Week 2. Graphics animation in R. Cdf-based stochastic inequality and ROC curve

My first statistical animation – cdf plot of the lognormal distribution. Stochastic inequality between distributions/samples and their comparison using cumulative distribution function (CDF). Receiver Operator Characteristic (ROC) curve and their animation in R. False positive and false negative rates for the BMI-based heart attack prognosis. Binormal ROC curve and misclassification error. Area under ROC curve (AUC), its interpretation and estimation. Optimal threshold under various cost scenarios.

R code: `cdf.dyn`, `salary`, `bph.ROC`

Data: `Vermont.txt, Connecticut.txt, bp.csv, mortgageROC.csv`

R animation for illustration of cdf

Theoretical cdf

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Empirical cdf

$$\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i \le x) = \text{ the proportion of observations less or equal to } x.$$

This estimator is unbiased and consistent.

Run `cdf.dyn(job=1)`

If X has a lognormal distribution, that is, $X = e^Y$ where $Y \sim \mathcal{N}(\mu, \sigma^2)$ then the cdf of X is

$$F_X(x) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right) = \texttt{pnorm}(\log(\texttt{x}), \texttt{mean} = \mu, \texttt{sd} = \sigma)$$

ImageMagic software download: *https://imagemagick.org/script/download.php*

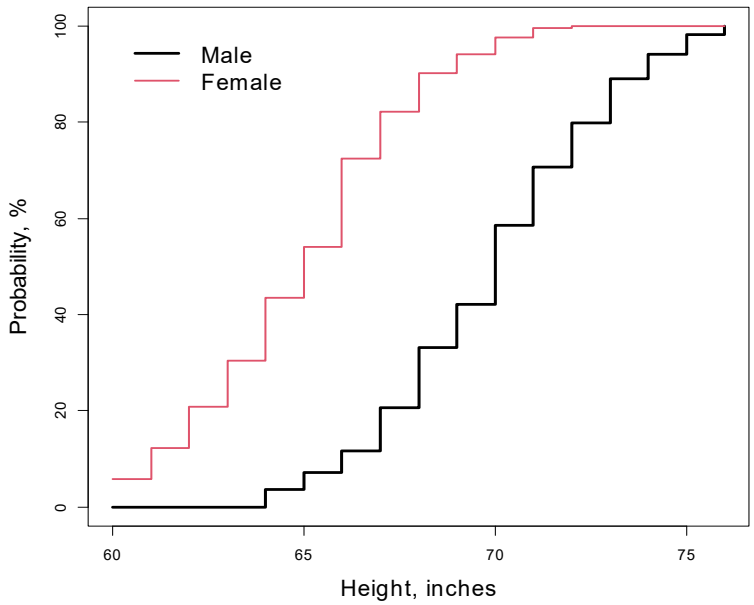
CDF for the uniform sample comparison

Section 5.1

How to compare samples using the cumulative distribution function (cdf)?

- What we mean by saying that the prices in one grocery store are higher than in the other?
- What we mean by saying that females are shorter than males?
- What we mean by saying that drug A is better than drug B in terms of time to relapse?

Mean is not a good quantity to make judgment on what is less and what is large.
When we say that females are shorter, how to interpret the sum of heights of females?



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 all x.

