Stat 8003, Homework 2

Group G: Andrew Schneider, Abdulsalam Hdadi, David Dobor

September 8, 2014

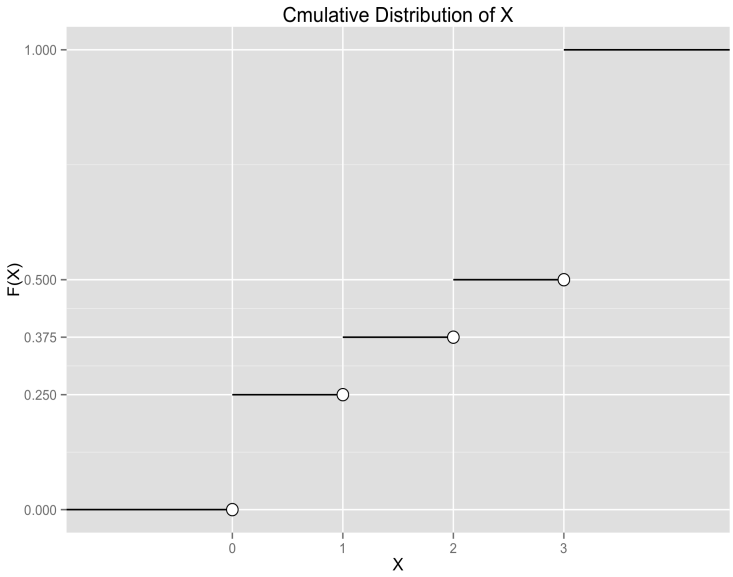
Question 2.1. Suppose that X is a discrete random variable with $P(X = 0) = 0.25$, $P(X = 1) = 0.125, P(X = 2) = 0.125,$ and $P(X = 3) = 0.5,$ Calculate the cdf of X and graph the `cdf` using R.

Answer: The `cdf` of X:

$$F(X) = \begin{cases} 0 & x < 0 \\ 0.25 & 0 \leq x < 1 \\ 0.375 & 1 \leq x < 2 \\ 0.5 & 2 \leq x < 3 \\ 1 & 3 \leq x < \infty \end{cases}$$

This right-continuous function changes it's value at each of it's 'jump points' by exactly the value of it's `pmf` at the same point. For example, the jump at point $x = 2$ is $0.5 - 0.375 = 0.125$, which is the same as the value of `pmf` at $x = 2$: $P(X = 2) = 0.125$.

This `cdf` looks like this:



Question 2.2. A certain type of cancer is known to be present in 2 percent of the population of males in their fifties. A test for the disease is advertised by a pharmaceutical company to be 3% false negative and 1% false positive.

1. Compute the probability that you have cancer if you are tested positive.
2. To make sure that you really have cancer an invasive and expensive surgery is needed. Your health insurance company is not willing to pay for this unless the pharmaceutical company improves its test in such way that that at least 90% of people who are tested positive actually have the disease. How low should be the rate of false positive for the test to reach this goal? (Assume that the rate of false negative remains the same).

Answer: We'll use the following notation for the given data:

- The probability of the disease being present in the population of males over fifty: $P(D) = .02$. (Thus 98% of this population is disease-free: $P(D^c) = .98$.)
- The probability of the test indicating that there is *no disease* when in fact the disease is present (false negative): $P(T- \mid D) = .03$. (Thus the probability of a *true positive*, i.e. the test indicating that there *is* disease when in fact the disease is present, $P(T+ \mid D)$, is then 97%.)
- The probability of the test indicating that there is *disease* when the disease is *not* present (false positive): $P(T+ \mid D^c) = .01$
- We would like the probability $P(D \mid T+)$ of a person having the decease when the test result is positive to be at least 90%.

By Bayes' rule, the probability of the disease being present when the test indicates that it is present is:

$$\begin{aligned} P(D \mid T+) = \frac{P(T+ \mid D)P(D)}{P(T+)} &= \frac{P(T+ \mid D)P(D)}{P(T+ \mid D)P(D) + P(T+ \mid D^c)P(D^c)} \\ &= \frac{.97 \times .02}{.97 \times .02 + .01 \times .98} \\ &= 0.664 \end{aligned}$$

We would like to increase this probability from 66% to at least 90% by decreasing the probability of the false positives. That is, we would like to make the $P(T+|D^c)$ term in the denominator smaller. So let's solve the following inequality for x:

$$\begin{aligned} .9 &< \frac{.97 \times .02}{.97 \times .02 + x \times .98} \\ .9 &< \frac{.0194}{.0194 + .98x} \end{aligned}$$

$$
\begin{aligned}
.01746 + .882x &< .0194 \\
.882x &< .00194 \\
x &< 0.002199546
\end{aligned}
$$

Thus, the false positive rate should be smaller than about *two in a thousand.*

Question 2.3. Suppose there is a continuous random variable X with `cdf` $F(x)$. Let $Y = F(X)$. What is the distribution of Y?

