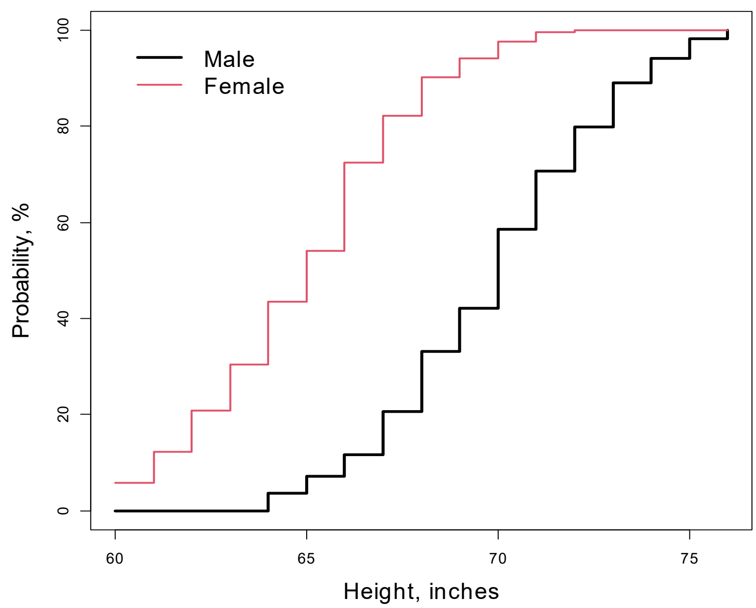
Week 6. AUC, binormal ROC curve, and LDA

Section 5.1

CDF for the uniform sample comparison

What sense do we put in by saying that men are taller than women? One store is cheaper than another? Or area A is colder than area B, etc.

How to compare samples?



Women are uniformly shorter than men because the proportion of women shorter than x is bigger than proportion of men: $F_{\text{woman}}(x) \geq F_{\text{man}}(x)$ for any x.

How to plot a cdf in R? `plot(sort(X),(1:n)/n,type="s")` or `line(sort(X),(1:n)/n,type="s")` to add the cdf on the existing plot.

