})}$$

Again, we can note that these terms were already defined earlier in the lecture:

$$\frac{P(\bar{T}|\bar{D})P(\bar{D})}{P(\bar{T}|D)P(D) + P(\bar{T}|\bar{D})P(\bar{D})} = \frac{(\text{specificity})(1 - \text{prior probability})}{(1 - \text{sensitivity})(\text{prior prob}) + (\text{specificity})(1 - \text{prior prob})}$$

Using our earlier results of the sensitivity of treadmill test = 0.8 and specificity of the treadmill test = 0.9, and assuming P(CHD) = 0.2 (20% CHD in population being tested), our PPV is:

$$PPV = P(D|T) = \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})} =$$

Interpretation: *If* the treadmill test is positive, there's a _____% chance that an individual in the population with 20% prevalence of CHD actually *has* CHD.

And our NPV is:

$$NPV = P(\bar{D}|\bar{T}) = \frac{P(\bar{T}|\bar{D})P(\bar{D})}{P(\bar{T}|D)P(D) + P(\bar{T}|\bar{D})P(\bar{D})} =$$

Interpretation: *If* the treadmill test is negative, there's a _____% chance that an individual in the population with 20% prevalence of CHD actually does *not* have CHD.

Is the TT a good test?

Looking at the movement of $P(D)$ to PPV and of $P(\overline{D})$ to NPV is one way to frame the answer.

Likelihood Ratios (*aka* Bayes factors) are also useful here, and are defined as positive likelihood ratio (LR+) and negative likelihood ratio (LR-):

$$LR+= \frac{P(T|D)}{P(T|\overline{D})} = \frac{\text{sensitivity}}{1-\text{specificity}} = \frac{TPR}{FPR}$$

$$LR-= \frac{P(\overline{T}|D)}{P(\overline{T}|\overline{D})} = \frac{1-\text{sensitivity}}{\text{specificity}} = \frac{FNR}{TNR}$$

LR+ is the number of true positive results per false positive result. Large ratios are desirable and it serves as a prevalence-free measure of the strength of a positive test.

LR- is the number of false negative results per true negative result. Small ratios are desirable and it serves as a prevalence-free measure of the strength of a negative test.

Using the likelihood ratio of a test we are also able to take advantage of Bayes' theorem to calculate the posterior odds of disease (or no disease). Recall from Lecture 7, Section B4, we had that

$$\begin{aligned}&\text{Posterior odds of } \mathrm{H}_0 \text{ (after observing data)}\\ &= \text{Prior odds of } \mathrm{H}_0 \text{ (before observing data)} \times \frac{\mathrm{P}(\text{Data}|\mathrm{H}_0)}{\mathrm{P}(\text{Data}|\mathrm{H}_1)}\end{aligned}$$

In our context, we replace H_0 with D or $\overline{D}$.

Let $O(D)$ represent the odds of having the disease. We can note that our marginal probability of having the disease (i.e., the prevalence), $P(D)$, can be used to calculate the corresponding odds:

$$O(D) = \frac{P(D)}{1 - P(D)}$$

This can also be solved so that we can calculate the probability if we are given the odds:

$$P(D) = \frac{O(D)}{1 + O(D)}$$

Given the prevalence, the prior odds for D can be calculated as $\frac{P(D)}{1-P(D)}$ and the prior odds for $\overline{D}$ can be calculated as $\frac{P(\overline{D})}{1-P(\overline{D})} = \frac{1-P(D)}{P(D)}$.

Therefore, the posterior odds of having (or not having) the disease are:

$$\text{Posterior odds of } D = \text{Prior odds of } D \times (\text{LR}+) = \frac{P(D)}{1-P(D)} \times \text{LR}+$$

$$\text{Posterior odds of } \overline{D} = \text{Prior odds of } \overline{D} \times (\text{LR}-)^{-1} = \frac{1-P(D)}{P(D)} \times \frac{1}{\text{LR}-}$$

Where the interpretation of the posterior odds of D $[\overline{D}]$ are that, after observing the test results for an individual, we are X:1 in favor of D $[\overline{D}]$ (i.e., in favor of having [not having] the disease), where X is our posterior odds.

Odds range from 0 to ∞ and represent a ratio of outcomes.

Probabilities range from 0 to 1 and, depending on your target audience, may be more easily understood.

