Lecture 9: Generating random variables – transformations and wrap-up

Ciaran Evans

Logistics

- Additional office hours added Mondays 2-3pm
- Upcoming due dates:
    - HW 1 resubmissions due today (end of the day)
    - HW 3 due Friday
    - Challenge 1 due Friday
- Feedback requested (google form)
- Project 1 released later this week

Recap: acceptance-rejection sampling

- Want to sample continuous r.v. $X \sim f$
- Can easily sample from a different density: $Y \sim g$, such that $\frac{f(t)}{g(t)} \leq c$

1. Sample $Y \sim g$
2. Sample $U \sim Uniform(0, 1)$
3. If $U \leq \dfrac{f(Y)}{cg(Y)}$, set $X = Y$. Otherwise, return to step 1.

Question: What are some potential downsides to the acceptance-rejection sampling method?

- If we keep rejecting, requires a lot of iterations
  could happen when f and g are very different
- Need to choose our candidate density g

Inefficiency in acceptance-rejection sampling

1. Sample $Y \sim g$
2. Sample $U \sim Uniform(0, 1)$
3. If $U \leq \frac{f(Y)}{cg(Y)}$, set $X = Y$. Otherwise, return to step 1.

$c = \max_t \frac{f(t)}{g(t)}$

c increases as f & g are more different

$$P(\text{accept } Y | Y = y) = P\left(U \leq \frac{f(y)}{cg(y)}\right) = ?? \quad \frac{f(y)}{cg(y)}$$

conditional probability

$$P(\text{accept } Y) = ?? \quad \frac{1}{c}$$

(decreases as f & g are more different)

marginal probability

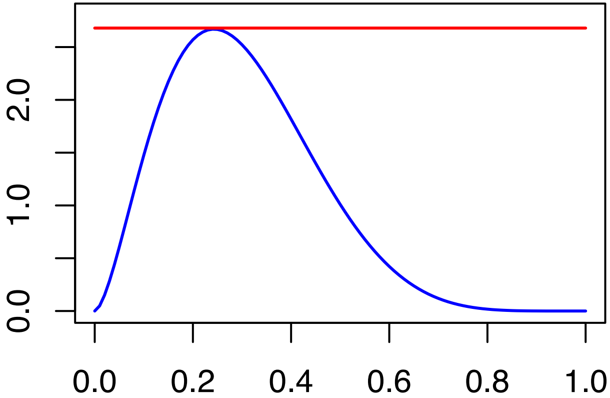
Law of total probability:

$$P(\text{accept } Y) = \int_{-\infty}^{\infty} P(\text{accept } Y | Y = y)\, g(y)\, dy$$

$$= \int_{-\infty}^{\infty} \frac{f(y)}{c\cancel{g(y)}} \cancel{g(y)}\, dy = \boxed{\frac{1}{c}}$$

Inefficiency in acceptance-rejection sampling

Beta(2.7, 6.3) example from class activity:



- $P(\text{accept } Y) = \frac{1}{c}$
- samples are independent

$\Rightarrow E[\text{time to get 1 sample from } f]$

$= \frac{1}{1/c}$

$= c$

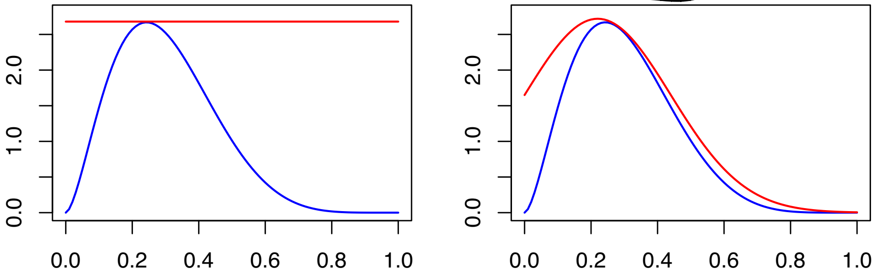
Here $c = 2.7$. About many samples from g would I need to get 1000 samples from f?

$c \cdot 1000$ (on average)

≈ 2700

Inefficiency in acceptance-rejection sampling

$c = 2.7$ $\quad$ $c = 1.26$ smaller value



Question: Which of these two candidate densities g would you prefer?

require $\dfrac{f(t)}{g(t)} \le c \quad \forall t$

$\Rightarrow \dfrac{f(t)}{cg(t)} \le 1 \quad \forall t$

Drawbacks of acceptance-rejection sampling

1. Need to find a suitable candidate g
2. Requires more samples $Y \sim g$ than we get from target f (because we reject some samples)
   - Want g to be as close as possible to f, to accept as many samples as possible
3. Calculating $f(Y)$ for candidate draws $Y \sim g$ may be expensive for some distributions

Project 1: Modifying the acceptance-rejection method to address these drawbacks

Today: Another approach to generating random variables

Generating a Normal random variable

Suppose we want to simulate $X \sim N(0, 1)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\}$$

$$F_X(t) = ? \qquad \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} dx$$

No closed form solution!

=> can't use inverse transform method

Today: Find a <u>different</u> transformation!