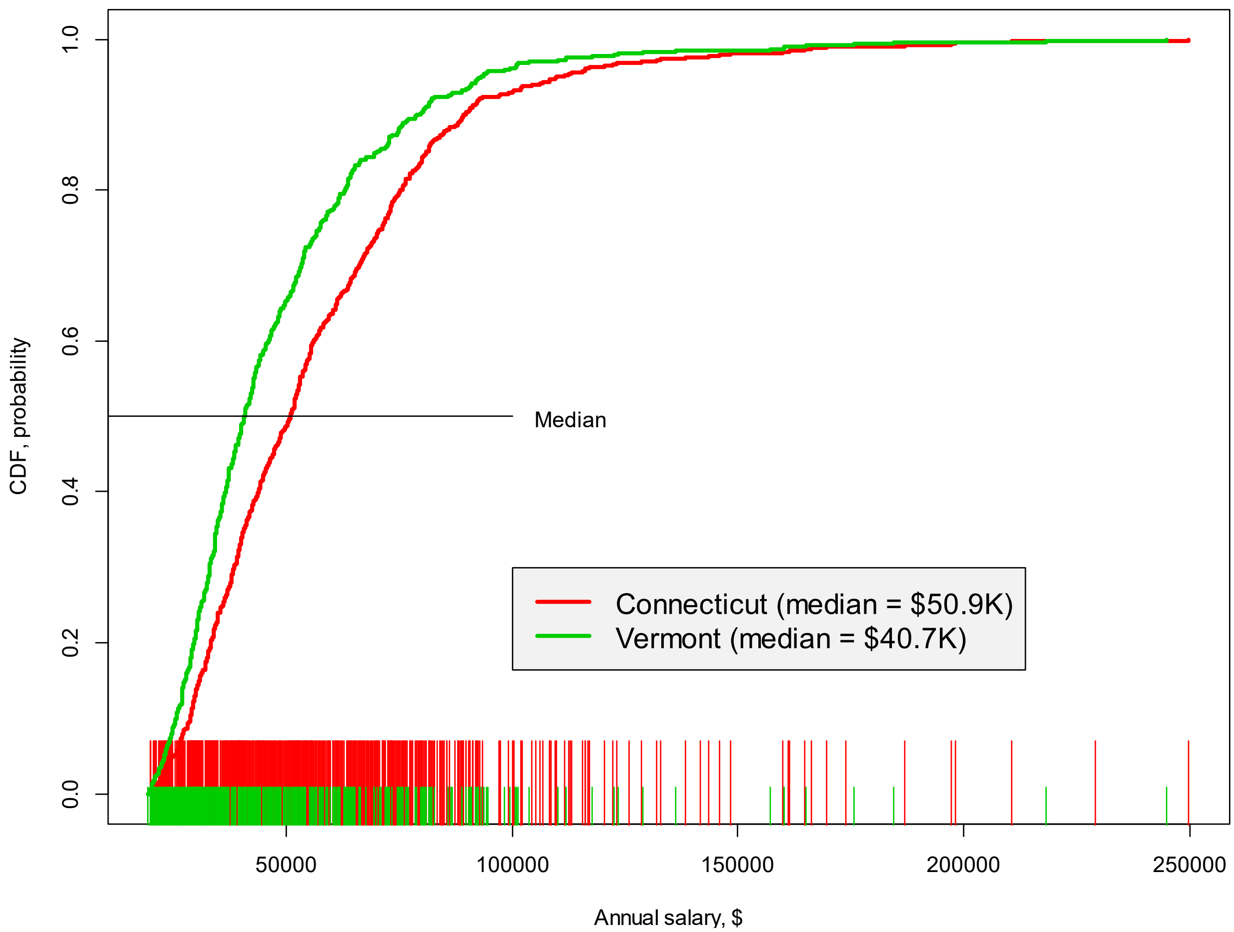
Definition 1 ***Stochastic inequality.*** *Let X and Y be two random variables with cdfs $F_X(x)$ and $F_Y(x)$. We say that Y is* ***uniformly*** *smaller than X, or symbolically, $Y \preceq X$, if $F_Y(x) \geq F_X(x)$ for all x. The same definition applies to empirical cdfs. We say that Y is uniformly smaller than X if the proportion of data Y smaller than x is bigger than the proportion of data X smaller than x for all x.*

Prove that $Y \preceq X$ implies $\text{median}_Y < \text{median}_X$, but the reverse is not true.

See my book for the proof that $Y \preceq X$ implies $\text{mean}_Y < \text{mean}_X$.

Example. *Salary comparison Vermont versus Connecticut.*

See the R code `salary`



ROC curve for quantification of $Y \preceq X$

Supervised classification problem via threshold: two samples of quantity are collected for cases (X) and controls (Y) under assumption that $Y \preceq X$. How to quantify the error of classification based on the threshold and how to choose the optimal threshold?

Example 2 *Identification of controls Y (low blood pressure) versus high blood pressure X (risk factor for heart attack). What is the threshold (x) for BP below which the patient is ok? Since in general Y have smaller BP we use the fact that $Y \preceq X$. We want to identify 'normal/control' patient, that is, the patient who will never face heart attack (correct identification).*

Definitions:

Sensitivity=true positive=correct identification of 'control' individuals.

False negative=1-Sensitivity=incorrect identification of control: based on his/her BP we predict 'no heart attack' but the patient gets a heart attack