Given the relationship between odds and probabilities on the previous slides, we can calculate the *posterior probability* of an outcome given our observed test response and interpret it like we do other probabilities:

$$\text{Posterior probability of D} = \frac{\text{Posterior odds of D}}{1 + \text{Posterior odds of D}}$$

$$\text{Posterior probability of } \overline{\text{D}} = \frac{\text{Posterior odds of } \overline{\text{D}}}{1 + \text{Posterior odds of } \overline{\text{D}}}$$

Using our earlier results of the sensitivity of treadmill test = 0.8 and specificity of the treadmill test = 0.9, and assuming P(CHD) = 0.2 (20% CHD in population being tested), what will our LR+, LR-, posterior odds of D and $\overline{D}$, and posterior probabilities of D and $\overline{D}$ be?

LR+ for the treadmill test:

Interpretation: There are ____ true positives per one false positive treadmill test result.

LR- for the treadmill test:

Interpretation: There are ____ false negatives per one true negative treadmill test result.

Posterior odds and posterior probability of D:

Interpretation: After observing a positive TT, our odds in favor of CHD are ___:1, or the (posterior) probability of having the disease is ____% (a change from our prevalence of 20%).

Posterior odds and probability of $\overline{D}$:

Interpretation: After observing a negative TT, our odds in favor of *not* having CHD are ___:1, or the (posterior) probability of *not* having the disease is ____% (a change from our prior probability of 80% not having CHD).

Test	Gold Standard/ Disease Status		Total
	Positive (D)	**Negative ($\bar{D}$)**	
Positive (T)	a	b	a+b
Negative ($\bar{T}$)	c	d	c+d
Total	a+c	b+d	N (a+b+c+d)

$$\text{Sensitivity} = P(T|D) = \frac{P(T \cap D)}{P(D)} = \frac{a}{a+c} \times 100$$

$$\text{Specificity} = P(\bar{T}|\bar{D}) = \frac{P(\bar{T} \cap \bar{D})}{P(\bar{D})} = \frac{d}{b+d} \times 100$$

$$\text{FNR} = P(\bar{T}|D) = \frac{P(\bar{T} \cap D)}{P(D)} = \frac{c}{a+c} \times 100$$

$$\text{FPR} = P(T|\bar{D}) = \frac{P(T \cap \bar{D})}{P(\bar{D})} = \frac{b}{b+d} \times 100$$

$$\text{PPV} = P(D|T) = \frac{P(T \cap D)}{P(T)} \overset{***}{=} \frac{a}{a+b} \times 100$$

$$\text{NPV} = P(\bar{D}|\bar{T}) = \frac{P(\bar{T} \cap \bar{D})}{P(\bar{T})} \overset{***}{=} \frac{d}{c+d} \times 100$$

$$LR+ = \frac{P(T|D)}{P(T|\bar{D})} = \frac{\text{Se}}{1-\text{Sp}} = \frac{a(b+d)}{b(a+c)}$$

$$LR- = \frac{P(\bar{T}|D)}{P(\bar{T}|\bar{D})} = \frac{1-\text{Se}}{\text{Sp}} = \frac{c(b+d)}{d(a+c)}$$

****Note: This relationship for PPV and NPV with* a, b, c, d *only applies if the 2x2 table margins are NOT fixed (i.e., you just sample the population instead of setting enrollment at 50 with disease and 50 without a disease even though the prevalence is 20%).*

For a fascinating non-biomedical application of Bayes' Theorem, take a look at this page of Wikipedia and its links on the subject of Bayesian search theory for locating lost objects: http://en.wikipedia.org/wiki/Bayesian_search_theory.

http://en.wikipedia.org/wiki/USS_Scorpion_(SSN-589)

E) ROC curves – receiver operating characteristic curves*

Given some continuous test or screening measure, how can we identify the appropriate threshold to identify who is "positive" for the disease and who is "negative"?