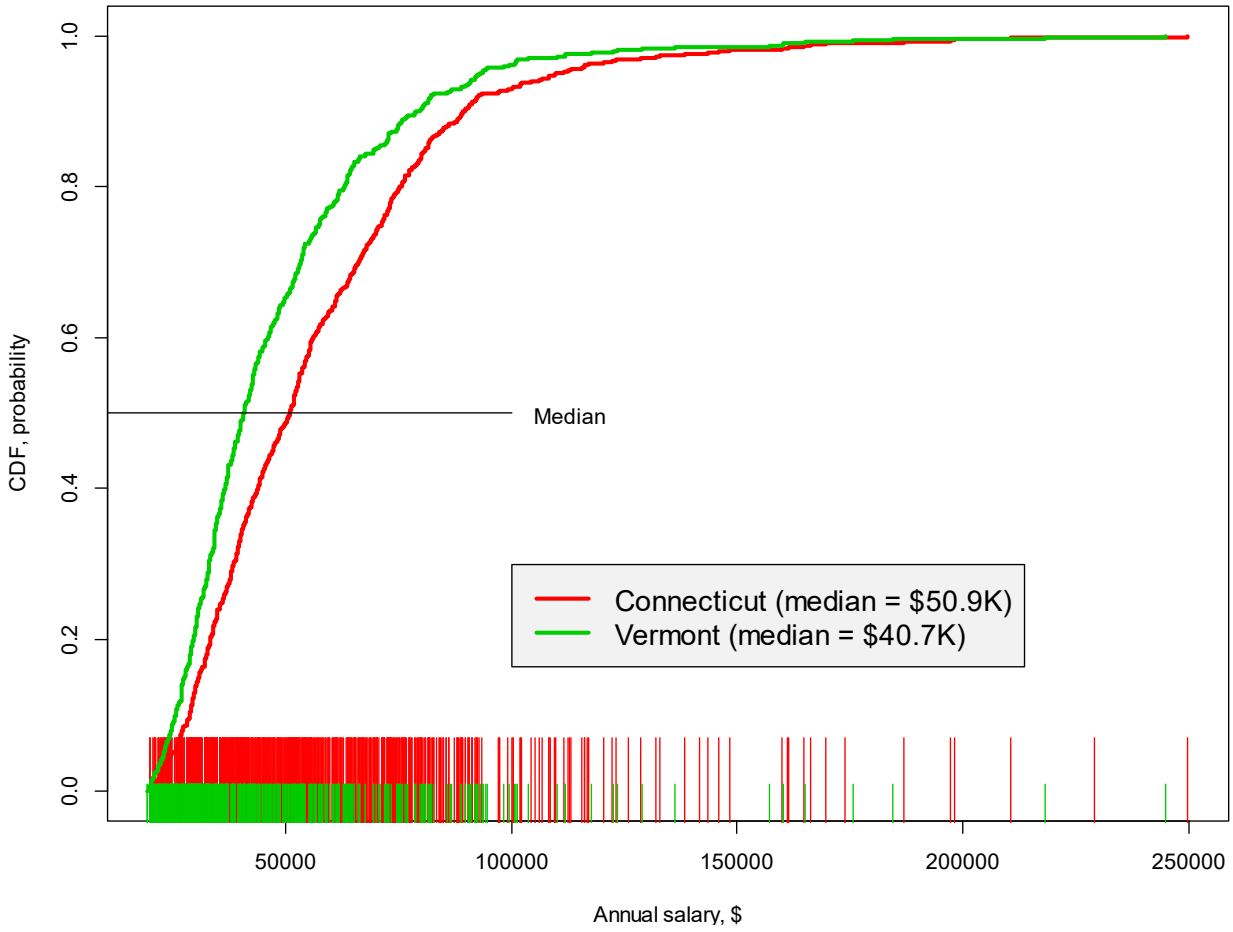
Definition 1 ***Stochastic inequality***. *Let X and Y be two random variables with cdfs $F_X(x)$ and $F_Y(x)$. We say that Y is stochastically (uniformly) smaller than X, or symbolically, $Y \preceq X$, if*

$$F_Y(x) \geq F_X(x) \text{ for all } x.$$

The same definition applies to empirical cdfs. We say that Y is uniformly smaller than Y if the proportion of data Y smaller than x is bigger than the proportion of data X smaller than x.

Y is smaller than X if the cdf of Y is above the cdf of X.

If $Y < X$ then median(Y) < median(X) and $\overline{Y} < \overline{X}$.
Run `cdf.dyn(job=2)`
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 classification variable (or attribute) 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 the women gender based on height. Given a set of heights of women* $Y = \{Y_1, ..., Y_m\}$ *and a set of heights of men* $X = \{X_1, ..., X_k\}$. *How the threshold* x, *such that if height* $< x$ *say it's woman and otherwise man, affects the error of gender identification. What is the optimal threshold* x?