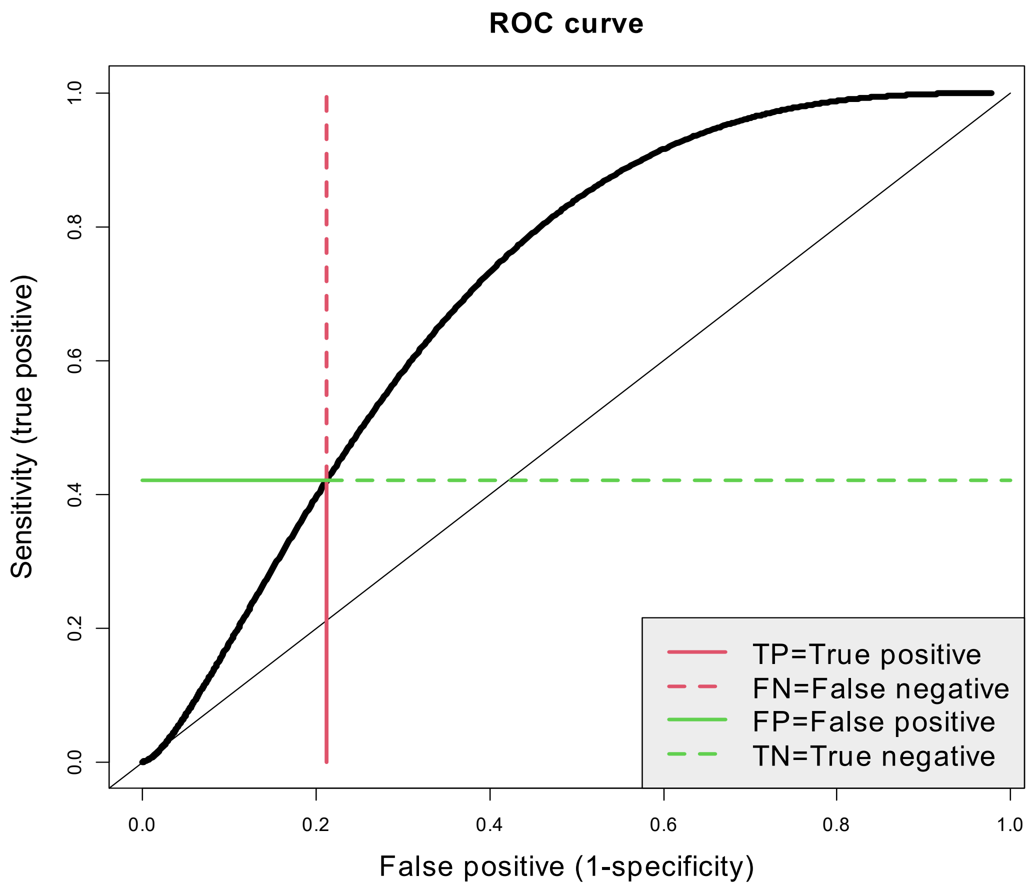
Specificity=correct identification of case, 'heart attack'

False positive=1-Specificity=incorrect identification of case: based on his/her BP we predict 'heart attack' but the patient will never gets heart attack

If the threshold is x we say that an individual is normal if his/her BP is smaller than x.

Total error = False negative + false positive
=(1-Sensitivity)+ false positive

The ROC curve is derived by plotting the sensitivity (y-axis) versus the false positive (x-axis) as the functions of the threshold which runs from $-\infty$ to $+\infty$.



Properties of the ROC curve:

1. The ROC curve can be plotted as $F_Y(x)$ on the y-axis and $F_X(x)$ on the x-axis. For empirical ROC curve x must be plotted on the union of values from X and Y.
2. It starts from (0,0) and goes up to (1,1). For empirical ROC curve it is a stepwise function.
3. It is invariant to an increasing transformation of X and Y, that is, the ROC curve build on X and Y is the same as build on $g(X)$ and $g(Y)$ where g is an increasing function, such as $\ln$.
4. The ROC curve is above the 45° if and only if $Y \prec X$, that is, $F_X(x) < F_Y(x)$ for every x.
5. Probability of correct identification, AUC=Pr($Y < X$). Interpretation: AUC is the proportion that a randomly chosen patient who will never have a heart attack has BP smaller than the a randomly chosen patient who will have a heart attack.

6. (a) The point on the curve where the tangent line has the 45° angle corresponds to the threshold, which minimizes the sum of two errors (total error = false negative+false positive). (b) This is the threshold where distance between the two cdfs is maximum, and this is the point where the two pdfs intersect.