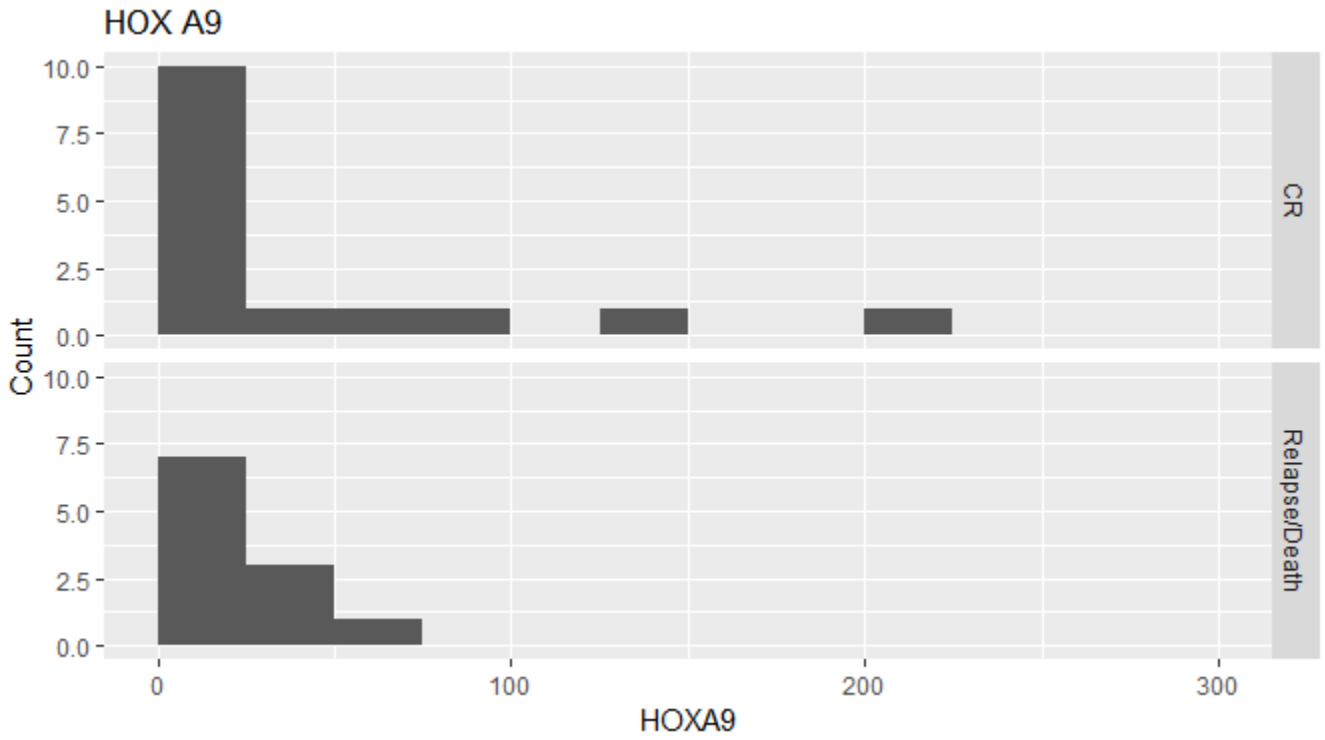
One answer is to consider that there is a natural tradeoff between sensitivity and specificity: as a test becomes more sensitive, the higher the false positive rate will be. Thus, specificity decreases. The cutoff or criterion value for a diagnostic or screening test will determine the sensitivity and specificity of the test.

ROC curves are a visual way of showing the sensitivity/specificity tradeoff as the cutoff varies. They can be used to identify optimal cutoffs, and to compare one test to another, based on the area under the ROC curve. They can also be used to compare combinations of tests.

To obtain ROC curves we take each possible cutoff value and create a 2x2 table showing the classification of positive (above the cutoff), or negative (below the cutoff) vs. the outcome of interest (e.g. gold standard: +/-, disease: yes/no, etc.). From each table the corresponding sensitivity and specificity are obtained. These values are then plotted for several values of the cutoff.