Box-Muller Transformation

Let $U_1, U_2 \overset{iid}{\sim} \text{Uniform}(0,1)$

iid : independent, identically distributed

Let
$$X_1 = \sqrt{-2\log U_1}\cos(2\pi U_2)$$
$$X_2 = \sqrt{-2\log U_1}\sin(2\pi U_2)$$

Then: $X_1, X_2 \overset{iid}{\sim} N(0,1)$

Pf : HW 3

Box-Muller in practice

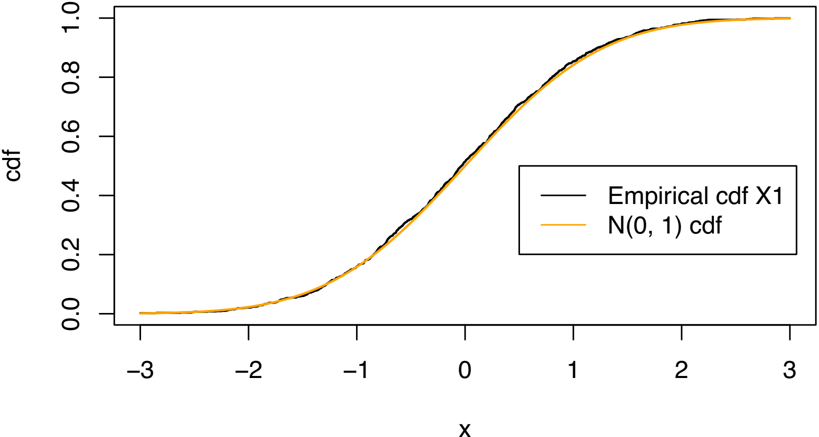
```
# generate 1000 samples
u1 <- runif(1000)
u2 <- runif(1000)
x1 <- sqrt(-2*log(u1)) * cos(2*pi*u2)
x2 <- sqrt(-2*log(u1)) * sin(2*pi*u2)
```

Question: How can I check that the samples match the desired $N(0,1)$ distribution?

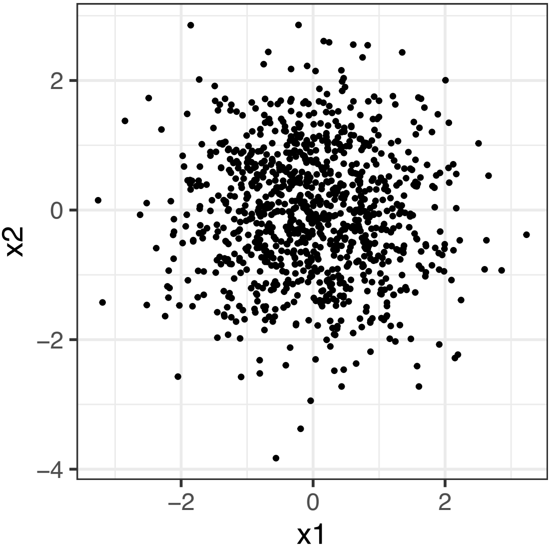
- check mean & variance
- plot empirical cdf
- Kolmogorov-Smirnov (KS) test
- histograms, QQ plots, etc.

Box-Muller in practice

```
# generate 1000 samples
u1 <- runif(1000)
u2 <- runif(1000)
x1 <- sqrt(-2*log(u1)) * cos(2*pi*u2)
x2 <- sqrt(-2*log(u1)) * sin(2*pi*u2)
```



Box-Muller in practice



Recall:
if variables are independent,
correlation = 0

correlation = 0 is a necessary (but not sufficient) condition for independence

Other Normals

$$\mathbb{E}[e^{tZ}] = e^{\frac{1}{2}t^2}$$

$$\mathbb{E}[e^{(\sigma t)Z}] = e^{\frac{1}{2}\sigma^2 t^2}$$

Suppose that $Z \sim N(0,1)$. How do I get $X \sim N(\mu, \sigma^2)$?

$$X = \mu + \sigma Z \qquad \sim N(\mu, \sigma^2)$$

$$\Longleftrightarrow \quad \underbrace{\frac{X-\mu}{\sigma}}_{\text{Z-score}} \sim N(0,1)$$

Pf: moment generating functions

mgf of a $N(\mu,\sigma^2) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

WTS: $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

$$M_X(t) = \mathbb{E}[e^{tX}] = \mathbb{E}[e^{t(\mu + \sigma Z)}] = \mathbb{E}[e^{\mu t} e^{\sigma t Z}]$$

$$= e^{\mu t}\,\mathbb{E}[e^{\sigma t Z}] = e^{\mu t} e^{\frac{1}{2}\sigma^2 t^2} = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

//

A few other transformations

- If $X \sim N(\mu, \sigma^2)$, then $e^X \sim Lognormal(\mu, \sigma^2)$
- If $Z_1, ..., Z_k \overset{iid}{\sim} N(0,1)$, then

$$\sum_{i=1}^{k} Z_i^2 \sim ? \qquad \chi^2_k \qquad (311/611)$$

- If $V_1 \sim \chi^2_{d_1}$ and $V_2 \sim \chi^2_{d_2}$ are independent, then

$$\frac{V_1/d_1}{V_2/d_2} \sim ? \qquad F_{d_1,d_2} \qquad (311/611) \quad (312/612)$$

≥ 0 $\qquad$ ≥ 0

- If $Y_1 \sim Gamma(\alpha, \theta)$ and $Y_2 \sim Gamma(\beta, \theta)$ are independent, then

$0 \leq \frac{Y_1}{Y_1+Y_2} \leq 1$

$$\frac{Y_1}{Y_1 + Y_2} \sim ? \qquad Beta(\alpha, \beta)$$

Summary (so far)

Methods to generate random variables, in rough order of preference:

1. Use inverse transform method (if inverse cdf is tractable)
2. Find a different transformation (if possible)
3. Acceptance-rejection sampling (perhaps with modifications)

Homework 3

https://sta379-s25.github.io/homework/hw3.html

- Practice generating random variables
- Accept and submit coding portion of assignment on GitHub Classroom
- Collaboration encouraged on homework, but everyone must submit their own work and acknowledge collaborators