Proof. (a) Since total error = false negative+false positive and false negative=1 - Sensitivity we have

$$\frac{d}{d\text{FP}}\text{Total error} = \frac{d}{d\text{FP}}(1 \text{ - Sensitivity+FP})$$

and

$$\frac{d}{d\text{FP}}\text{Total error} = \text{ - } \frac{d}{d\text{FP}}\text{Sensitivity} + 1 = 0$$

This implies

$$\frac{d}{d\text{FP}}\text{Sensitivity} = 1.$$

But ROC is Sensitivity = Sensitivity(FP), so that

$$\frac{d}{d\text{FP}}\text{ROC} = \frac{d}{d\text{FP}}\text{Sensitivity} = 1.$$

(b) The distance between the cdfs is Sensitivity - FP and the optimal threshold minimizes 1-Sensitivity + FP = 1-(Sensitivity - FP). Therefore it maximizes Sensitivity - FP. Since the optimal threshold is where

$$(F_Y(x) - F_Y(x))' = 0$$

where have

$$f_Y(x) = f_Y(x),$$

where f stands for the pdf.

Example 3 ***Cost reduction.*** *The cost of overlooking a future heart attack patient is \$50K (bypass surgery) and the cost of the false detection is \$1K (buy drugs and more doctor visits). Find the optimal threshold which minimizes the total cost.*

Example 4 *Identification of "low risk stroke patient" (normal, Y) versus "high risk stroke patient" (case, X) using his/her blood pressure. What is the threshold (x) below which we say that the individual is normal, that is, is no risk of stroke? What is an optimal x?*

Definition 5 *The ROC curve is the plot of one cdf versus another: when $Y \preceq X$ (or close to this): plot F_Y on the y-axis and F_X on the x-axis.*

Definitions:

Sensitivity=true positive=TP=correct identification of Y
False negative=1-Sensitivity=FN=incorrect identification of Y
Specificity=true negative=TN=correct identification of X
False positive=1-Specificity=FP=incorrect identification of X

In Example 4 the goal is to identify normals (# means the 'number'**)**

TP=proportion of people with blood pressure $\leq x$ among normal (no stroke) individuals = #(BP $\leq$ x & Normal) / #Normal

FN=proportion of people with blood pressure $> x$ among normal (no stroke) individuals = #(BP $>$ x & Normal) / #Normal

TN=proportion of people with blood pressure $> x$ among patients with stroke = #(BP $>$ x & Stroke) / #Stroke

FP=proportion of people with blood pressure $\leq x$ among patients with stroke = #(BP $\leq$ x & Stroke) / #Stroke

$$\mathtt{FN} = 1 - \mathtt{TP}, \qquad \mathtt{FP} = 1 - \mathtt{TN}$$

Definition 6 *At each threshold x we have*

$$\textit{Sensitivity} = \frac{TP}{TP+FN}, \quad \textit{Specificity} = \frac{TN}{TN+FP}.$$

Remark 7 *Sometimes, it's more convenient to reverse the identification, say, identify patients with stroke. Then to make the ROC an increasing function we plot*

$$1 - F_Y(x) = \Pr(Y > x) \textit{ versus } 1 - F_X(x) = \Pr(X > x).$$

AUC=Area Under ROC Curve

$$\mathrm{AUC} = \Pr(Y < X)$$

Proof. Let $f_X(x)$ and $f_Y(y)$ be pdfs of X and Y that are independent. Then

$$\Pr(Y < X) = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{x} f_Y(y)dy \right) f_X(x)dx = \int_{-\infty}^{\infty} F_Y(x) f_X(x)dx.$$

Now take change of variable, $F_X(x) = p$ that implies $f_X(x)dx = dp$ and $x = F_X^{-1}(p)$. This yields

$$\Pr(Y < X) = \int_0^1 F_Y(F_X^{-1}(p))dp = \int_0^1 R(p)dp,$$

the area under the ROC curve, where $R(p)$ is interpreted as sensitivity and p as false positive.

Estimation of AUC

$$\widehat{\mathrm{AUC}} = \frac{1}{nm} \sum_{i,j} 1(Y_i < X_j)$$

Interpretation: ROC=the chance that a randomly chosen observation from population Y is smaller than a randomly chosen observation from population X.

Two methods of AUC computation in R having two samples $\{Y_i, j = 1...m\}$ and $\{X_j, j = 1, ..., n\}$ and **:**

1. Sum of areas of rectangles under the ROC curve:

$$AUC = \sum_{i=2}^{n} S_i (FP_i - FP_{i-1})$$

2. Vectorized `Ylong=rep(Y,times=n);Xlong=rep(X,each=m);AUC=mean(Ylong<Xlong)`

Choosing a threshold for the optimal binary classification problem

AUC is the quality of the predictor/classifier. How to make prediction? What is the optimal threshold?