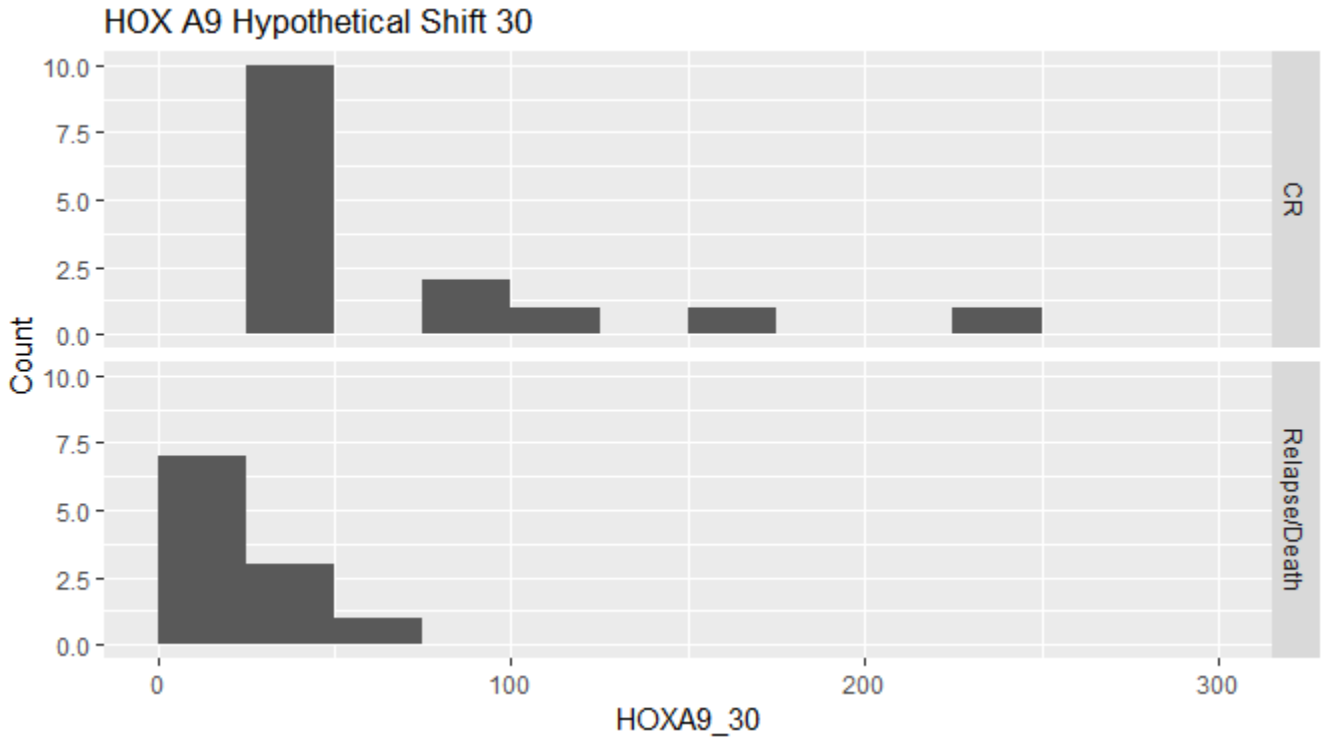
*Developed during WWII to describe how well operators of radar were at detecting blips on screen as true enemy targets, friendly vessels, or noise.

e.g. In a study of biomarkers for acute myeloid leukemia (AML), several genes in the HOX family were examined as possible predictors of response to therapy. Let's take a look at the expression (as measured by reverse transcript polymerase chain reaction: RT-PCR) of one of these genes in patients who responded to therapy vs. those who didn't.