Rule of thumb: identify 'small value' variable.
The ROC curve distinguish two types of identification errors: false positive and false negative.

Example 3 *Identification of normal mammography: Given the size of a blob on the mammogram identify normal patient. The rule: size* $< x$ *means normal and size* $> x$ *means abnormal/case. How the choice of* x *affects sensitivity (correct identification of normal) and specificity (correct identification of case).*

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

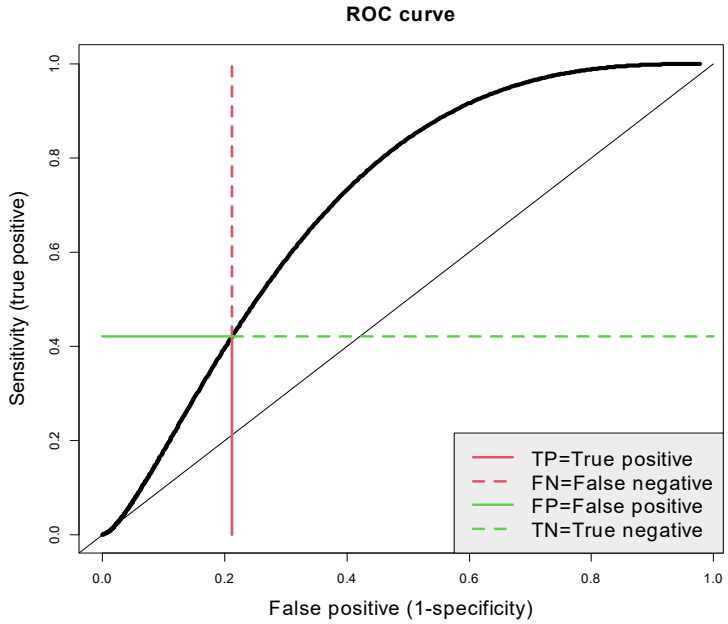
$$\texttt{FN} = 1 - \texttt{TP}, \qquad \texttt{FP} = 1 - \texttt{TN}$$