Classification problem: We are given data/sample from population Y and X. Under assumption that $Y \preceq X$, so that $F_Y(x) \geq F_X(x)$, we want to find an optimal threshold h, so that for a future observation Z for which we don't know what population it belongs the classification is: if $Z < h$ then we say that Z belongs to Y and otherwise to X. The total classification error is

$$\begin{aligned} T(h) &= 1 - \text{Sensitivity} + \text{False positive} \\ &= (1 - F_Y(h)) + F_X(h). \end{aligned}$$

The optimal threshold h is for which $T(h)$ takes minimum. The minimum exists: It is easy to see that $T(-\infty) = T(\infty) = 1$ and if $F_Y(x) \geq F_X(x)$ we have $T(h) \leq 1$ for all h.

Choosing an optimal threshold when the consequences of the two errors are not the same

To address the fact that errors have different consequences/cost we must specify the cost of two errors, w_S and w_{FP}. Then we find the threshold, which minimizes the total weighted cost:

$$\min_h \left[w_S \times (\text{1-Sensitivity}(h)) + w_{FP} \times FP(h)\right]$$

Plotting the ROC curve and the total error as a function of the threshold in R

Let Y is array/sample from the Y population and X from the X population.

```
h=sort(c(Y,X)
n=length(h)
AUC=0
sens=fp=TotEr=rep(0,n)
#default wS=wFP=1
for(i in 1:n)
{
    sens[i]=mean(Y<=h[i])
    fp[i]=mean(X<=h[i])
    if(i>1) AUC=AUC+sens[i]*(fp[i]-fp[i-1])
    TotEr[i]=TotEr[i]+wS*(1-sens[i])+wFP*fp[i]
}
h.opt=mean(h[TotEr==min(TotEr)])
```

Binormal ROC curve

The simplest ROC curve, hereafter referred to as the binormal ROC curve, is when two independent random variables are normally distributed, $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$. Without loss of generality, we can assume that $\mu_Y < \mu_X$. The inequality of the means is necessary but not sufficient to claim that $Y \prec X$. Moreover, it is easy to prove that $Y \prec X$ if and only if $\mu_Y < \mu_X$ and $\sigma_X^2 = \sigma_Y^2$. The cdfs are easily expressed through the standard normal cdf, Φ as

$$\begin{aligned} F_X(u) &= \Phi\left((u - \mu_X)/\sigma_X\right) \\ F_Y(u) &= \Phi\left((u - \mu_Y)/\sigma_Y\right) \end{aligned}$$

interpreted as the false positive (1− specificity) and sensitivity of the test. Thus the binormal ROC curve can be derived (and plotted) as a parametrically defined curve with the x-coordinate $F_X(u)$ and the y-coordinate $F_Y(u)$ when u runs from $-\infty$ to ∞. Alternatively, the binormal ROC curve can be defined as the sensitivity R expressed directly through the false positive rate p as

$$R(p) = \Phi\left(\frac{\mu_X - \mu_Y + \sigma_X \Phi^{-1}(p)}{\sigma_Y}\right), \quad 0 < p < 1.$$

Area under the binormal ROC curve in closed form:

$$\text{AUC} = \Pr(Y < X) = \Phi\left(\frac{\mu_X - \mu_Y}{\sqrt{\sigma_X^2 + \sigma_Y^2}}\right).$$