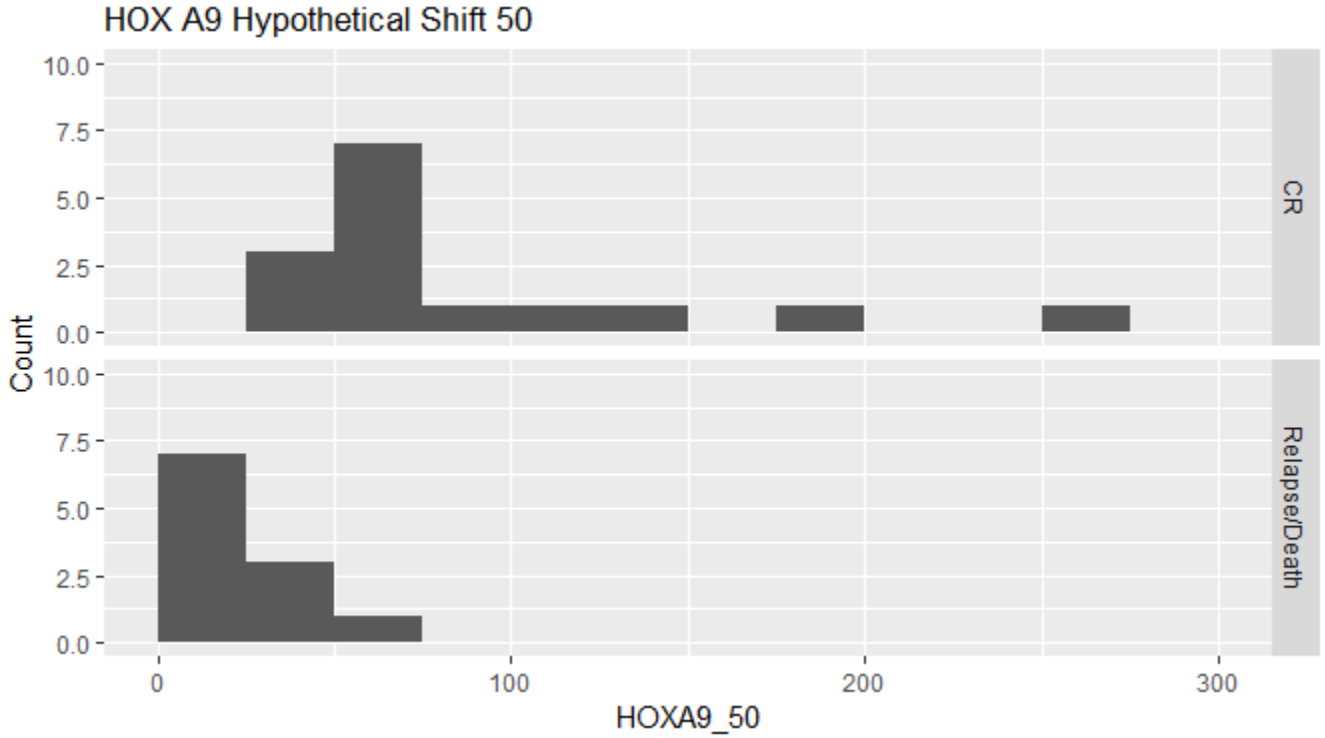
How much overlap do you see in the distribution of HOXA9 for those who had a complete response (CR) compared with those who relapsed or died? How well does HOXA9 discriminate between these two groups?

Hypothetical: If we shift the distribution of HOXA9 in the CR group by 30 expression units, what happens?



Now, how well does HOXA9 *discriminate* between the two groups?

Finally, hypothetically, let's shift the HOXA9 expression in the CR group by 50 units. How much overlap do you see?



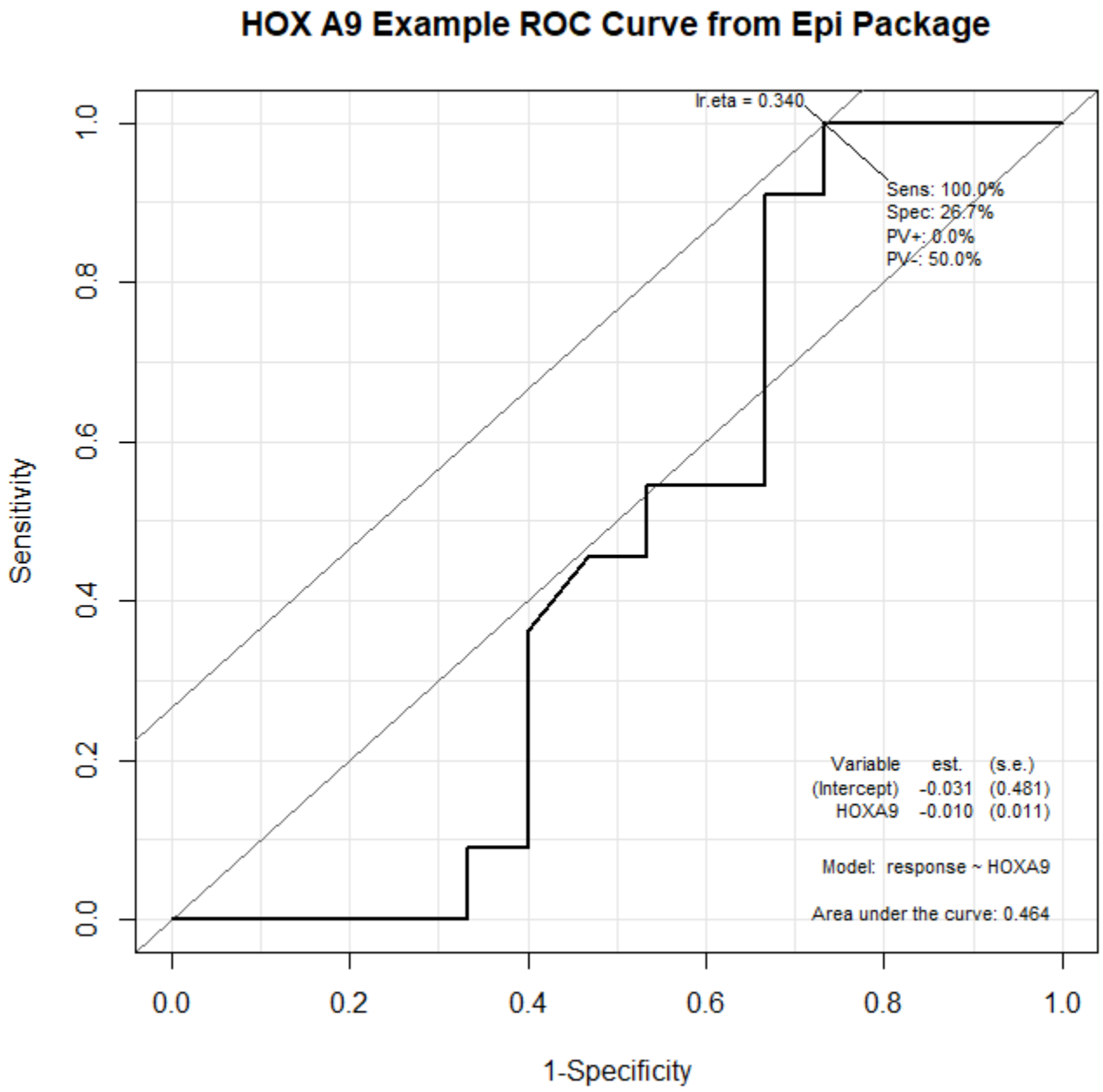
The ability of HOX A9 to discriminate between CR and relapse/death in these 3 scenarios can be summarized using an ROC curve. From an ROC curve, we can calculate the **area under the curve** (AUC).