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*

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



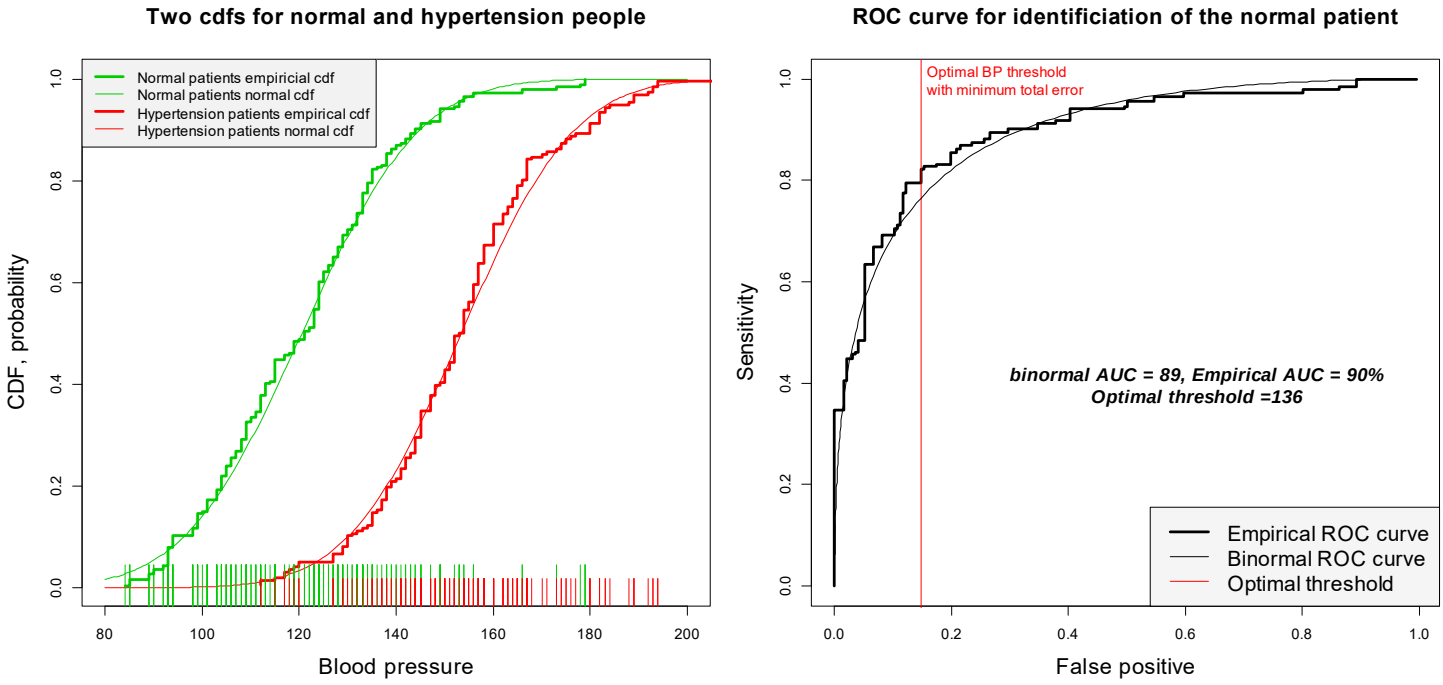


Figure 1:

Example 8 *File* `bp.csv` *contains blood pressure for normal patients (controls, high=0) and hypertension patients (high=1). (a) Use par(mfrow=c(1,2)) to plot the two cdfs at left and the ROC curve at right. Use different colors to depict two groups of patients. (b) Display the data-driven cdf and ROC curve and the binormal counterparts by estimating the means and SDs using* `mean` *and* `sd` *respectively. (c) Compute and display AUC. (d) Use axis(side=3) to display the corresponding threshold. (e) Compute and display as the vertical line the threshold which minimizes the sum of two errors.*

Solution. See the R function `bph.ROC`

Example 9 ***Statistical animation for cdf comparison and ROC curve.*** *Run* `cdf.dyn(job=2)` *and* `cdf.dyn(job=3)`

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

Total error = False negative + False positive

The ROC curve is derived by plotting Sensitivity (y-axis) versus False positive (x-axis) when the threshold 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 the union of values from X and Y.

2. The ROC curve starts from (0,0) and goes up to (1,1), that is the ROC curve is not a decreasing function at any point. The empirical ROC curve is a stepwise function.

3. ROC curve is invariant with respect to any 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. 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 positive + false negative).

How to choose an optimal threshold?

Theorem 10 *(a) The total sum of errors, FP+FN, is minimized for the threshold for which the tangent line on the ROC curve has the 45° slope. (b) The point where FP=FN is the intersection of the -45° line with the ROC curve.*

Proof. (a) Sensitivity, TP=ROC(FP). Since FP=1-TP we have to minimize FP+ (1-ROC(FP)). Differentiate with respect to FP

$$\frac{d}{dFP}(FP + (1 - ROC(FP))) = 1 - \frac{d}{dFP}ROC = 0$$

which implies

$$\frac{d}{dFP}ROC = 1.$$

(b) is obvious.

Homework 2

Use `mortgageROC.csv` data from Example 5.5 of the book.

1. (20 points). Plot the two empirical cdfs with family income on the log scale with actual numbers displayed in thousand dollars. Use `rug`, `legend`, and different colors.

2. (20 points). Create an R animation where at left you show empirical cdfs from the previous task and the growing stepwise ROC curve at right (use `type="s"`) as in `cdf.dyn(job=3)`. Submit as a standalone `gif` file.

3. (20 points). (a) Compute AUC and provide its layman interpretation. (b) Compute and provide a layman interpretation for True Positive, True negative, False Positive, and False negative when Sensitivity=0.8. (c) Compute and display the optimal threshold if the cost of overlooking a future defaulter is \$200K and the cost of denying the mortgage application who will not default in the future is \$100K.