Proof. We have

$$\Pr(Y < X) = \Pr(Y - X \leq 0).$$

But

$$Y - X \sim \mathcal{N}\left(\mu_Y - \mu_X, \sigma_Y^2 + \sigma_X^2\right)$$

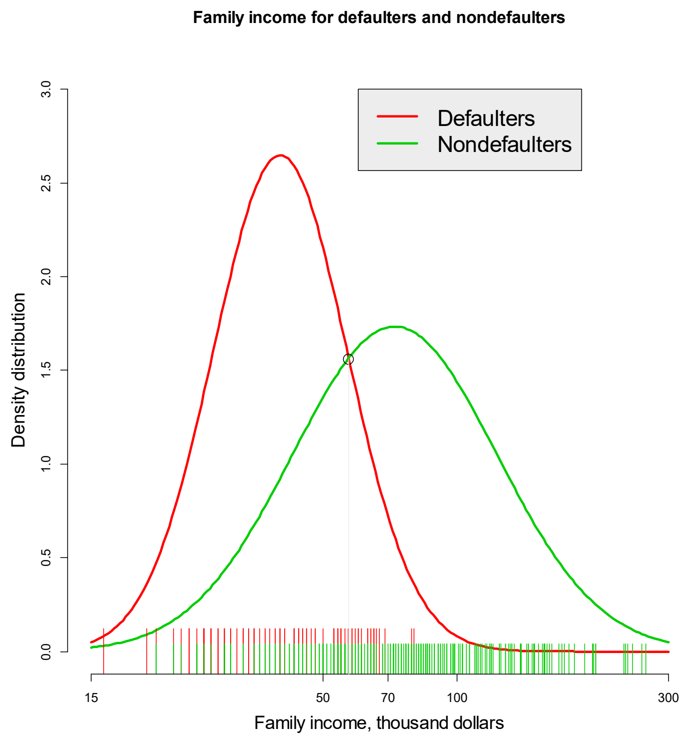
and therefore

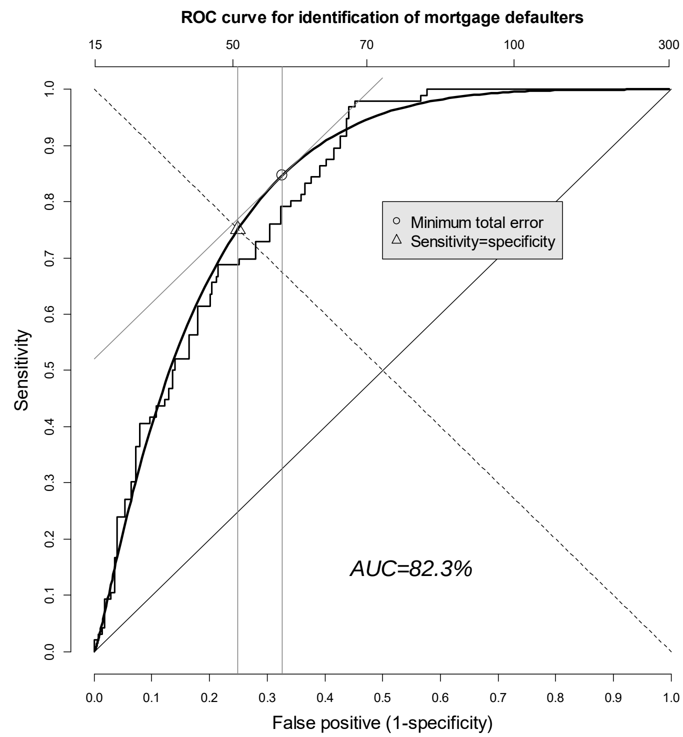
$$\Pr(Y - X \leq 0) = \Phi\left(\frac{0 - (\mu_Y - \mu_X)}{\sqrt{\sigma_X^2 + \sigma_Y^2}}\right) = \Phi\left(\frac{\mu_X - \mu_Y}{\sqrt{\sigma_X^2 + \sigma_Y^2}}\right).$$