The AUC can be interpreted as: ***the probability that, for a randomly selected pair of CR patient and patient who relapsed/died, HOX A9 will rank higher for a CR patient than for a patient who has died or relapsed.***

Tied HOX A9 values are broken by randomly assigning one observation to be a CR and one to be in the relapse/dead group, i.e. a weight of ½ goes to each observation.

One approach to calculating the AUC is by using the trapezoidal rule.

AUC can range from 0 to 1. A test with no better accuracy than random chance has an AUC=0.5. A test with perfect accuracy has an AUC of 1. If AUC < 0.5, your model is actually worse than random chance…



Using the "Epi" package in R we can generate ROC curves from data with the ROC function.

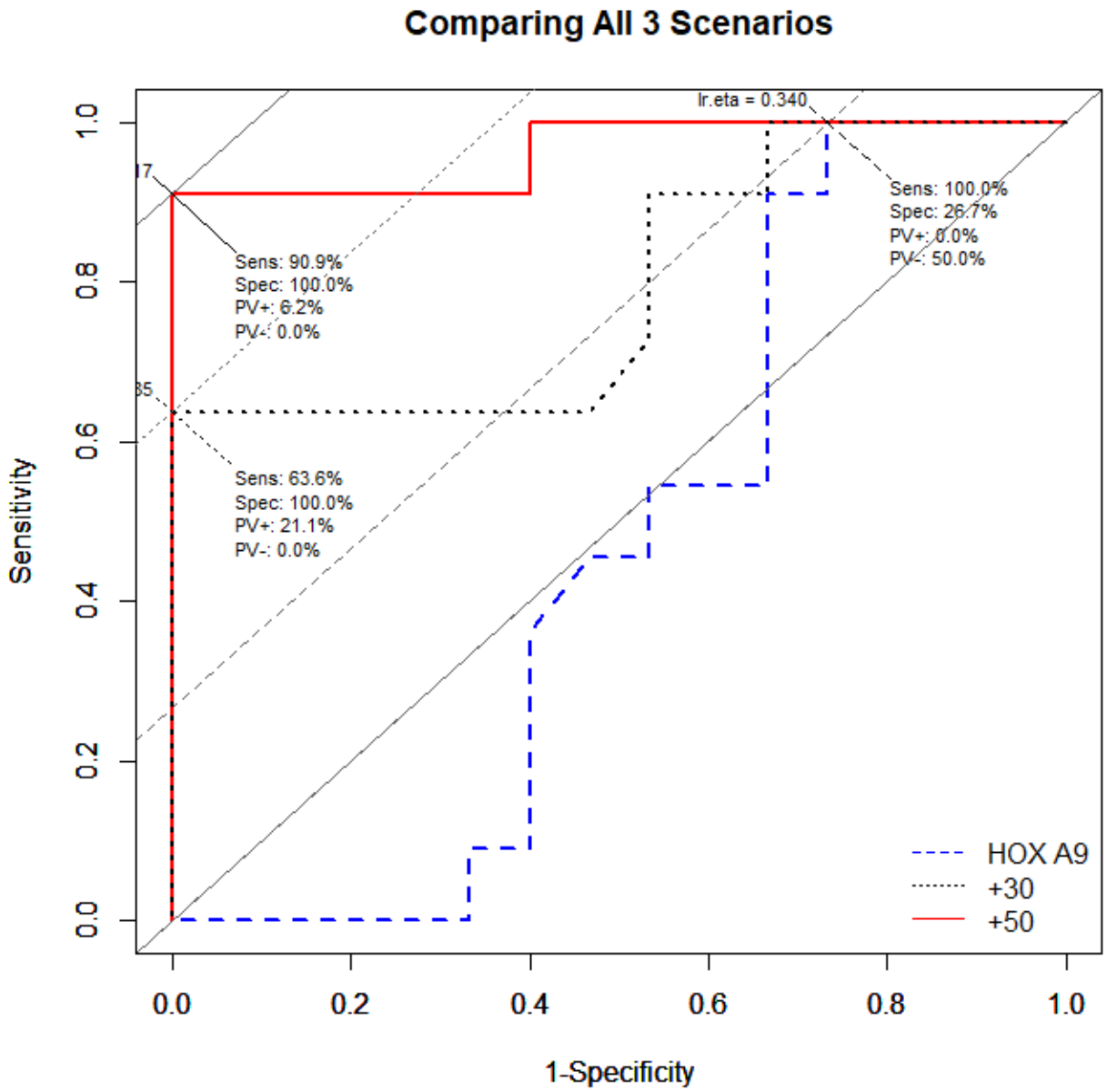
Here we see it has identified an "optimal cut point" where sens+spec is maximal.

