MATH 550: The Probability Integral Transform

Simulation and Transformation

Delivered by Peter Neal (October 2012)

Lecture materials produced by Dr Chris Sherlock.
This forms part of the MSc in Statistics at Lancaster University

Revision of Uniform(0,1)

$U \sim Unif(0,1)$ is a (continuous) random variable uniformly distributed between 0 and 1.

The pdf is $f_U(u) = 1$ if $0 \leq u \leq 1$, and $f_U(u) = 0$ otherwise.

The cdf is

$$F_U(u) = \begin{cases} 0 & \text{if} \quad u < 0 \\ u & \text{if} \quad 0 \leq u \leq 1 \\ 1 & \text{if} \quad u > 1. \end{cases}$$

Sketch these.

A coin toss

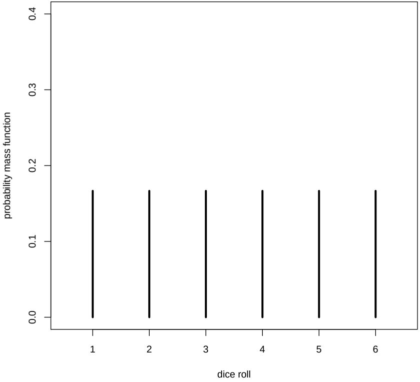
Suppose you wish to simulate the toss of a coin, but all that you are able to do is simulate uniform random numbers between 0 and 1, $Unif(0, 1)$.

You could say "if $U \leq 0.5$ then the coin is a head, and otherwise it is a tail".

Why is this reasonable?

A dice roll

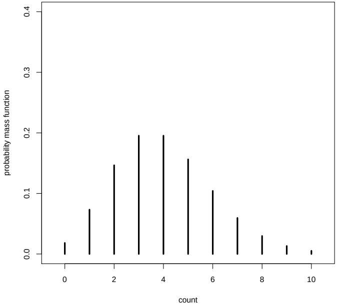
Now suppose you wish to simulate the roll of a dice D but, again, you are only able to simulate $U \sim Unif(0,1)$



Write a sensible set of conditions for D.

A Poisson(4) random variable

Next simulate from a Poisson(4) random variable P using only $U \sim Unif(0,1)$.

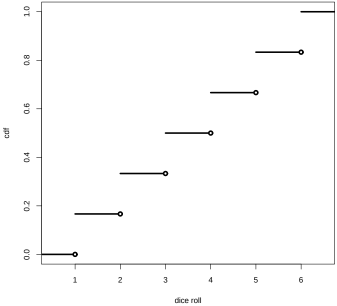


$P(P=0) \approx 0.0183$, $P(P=1) \approx 0.0733$, $P(P=2) \approx 0.1465$, $P(P=3) \approx 0.1954$ and $P(P=4) \approx 0.1954$...

Write down the decision rule for the first few values of P.

Using the CDF - dice

The CDF for a dice roll is (*Make sure you understand why!*)



In matching the probabilities we performed the following calculations

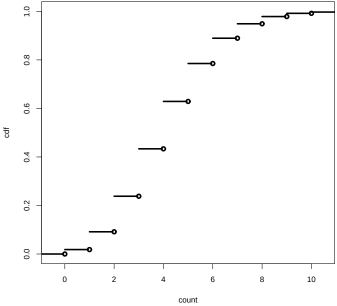
$$P(D=1) = P(D \leq 1) \quad = \quad P(U \leq 1/6)$$

$$P(D=2) = P(D \leq 2) - P(D \leq 1) \quad = \quad P(U \leq 2/6) - P(U \leq 1/6)$$

$$\cdots \qquad \cdots$$

Using the CDF - Poisson(4)

The CDF for a Po(4) r.v. is



In matching the probabilities we performed the following calculations

$$
\begin{aligned}
P(P=0) = P(P \leq 0) &= P(U \leq 0.0183) \\
P(P=1) = P(P \leq 1) - P(P \leq 0) &= P(U \leq 0.0916) - P(U \leq 0.0183) \\
P(P=2) = P(P \leq 2) - P(P \leq 1) &= P(U \leq 0.2381) - P(U \leq 0.0916)
\end{aligned}
$$

Simulating from a continuous r.v.

Above we made P and U correspond so that

$$F_P(p) = F_U(u).$$

(NB we implicitly forced the probability of non-integer P to be zero).

For a continuous random variable the probability of any given value (e.g. 0.817) is 0.

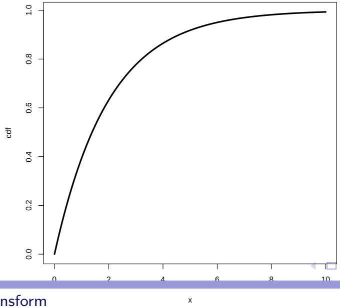
To simulate from a continuous random variable given $U \sim Unif(0,1)$ we *match the CDFs at all values*.

Simulating an Exp(2)

Suppose that you wish to simulate $Y \sim Exp(2)$ using only $U \sim Unif(0,1)$.