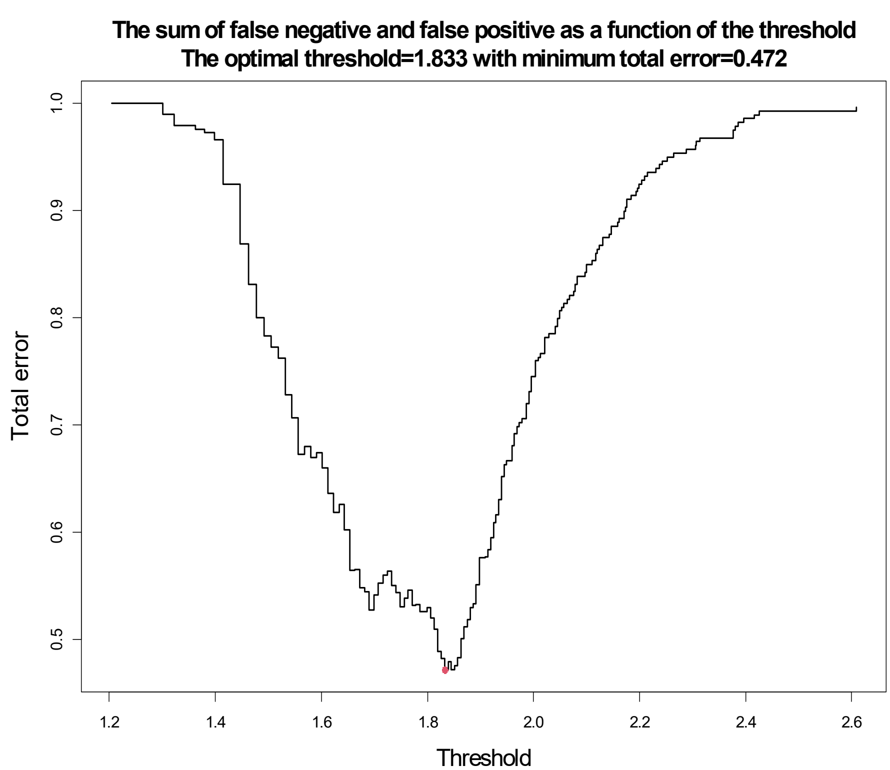
If Y and X are samples we estimate μ_X and μ_Y by the means and σ_X^2 and σ_Y^2 by respective variances.

Example: mortgageROC

See Example 5.5
   `mortgageROC.r` and `mortgageROC.csv`







How to compute/estimate AUC and optimal threshold?

Nonparametric and parametric statistical methodology

AUC

Empirical: add areas of small rectangles under the ROC curve

Theoretical/smoothed: Apply transformation for samples Y and X to make them close to normal, say, log-transformation. Then estimate

$$\begin{aligned}\widehat{\mu}_X &= \overline{X}, \widehat{\mu}_Y = \overline{Y}, \\ \widehat{\sigma}_X^2 &= \frac{1}{n_X - 1}\sum_{i=1}^{n_X}(X_i - \overline{X})^2, \widehat{\sigma}_Y^2 = \frac{1}{n_Y - 1}\sum_{i=1}^{n_Y}(Y_i - \overline{Y})^2\end{aligned}$$

and then

$$AUC = \Phi\left(\frac{\widehat{\mu}_X - \widehat{\mu}_Y}{\sqrt{\widehat{\sigma}_X^2 + \widehat{\sigma}_Y^2}}\right).$$

Optimal threshold

Empirical: after plotting the ROC curve plot (`type="s"`)

Total cost = (1-Sensetivity) + False positive

versus the threshold and find at what threshold value it takes minimum.

Theoretical/smoothed: Obtain binormal ROC curve as above and compute the optimal threshold as

$$\frac{A - \sqrt{A^2 - BC}}{B}$$

where

$$A = \mu_X\sigma_Y^2 - \mu_Y\sigma_X^2,\ B = \sigma_Y^2 - \sigma_X^2,\ C = \sigma_Y^2\mu_X^2 - \sigma_X^2\mu_Y^2 + \sigma_X^2\sigma_Y^2\ln(\sigma_X^2/\sigma_Y^2),$$

as in the book.

Homework 6

Presentation matters, display all necessary information on the graphs.

1. (10 points). Problem 5.5 (page 309).
2. (20 points). File `bp.csv` contains blood pressure (BP) for normal patients (controls, high=0) and hypertension patients (high=1). (a) Display two cdfs to demonstrate that BP among normal patients is uniformly smaller than among hypertension patients. (b) Display the data-driven ROC curve for the identification of normal patients and the superimposed binormal counterpart. (c) Compute and display AUCs using three methods: (1) empirical, as the sum or rectangles, (2) empirical, using vectorized computation, and (3) theoretical, using the formula. (d) The cost associated with overlooking a hypertension patient is \$10K and the cost of the false identification of hypertension is \$1K. Display the data-driven total cost and the superimposed continuous counterpart as a function of the threshold along with the respective optimal thresholds. (d) Display the optimal threshold on the binormal ROC curve and the respective BP scale using `axis(side=3)`.