It also includes a model summary of the logistic regression model, the model itself, and the AUC in the bottom right.

Since AUC<0.5, we should probably flip some (fair) coins instead…

Do you see the relationship between the AUC and the amount of overlap in the distribution of HOXA9 in the two patient groups?

Note: The AUC is also known as: the "c-index" (concordance index); a measure of discrimination; and the "predictive accuracy" of the biomarker. The statistic is equivalent to a non-parametric (rank-based) statistic that we'll talk about in a couple of weeks: the Mann-Whitney U-statistic (*aka* the Wilcoxon rank sum test).



R code

```
library(Epi) #load library

#Read in data set HOXgenes.txt
hox <- read.table(header=T,'~/HOXgenes.txt') #available on Canvas site

#Create new variable for defined response
hox$nevent <- hox$response
hox$nevent[hox$response == 0] <- 'CR'
hox$nevent[hox$response == 1] <- 'Relapse/Death'

#Create hypothetical variables for shifts of 30 and 50 in CR group
hox$HOXA9 <- hox$HOXA9 - 0.001 #add noise to make histograms more easily plotted
hox$HOXA9_30 <- hox$HOXA9_50 <- hox$HOXA9
hox$HOXA9_30[which(hox$response==0)] <- ( hox$HOXA9 + 30 )[which(hox$response==0)]
hox$HOXA9_50[which(hox$response==0)] <- ( hox$HOXA9 + 50 )[which(hox$response==0)]

###Create ROC plot on page 22
roc1 <- ROC(form = response ~ HOXA9, data=hox, plot="ROC", main='HOX A9 Example ROC Curve from Epi Packa
ge')

#Overlay all 3 for figure on page 23
# overlaying the ROC curves
par(col = "blue", lty = 2)
ROC(form = response ~ HOXA9, data = hox, plot = "ROC", PV = T, MI = F, grid = F, AUC = F)

par(new = T, col = "red", lty = 1)
ROC(form = response ~ HOXA9_50, data = hox, plot = "ROC", PV = T, MI = F, grid = F, AUC = F)

par(new = T, col = "black", lty = 3)
ROC(form = response ~ HOXA9_30, data = hox, plot = "ROC", PV = T, MI = F, grid = F, AUC = F, main='Compa
ring All 3 Scenarios')

legend('bottomright', col=c('blue','black','red'), lty=c(2,3,1), legend=c('HOX A9','+30','+50'),bty='n')
```