Answer: First, we note that the domain of the `r.v.` Y is $[0, 1]$ (since the range of the monotonically non-decreasing $F(X)$ is $[0, 1]$). Next,

$$P(Y \leq y) = P(F(X) \leq y) = P(X \leq F^{-1}(y)) = y$$

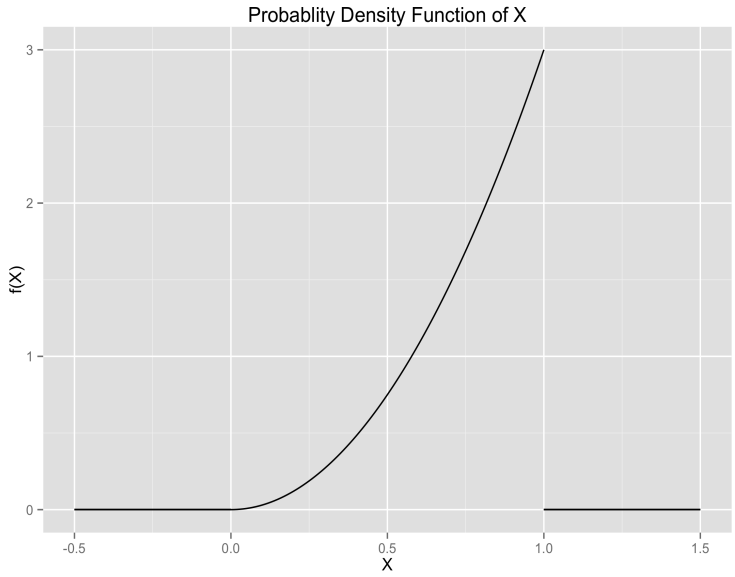
That is, Y has to be uniformly distributed over $[0, 1]$.

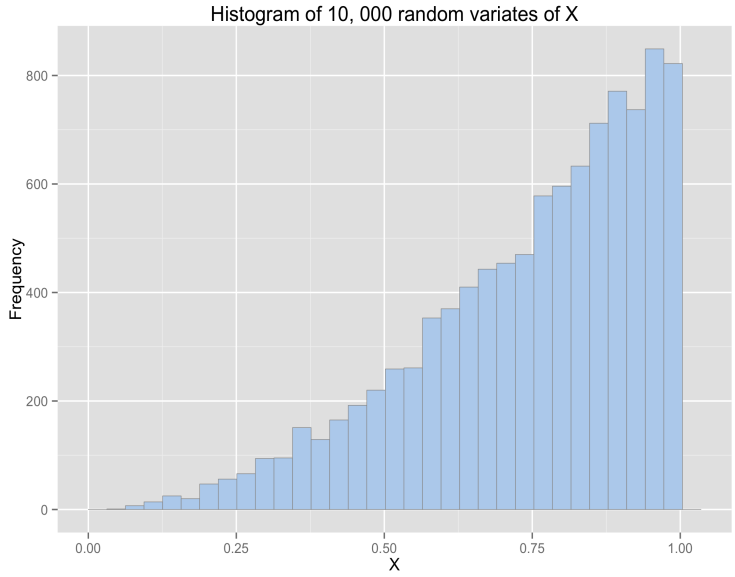
Question 2.4. Generate 10,000 random samples from a distribution with the `pdf` $f(x) = 3x^2 1(0 < x < 1)$. State your steps and program it in `R`. After generating these random numbers, plot its density and compare it with the actual pdf function. (Hint: Use the result from Problem 4? (should this be the previous problem, problem 3?.)

Answer: To generate random samples from the random variable X with `pdf` $f(x) = 3x^2 1(0 < x < 1)$, we

1. Compute the `cdf` of X, $F(X)$. This is simply $u = F(x) = 3x^2$ on the interval from $[0, 1]$ (and it's 1 whenever $x \geq 1$ and 0 whenever $x \leq 0$)
2. Then find its inverse: $F^{-1}(u) = u^{1/3}$ (where u ranges from 0 to 1)
3. Finally, we generate uniform `r.v.`s and transform them with this F^{-1} to obtain the needed variates

The `R` code that performs the above steps and generates the figures below can be found at `https://github.com/david-dobor/8003/blob/master/week2/q4hw1.R`





The shapes are similar, which is reassuring.