The pdf is $f_Y(y) = 1/2\ e^{-y/2}$ $(y \geq 0)$ and $f_Y(y) = 0$ $(y < 0)$ (*sketch it*).

The cdf is $F_Y(y) = 1 - e^{-y/2}$ $(y \geq 0)$ and $F_Y(y) = 0$ $(y < 0)$ (*check you understand this.*)



Simulating an Exp(2)

We must match the CDFs of Y and U.

$$\begin{aligned} P(Y \leq y) &= P(U \leq u) \\ F_Y(y) &= F_U(u) \\ 1 - e^{-y/2} &= u \\ y &= -2\log(1-u) \end{aligned}$$

So if for example you simulated $u = 0.843$ then your $Exp(2)$ random variable is $-2\log(1 - .843) = 3.703$.

This is (almost) exactly how `R` simulates exponential rvs.

A more complex example

The random variable X has density function

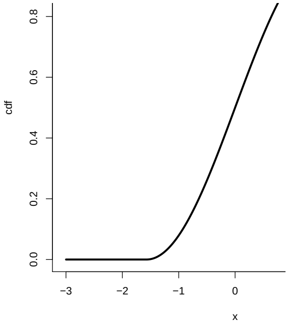
$$f_X(x) = \begin{cases} 0 & \text{if} \quad x <= -\pi/2 \\ \frac{1}{2}\cos x & \text{if} \quad -\pi/2 < x \leq \pi/2 \\ 0 & \text{if} \quad \pi/2 < x \end{cases}$$

How would you simulate from it? (Why the factor of $1/2$?).
Important: first sketch (or plot) the density function.

The cdf

The cumulative distribution function is

$$F_X(x) = \begin{cases} 0 & \text{if} \quad x \leq -\pi/2 \\ \frac{1}{2}\sin x + \frac{1}{2} & \text{if} \quad -\pi/2 < x \leq \pi/2 \quad (\textit{why?}) \\ & \text{if} \quad \pi/2 < x \end{cases}$$



Simulation

Simulate u, then set

$$\frac{1}{2} sinx + \frac{1}{2} = u \Leftrightarrow x = \sin^{-1}(2u - 1).$$

Normal distribution

Suppose that `rnorm()` is broken, but that the following functions work.

`runif()`, `dnorm()`, `pnorm()`, `qnorm()`.

How could you use two of the above functions to simulate a $Z \sim N(0, 1)$?

NB This is NOT how `R` simulates normal random variables - there are several much faster algorithms including the Box-Muller algorithm (just beyond the scope of this course).

The Probability Integral Transform

For any continuous random variable X and $k \in [0, 1]$,

$$P\left(F_X(X) \leq k\right) = k.$$

Proof *Students!*

i.e. $A := F_X(X) \sim Unif(0, 1)$.
We can convert from any continuous distribution to a uniform random variable - this transformation has the grand title of **The Probability Integral Transform**.

The method of simulation that we have studied is called the **inverse transformation method** because we have to solve $x = F_X^{-1}(u)$.

PP-plots

A student collects the following 15 data values, $\mathbf{x}$:
234,264,214,151,321,112,255,160,235,296,238,244,226,222,94.

Do they follow a Normal distribution?

If the data are Normal then $\hat{\mu} = \overline{x} = 215.9$ and $\hat{\sigma}^2 = s^2 = 4057.8$.

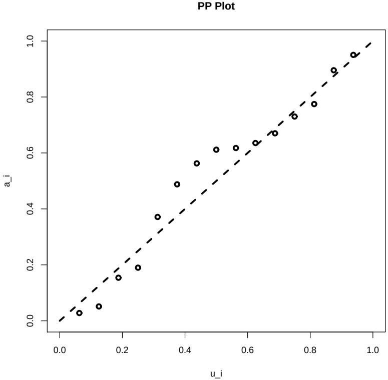
If the data are Normal then $A_i = F(X_i) \approx \Phi((X_i - \hat{\mu})/\hat{\sigma})$ should be uniformly distributed in $[0, 1]$.

Sort the a_i so that a_1 is the smallest and a_{15} is the largest.

If the a_i are uniformly distributed then this should be true:
$a_1 \approx u_1 = 1/16$, $a_2 \approx u_2 = 2/16$ ... $a_{15} \approx u_{15} = 15/16$.

Draw a picture to see why.

A plot of a_i against u_i is called a **PP Plot**.



NB(1) We can obtain confidence bands via the bootstrap.
NB(2) QQ plots (x_i vs $F^{-1}(u_i)$) are more useful for examining the tails of a distribution (read about them in your own time).

Copula Transformations

It is sometimes useful to be able to transform a random variable with one distribution into a random variable with another.

e.g. How do I turn a $N(1, 4)$ random variable Z into an Exp(2) random variable, Y?

By the PIT, $Z \sim N(1, 4) \Leftrightarrow U := \Phi((Z - 1)/2) \sim Unif(0, 1)$.

We already know how to simulate an Exp(2) from a Unif(0,1): $Y = -2 \log(1 - U)$.

The general formula for turning a continuous random variable X with cdf $F_X(x)$ into a continuous random variable Y with cdf $F_Y(y)$ is: