YOUNG-GEUN statistics

Statistical Computing

R Lab

O RLY?

Young-geun Kim

R Lab for Statistical Computing

Young-geun Kim
Department of Statistics, SKKU
dudrms33@g.skku.edu

21 Apr, 2019

Contents

Welcome

Statistical computing mainly treats useful simulation methods.

```
library(tidyverse)
```

`tidyverse` package family will be used in every chapter. Loading step is in `_common.R`, so it is not included in the text. Sometimes `data.table` library will be called for efficiency.

Statistical Computing

We first look at *random generation* methods. Lots of simulation methods are built based on this random numbers.

Sampling from a fininte population

Generating random numbers is like sampling. From finite population, we can sample data with or without replacement. For example of sampling with replacement, we toss coins 10 times.

```
sample(0:1, size = 10, replace = TRUE)
#>  [1] 1 0 0 1 0 1 1 0 1 1
```

Sampling without replacement: Choose some lottery numbers which consist of 1 to 100.

```
sample(1:100, size = 6, replace = FALSE)
#> [1] 61 83 50 74 34 35
```

Random generators of common probability distributions

`R` provides some functions which generate random numbers following famous distributions. Although we will learn some skills generating these numbers in basis levels, these functions do the same thing more elegantly.

```
gg_curve(dbeta, from = 0, to = 1, args = list(shape1 = 3, shape2 = 2)) +
  geom_histogram(
    data = tibble(
      rand = rbeta(1000, 3, 2),
      idx = seq(0, 1, length.out = 1000)
    ),
    aes(x = rand, y = ..density..),
    position = "identity",
    bins = 30,
    alpha = .45,
    fill = gg_hcl(1)
  )
```

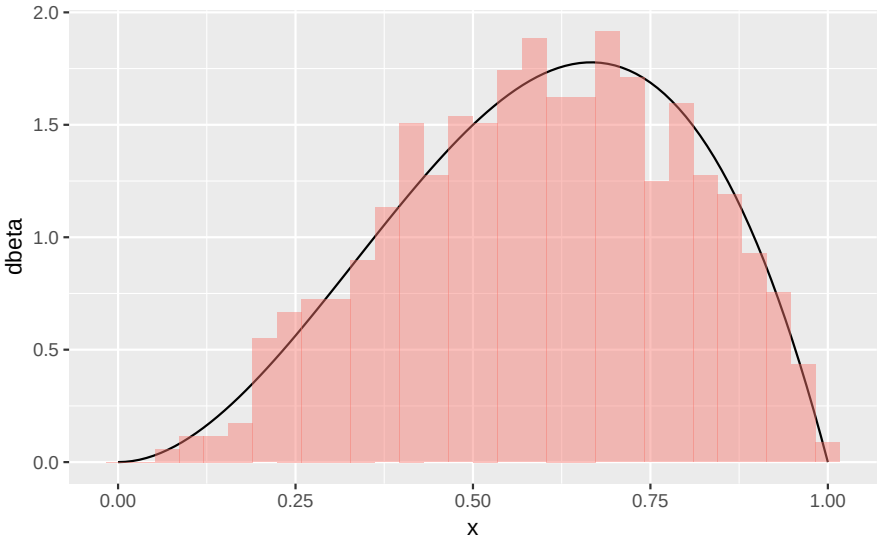


Figure 1: Beta(3,2) random numbers

Figure 1 shows that `rbeta()` function generate random numbers very well. Histogram is of the random number, and the curve is the true beta distribution.

Chapter 1

Methods for Generating Random Variables

1.1 Introduction

Most of the methods so-called *computational statistics* requires generation of random variables from specified probability distribution. In hand, we can spin wheels, roll a dice, or shuffle cards. The results are chosen randomly. However, we want the same things with computer. Here, `r`. As we know, computer cannot generate complete uniform random numbers. Instead, we generate **pseudo-random** numbers.

1.2 Pseudo-random Numbers

Definition 1.1 (Pseudo-random numbers)**.** Sequence of values generated deterministically which have all the appearances of being independent $unif(0,1)$ random variables, i.e.

$$x_1, x_2, \ldots, x_n \overset{iid}{\sim} unif(0,1)$$

- behave *as if* following $unif(0,1)$
- typically generated from an *initial seed*

1.2.1 Linear congruential generator

Then $u_1, u_2, \ldots, u_n \sim unif(0,1)$

Algorithm 1: Linear congruential generator

input : $a, c \in \mathbb{Z}_+$ and modulus m

Initialize x_0;

for $i \leftarrow 1$ **to** n **do**

  $x_i = (ax_{i-1} + c) \mod m$;

end

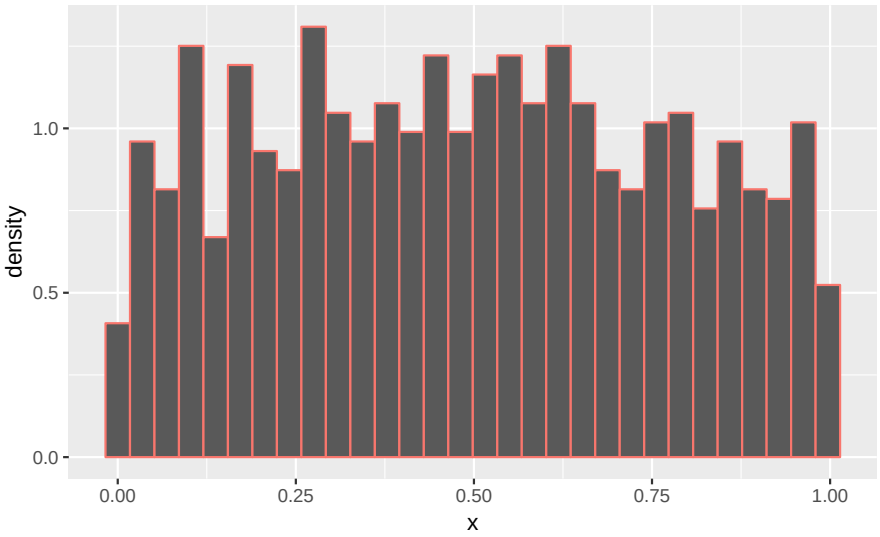
$u_i = \frac{x_i}{m} \in (0,1)$;

output: $u_1, u_2, \ldots, u_n \sim unif(0,1)$

```
lcg <- function(n, seed, a, b, m) {
  x <- rep(seed, n + 1)
  for (i in 1:n) {
    x[i + 1] <- (a * x[i] + b) %% m
  }
```

```
  x[-1] / m
}
```

```
tibble(
  x = lcg(1000, 0, 1664525, 1013904223, 2^32)
) %>%
  ggplot(aes(x = x)) +
  geom_histogram(aes(y = ..density..), bins = 30, col = gg_hcl(1))
```



1.2.2 Multiplicative congruential generator

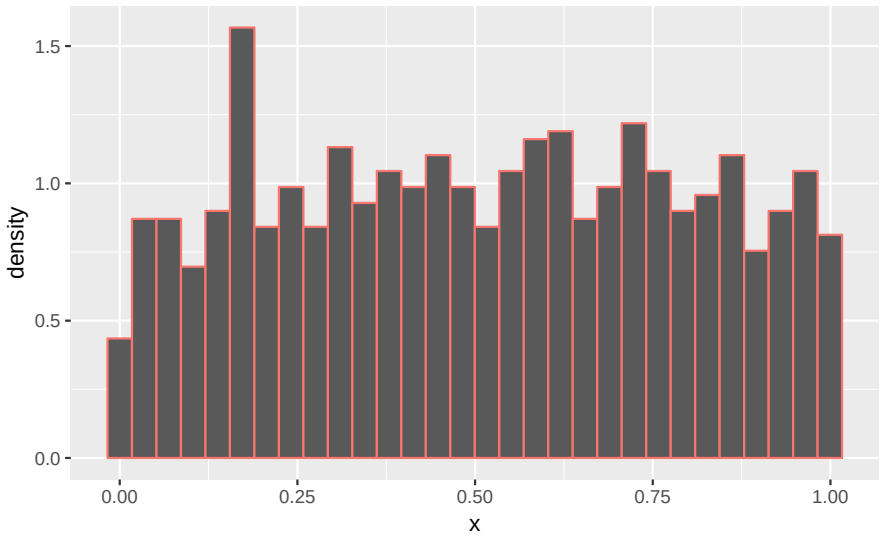
As we can expect from its name, this is congruential generator with $c = 0$.

Algorithm 2: Multiplicative congruential generator

input : $a, \in \mathbb{Z}_+$ and modulus m
Initialize x_0;
for $i \leftarrow 1$ **to** n **do**
  $x_i = ax_{i-1} \mod m$;
end
$u_i = \frac{x_i}{m} \in (0, 1)$;
output: $u_1, u_2, \ldots, u_n \sim unif(0, 1)$

We just set `b = 0` in our `lcg()` function. The **seed must not be zero**.

```
tibble(
  x = lcg(1000, 5, 1664525, 0, 2^32)
) %>%
  ggplot(aes(x = x)) +
  geom_histogram(aes(y = ..density..), bins = 30, col = gg_hcl(1))
```



1.2.3 Cycle

Generate LCG $n = 32$ with $a = 1$, $c = 1$, and $m = 16$ from the seed $x_0 = 0$.

```
lcg(32, 0, 1, 1, 16)
#>  [1] 0.0625 0.1250 0.1875 0.2500 0.3125 0.3750 0.4375 0.5000 0.5625 0.6250
#> [11] 0.6875 0.7500 0.8125 0.8750 0.9375 0.0000 0.0625 0.1250 0.1875 0.2500
#> [21] 0.3125 0.3750 0.4375 0.5000 0.5625 0.6250 0.6875 0.7500 0.8125 0.8750
#> [31] 0.9375 0.0000
```

Observe that we have the cycle after m-th number. Against this problem, we give different seed from every $(im + 1)$th random number.

1.3 The Inverse Transform Method

Definition 1.2 (Inverse of CDF)**.** Since some cdf F_X is not strictly increasing, we difine $F_X^{-1}(y)$ for $0 < y < 1$ by

$$F_X^{-1}(y) := inf\{x : F_X(x) \geq y\}$$

Using this definition, we can get the following theorem.

Theorem 1.1 (Probability Integral Transformation)**.** *If X is a continuous random variable with cdf $F_(x)$, then*

$$U \equiv F_X(X) \sim unif(0, 1)$$

Probability Integral Transformation. Let $U \sim unif(0, 1)$. Then

$$\begin{aligned} P(F_X^{-1}(U) \leq x) &= P(\inf\{t : F_X(t) = U\} \leq x) \\ &= P(U \leq F_X(x)) \\ &= F_U(F_X(x)) \\ &= F_X(x) \end{aligned}$$

□

Thus, to generate n random variables $\sim F_X$, we can use *uniform random numbers*.

Algorithm 3: Inverse transformation method

input : analytical form of F_X^{-1}
for $i \leftarrow 1$ **to** n **do**
  $u_i \overset{iid}{\sim} unif(0,1)$;
  $x_i = F_X^{-1}(u_i)$;
end
output: $x_1, x_2, \ldots, x_n \overset{iid}{\sim} F_X$

Note that in `R`, vectorized operation would be better, i.e. generate `runif(n)` and plug it into given inverse cdf.

1.3.1 Continuous case

Denote that the *probability integral transformation* holds for a continuous variable. When generating continuous random variable, applying above algorithm might work.

Example 1.1 (Exponential distribution)**.** If $X \sim Exp(\lambda)$, then $F_X(x) = 1 - e^{-\lambda x}$. We can derive the inverse function of cdf

$$F_X^{-1}(u) = \frac{1}{\lambda} \ln(1-u)$$

Note that

$$U \sim unif(0,1) \Leftrightarrow 1 - U \sim unif(0,1)$$

Then we just can use U instead of $1 - U$.

```
inv_exp <- function(n, lambda) {
  -log(runif(n)) / lambda
}
```

If we generate $x_1, \ldots, x_{500} \sim Exp(\lambda = 1)$,

```
gg_curve(dexp, from = 0, to = 10) +
  geom_histogram(
    data = tibble(x = inv_exp(500, lambda = 1)),
    aes(x = x, y = ..density..),
    bins = 30,
    fill = gg_hcl(1),
    alpha = .5
  )
```

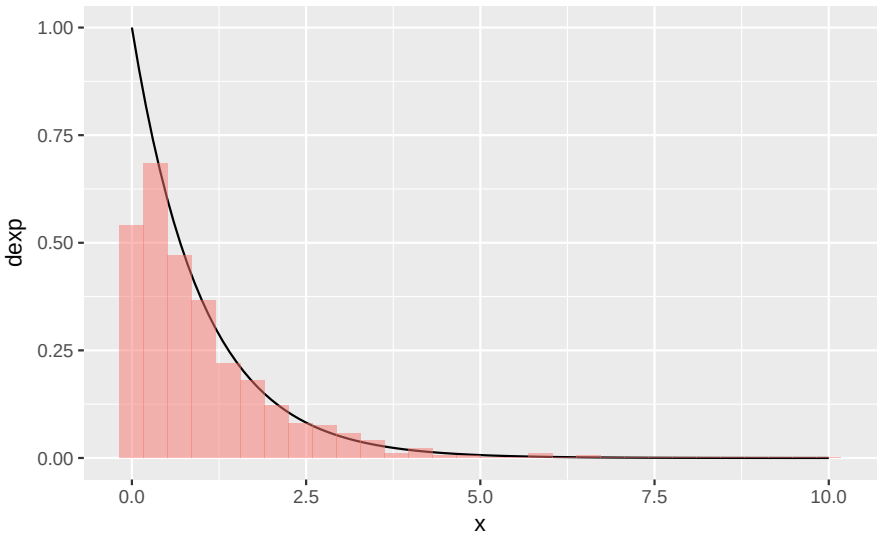


Figure 1.1: Inverse Transformation: Exp(1)

1.3.2 Discrete case

Algorithm 4: Inverse transformation method in discrete case

input : analytical form of F_X

for $i \leftarrow 1$ **to** n **do**

  $u_i \overset{iid}{\sim} unif(0,1)$;

  Take x_i s.t. $F_X(x_{i-1}) < U \leq F_X(x_i)$;

end

output: $x_1, x_2, \ldots, x_n \overset{iid}{\sim} F_X$

Table 1.1: Example of a Discrete Random Variable

x	0.0	1.0	2.0	3.0	4.0
p	0.1	0.2	0.2	0.2	0.3

Example 1.2 (Discrete Random Variable)**.** Consider a discrete random variable X with a mass function as in Table 1.1.

i.e.

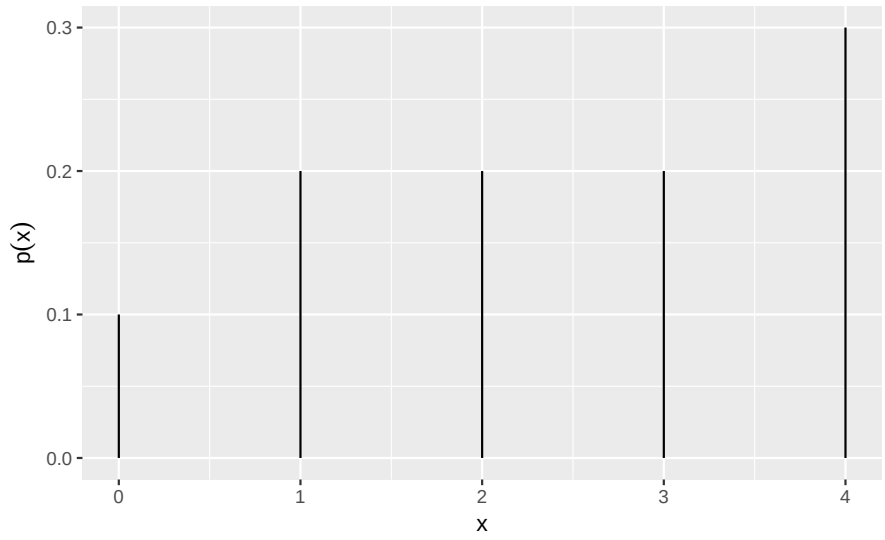


Figure 1.2: Probability Mass Function

Then we have the cdf

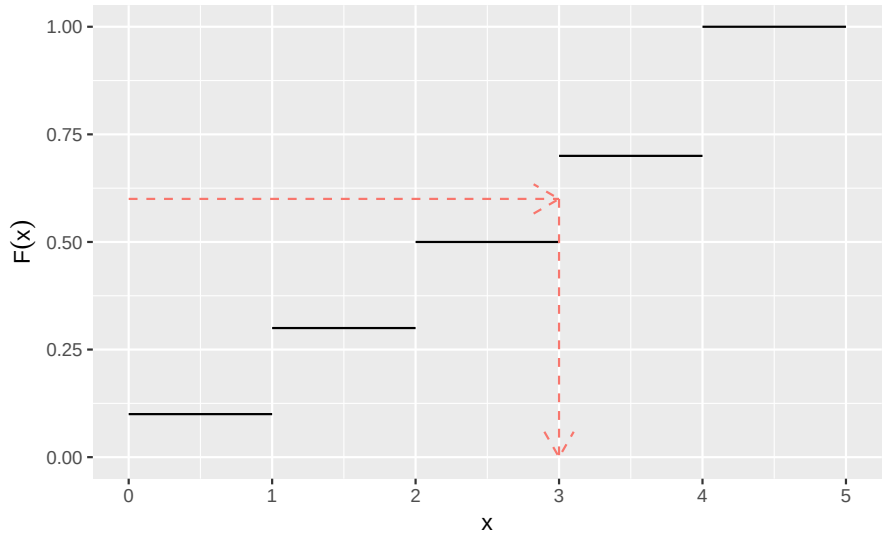


Figure 1.3: CDF of the Discrete Random Variable: Illustration for discrete case

Remembering the algorithm, we can implement `dplyr::case_when()` here.

```
rcustom <- function(n) {
  tibble(u = runif(n)) %>%
    mutate(
      x = case_when(
        u > 0 & u <= .1 ~ 0,
        u > .1 & u <= .3 ~ 1,
        u > .3 & u <= .5 ~ 2,
        u > .5 & u <= .7 ~ 3,
        TRUE ~ 4
      )
```

```
    ) %>%
    select(x) %>%
    pull()
}
```

```
tibble(
  x = rcustom(100)
) %>%
  ggplot(aes(x = x)) +
  geom_histogram(aes(y = ..ndensity..), binwidth = .1)
```

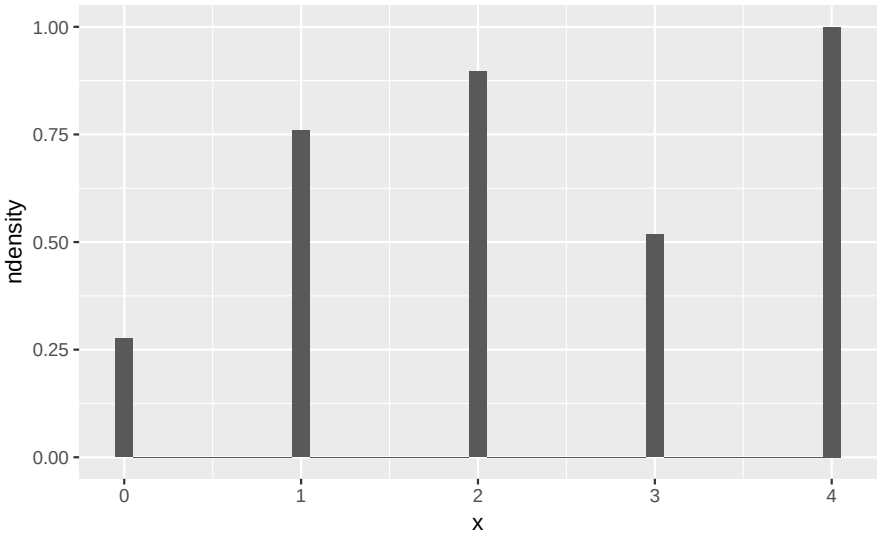


Figure 1.4: Generated discrete random numbers

See Figure 1.2 and 1.4. Comparing the two, the result can be said okay.

1.3.3 Problems with inverse transformation

Examples 1.1 and 1.2. We could generate these random numbers because we aware of

1. analytical F_X
2. F^{-1}

In practice, however, not all distribution have analytical F. Numerical computing might be possible, but it is not efficient. There are other approaches.

1.4 The Acceptance-Rejection Method

Acceptance-rejection method does not require analytical form of cdf. What we need is our *target* density (or mass) function and *proposal* density (or mass) function. Target function is what we want to generate. Propsal function is of any random variable that is *easy to generate random numbers*. From this approach, we can generate any distribution while computation is not efficient.

pdf or pmf	target or proposal
f	target
g	proposal - easy to generate random numbers

First of all, g should satisfy that

$$sptf \subseteq sptg$$

Next, for some (pre-specified) $c > 0$

$$\forall x \in sptf : \frac{f(x)}{g(x)} \leq c$$

Algorithm 5: Acceptance-rejection algorithm

input : target f, proposal g, and c

1 **for** $i \leftarrow 1$ **to** n **do**
2 $\quad Y \sim g(y)$;
3 $\quad U \sim unif(0,1) \perp\!\!\!\perp Y$;
4 $\quad$ **if** $U \leq \frac{f(Y)}{cg(Y)}$ **then**
5 $\quad\quad$ Accept $x_i = Y$;
6 $\quad$ **else**
7 $\quad\quad$ go to Line 2;
8 $\quad$ **end**
9 **end**

output: $x_1, x_2, \ldots, x_n \overset{iid}{\sim} f(x)$

1.4.1 Efficiency

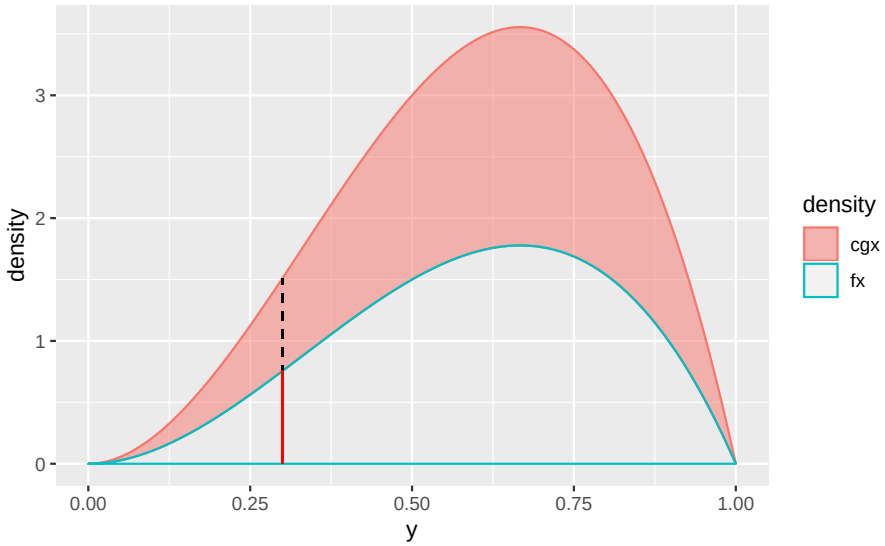


Figure 1.5: Property of AR method

See Figure 1.5. This illustrates the motivation of A-R method. Lower one is $f(x)$ and the upper one is $cg(x)$ which covers f. We can see that

$$0 < \frac{f(x)}{cg(x)} \leq 1$$

The algorithm takes random number from $Y \sim g$ in each recursive step i, which is represented as a line in the figure. At this value, the algorithm accept Y as random number of f if

$$U \leq \frac{f(Y)}{cg(Y)}$$

Suppose that we choose a point at random on a line drawn in the figure 1.5. If we get the red line, we accept. Otherwise, we reject. In other words, the *colored area is where we reject the given value.* The smaller the area is, the more efficient the algorithm will be.

Proposition 1.1 (Properties of A-R Method)**.** *See Figure 1.5.*

1. $\frac{f(Y)}{cg(Y)} \perp\!\!\!\perp U$
2. $0 < \frac{f(x)}{cg(x)} \leq 1$
3. *Let N be the number of iterations needed to get an acceptance. Then*

$$N \sim Geo(p) \quad where\, p \equiv P\left(U \leq \frac{f(Y)}{cg(Y)}\right)$$

and so

$$\begin{cases} P(N = n) = p(1-p)^{n-1} I_{\{1,2,\dots\}}(n) \\ E(N) = average\ number\ of\ iterations = \frac{1}{p} \end{cases}$$

4. $X \sim Y \mid U \leq \frac{f(Y)}{cg(Y)}$, *i.e.*

$$P\left(Y \leq y \mid U \leq \frac{f(Y)}{cg(Y)}\right) = F_X(y)$$

Remark (Efficiency)*.* Efficiency of the A-R method depends on $p = P\left(U \leq \frac{f(Y)}{cg(Y)}\right)$. In fact,

$$E(N) = \frac{1}{p} = c$$

The algorithm becomes efficient for small c.

Proof. Note that

$$P\left(U \leq \frac{f(y)}{cg(y)}, Y = y\right) = P\left(Y \leq \frac{g(y)}{cg(y)} \mid Y = y\right) P(Y = y)$$

Since $U \sim unif(0,1)$, $P\left(Y \leq \frac{g(y)}{cg(y)} \mid Y = y\right) = \frac{f(y)}{cg(y)}$.

By construction, $P(Y = y) = g(y)$.

It follows that

$$\begin{aligned} p = P\left(U \leq \frac{f(y)}{cg(y)}\right) &= \int_{-\infty}^{\infty} P\left(U \leq \frac{f(y)}{cg(y)}, Y = y\right) dy \\ &= \int_{-\infty}^{\infty} \frac{f(y)}{cg(y)} g(y) dy \\ &= \frac{1}{c} \int_{-\infty}^{\infty} f(y) dy \\ &= \frac{1}{c} \end{aligned}$$

Hence,

$$E(N) = \frac{1}{p} = c$$

We can say that the method is efficient when the acceptance rate p is large, i.e. c small. □

Corollary 1.1 (Efficiency of A-R Method)**.** *A-R method is efficient when*

$g(\cdot)$ *is close to* $f(\cdot)$ *and*

have small c.

Corollary 1.2 (Choosing c)**.** *To enhance the algorithm, we might choose c which satisfy*

$$c = \max\left\{\frac{f(x)}{g(x)} : x \in sptf\right\}$$

1.4.2 Examples

Example 1.3 (Beta(a,b))**.** Let $X \sim Beta(a, b)$. Then the pdf of X is given by

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} I_{(0,1)}(x)$$

Solution (Generating Beta(a,b) with A-R method). Consider proposal density $g(x) = I_{(0,1)}(x)$, i.e. $unif(0, 1)$.

To determine the optimal c s.t.

$$c = \max\left\{\frac{f(x)}{g(x)} : x \in (0,1)\right\}$$

find the maximum of

$$\frac{f(x)}{g(x)} = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$$

Solve

$$\begin{aligned}\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) &= \frac{1}{B(a,b)}\Big((a-1)x^{a-2}(1-x)^{b-1} - (b-1)x^{a-1}(1-x)^{b-2}\Big) \\ &= \frac{x^{a-2}(1-x)^{b-2}}{B(a,b)}\Big((a-1)(1-x) - (b-1)x\Big) \\ &= \frac{x^{a-2}(1-x)^{b-2}}{B(a,b)}\big(a-1-(a+b-2)x\big) \quad = 0\end{aligned}$$

It follows that

$$\frac{f(x)}{g(x)} \leq \frac{f(\frac{a-1}{a+b-2})}{g(\frac{a-1}{a+b-2})} = c$$

if $\frac{a-1}{a+b-2} \neq 0, 1$

```
ar_beta <- function(n, a, b) {
  opt_x <- (a - 1) / (a + b - 2)
  opt_c <- dbeta(opt_x, shape1 = a, shape2 = b) / dunif(opt_x)
  X <- NULL
  N <- 0
  while (N <= n) {
    Y <- runif(n)
    U <- runif(n)
    X <- c(X, Y[U <= dbeta(Y, shape1 = a, shape2 = b) / opt_c])
    N <- length(X)
    if ( N > n ) X <- X[1:n]
  }
  X
}
```

Now we try to compare this A-R function to `R rbeta` function.

```
gen_beta <-
  tibble(
    ar_rand = ar_beta(1000, 3, 2),
    sam = rbeta(1000, 3, 2)
  ) %>%
  gather(key = "den", value = "value")
```

```
gg_curve(dbeta, from = 0, to = 1, args = list(shape1 = 3, shape2 = 2)) +
  geom_histogram(
    data = gen_beta,
    aes(x = value, y = ..density.., fill = den),
    position = "identity",
    bins = 30,
    alpha = .45
  ) +
  scale_fill_discrete(
    name = "random number",
    labels = c("AR", "rbeta")
  )
```

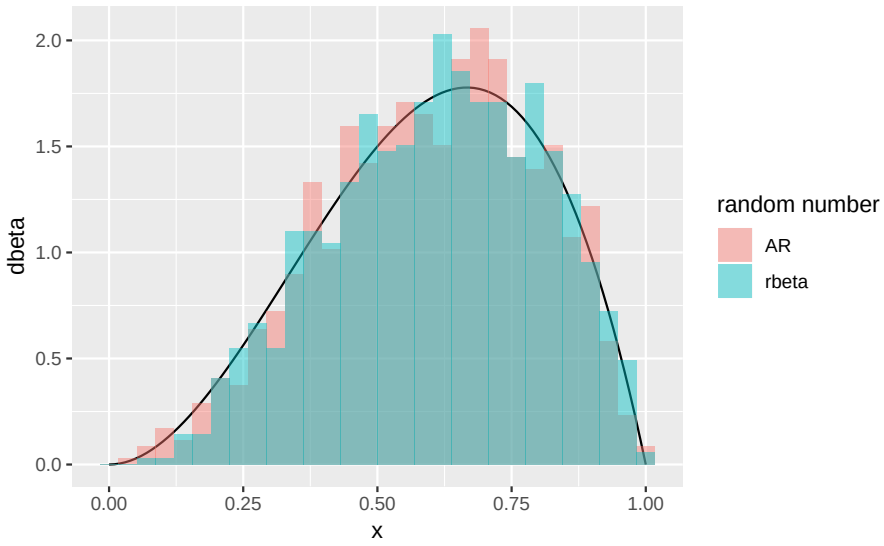


Figure 1.6: Beta(3,2) Random numbers from each function

In the Figure 1.6, the both histograms are very close to the true density curve. To see more statistically, we can draw a Q-Q plot.

```
gen_beta %>%
  ggplot(aes(sample = value)) +
  stat_qq_line(
    distribution = stats::qbeta,
    dparams = list(shape1 = 3, shape2 = 2),
    col = I("grey70"),
    size = 3.5
  ) +
  stat_qq(
    aes(colour = den),
    distribution = stats::qbeta,
    dparams = list(shape1 = 3, shape2 = 2)
  ) +
  scale_colour_discrete(
    name = "random number",
    labels = c("AR", "rbeta")
  )
```

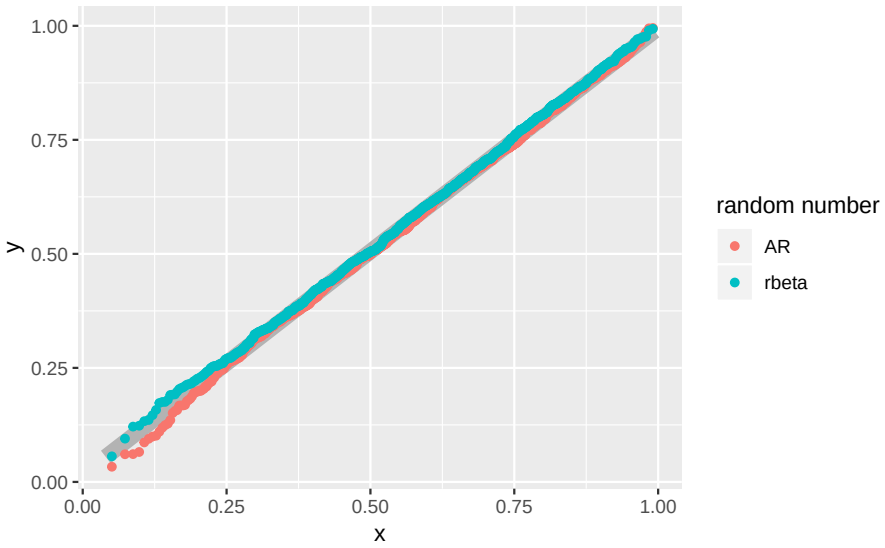


Figure 1.7: Q-Q plot for Beta(3,2) random numbers

See Figure 1.7. We have got series of numbers that are sticked to the beta distribution line.

Example 1.4 (A-R Method for Discrete case)**.** A-R method can be also implemented to discrete case such as Example 1.2.

Table 1.3: Example of a Discrete Random Variable

x	0.0	1.0	2.0	3.0	4.0
p	0.1	0.2	0.2	0.2	0.3

Solution (Generating discrete random numbers using A-R methods). Consider proposal $g(x) \sim$ Discrete unif$(0, 1, 2, 3, 4)$, i.e.

$$g(0) = g(1) = \cdots = g(4) = 0.2$$

Then we set

$$c = \max\left\{\frac{p(x)}{g(x)} : x = 0, \ldots, 4\right\} = \max\left\{0.5, 1, 1.5\right\} = 1.5$$

1.5 Transfomation Methods

1.5.1 Continuous

Proposition 1.2 (Transformation between continuous random variables)**.** *Relation between random variables enables generating target numbers from the others.*

1. $Z_1, \ldots, Z_n \overset{iid}{\sim} N(0,1) \Rightarrow \sum Z_i^2 \sim \chi^2(n)$
2. $Y_1 \sim \chi^2(m) \perp\!\!\!\perp Y_2 \sim \chi^2(n) \Rightarrow \frac{Y_1/m}{Y_2/n} \sim F(m,n)$
3. $Z \sim N(0,1) \perp\!\!\!\perp Y \sim \chi^2(n) \Rightarrow \frac{Z}{\sqrt{Y/n}} \sim t(n)$

4. $Y_1, \ldots, Y_n \overset{iid}{\sim} Exp(\lambda) \Rightarrow \sum Y_i^2 Gamma(n, \lambda)$
5. $U \sim unif(0,1) \Rightarrow (b-a)U + a \sim unif(a,b)$
6. $U \sim Gamma(r,\lambda) \perp\!\!\!\perp V \sim Gamma(s,\lambda) \Rightarrow \frac{U}{U+V} \sim Beta(r,s)$
7. $Z \sim N(0,1) \Rightarrow \mu + \sigma Z \sim N(\mu, \sigma^2)$
8. $Y \sim N(\mu, \sigma^2) \Rightarrow e^Y \sim LogNormal(\mu, \sigma^2)$

Example 1.5 (Generating Beta(a, b) using rgamma)**.** From Proposition 1.2, we can generate $Beta(a,b)$ random numbers using $Gamma(a,1)$ and $Gamma(b,1)$.

```
trans_beta <- function(n, shape1, shape2) {
  u <- rgamma(n, shape = shape1, rate = 1)
  v <- rgamma(n, shape = shape2, rate = 1)
  u / (u + v)
}
```

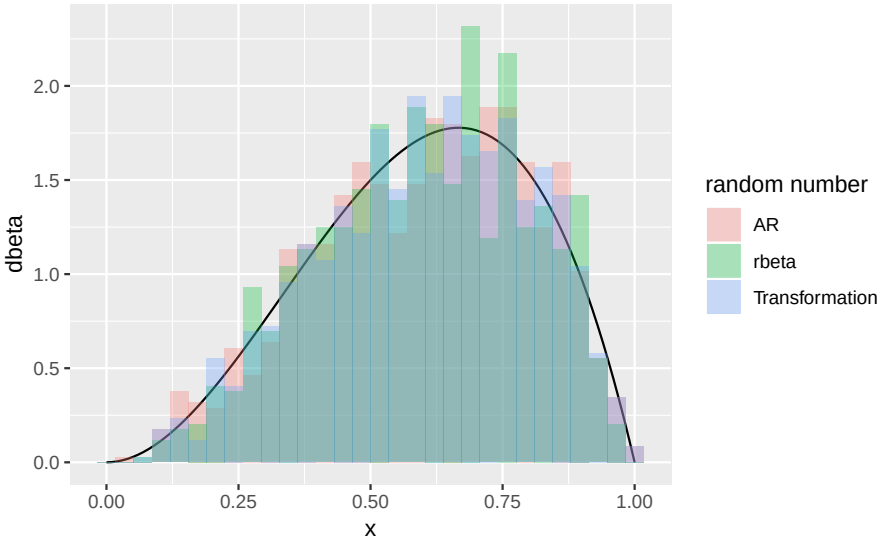


Figure 1.8: Beta(3,2) Random numbers from each function, including transformation method

1.5.2 Box-Muller transformation

Denote that Gaussian cdf has no closed form of F_X^{-1}. Using polar coordiantes, we can generate Normal random numers.

Theorem 1.2 (Box-Muller transformation)**.** *Let* $U_1, U_2 \overset{iid}{\sim} unif(0,1)$. *Then*

$$\begin{cases} Z_1 = \sqrt{-2\ln U_2}\cos(2\pi U_1) \\ Z_2 = \sqrt{-2\ln U_2}\sin(2\pi U_1) \end{cases}$$

Proof. Write

$$(Z_1, Z_2)^T \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

Then the joint pdf is given by

$$f_{Z_1,Z_2}(x_1,x_2) = \frac{1}{2\pi}\exp\left(-\frac{x_1^2+x_2^2}{2}\right)$$

Consider polar coordiate transformation (R,θ): $x_1 = R\cos\theta$ and $x_2 = R\sin\theta$.

Since it is also random vector,

$$\begin{aligned} f_{R,\theta}(r,\theta) &= f_{Z_1,Z_2}(x_1,x_2)|J| \\ &= \frac{1}{2\pi}\exp\left(-\frac{x_1^2+x_2^2}{2}\right)\left|\begin{array}{cc}\frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta}\\ \frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \theta}\end{array}\right| \\ &= \frac{1}{2\pi}\exp\left(-\frac{r^2}{2}\right)\left|\begin{array}{cc}\frac{\partial x_1}{\partial r} & \frac{\partial x_1}{\partial \theta}\\ \frac{\partial x_2}{\partial r} & \frac{\partial x_2}{\partial \theta}\end{array}\right| \\ &= \frac{r}{2\pi}\exp\left(-\frac{r^2}{2}\right) \end{aligned}$$

Then each marginal density function can be computed as

$$\begin{aligned} f_\theta(\theta) &= \int_0^\infty \frac{r}{2\pi}\exp\left(-\frac{r^2}{2}\right)dr \\ &= \frac{1}{2\pi}I_{(0,2\pi)}(\theta) \\ &\overset{d}{=} unif(0,2\pi) \end{aligned}$$

$$\begin{aligned} f_R(r) &= \int_0^\theta \frac{r}{2\pi}\exp\left(-\frac{r^2}{2}\right)d\theta \\ &= r\exp\left(-\frac{r^2}{2}\right)I_{(0,\infty)}(r) \end{aligned}$$

Thus,

$$f_{R,\theta} = f_\theta f_R \Rightarrow R \perp\!\!\!\perp \theta$$

It follows from inverse transformation theorem that

$$Z_1 = R\cos\theta = \sqrt{-2\ln U_2}\cos(2\pi U_1)$$

and that

$$Z_2 = R\sin\theta = \sqrt{-2\ln U_2}\sin(2\pi U_1)$$

where $U_1, U_2 \overset{iid}{\sim} unif(0,1)$ □

Algorithm 6: Box-Muller transformation

for $i \leftarrow 1$ **to** n **do**
  $U_1, U_2 \overset{iid}{\sim} unif(0,1)$;
  $z_{2i-1} = \sqrt{-2\ln U_2}\cos(2\pi U_1)$;
  $z_{2i} = \sqrt{-2\ln U_2}\sin(2\pi U_1)$;
end
output: $z_1, \ldots, z_n \overset{iid}{\sim} N(0,1)$

```
bmnorm <- function(n, mean = 0, sd = 1) {
  n_bm <- ceiling(n / 2)
  tibble(
    theta = runif(n = n_bm, max = 2 * pi),
    R = sqrt(-2 * log(runif(n_bm)))
  ) %>%
    mutate(
      x1 = R * cos(theta),
      x2 = R * sin(theta)
    ) %>%
    gather(x1, x2, key = "key", value = "value") %>%
    mutate(value = mean + sd * value) %>%
    select(value) %>%
    pull()
}
```

```
gg_curve(dnorm, from = 0, to = 6, args = list(mean = 3, sd = 1)) +
  geom_histogram(
    data = tibble(x = bmnorm(1000, mean = 3, sd = 1)),
    aes(x = x, y = ..density..),
    bins = 30,
    fill = gg_hcl(1),
    alpha = .5
  )
```

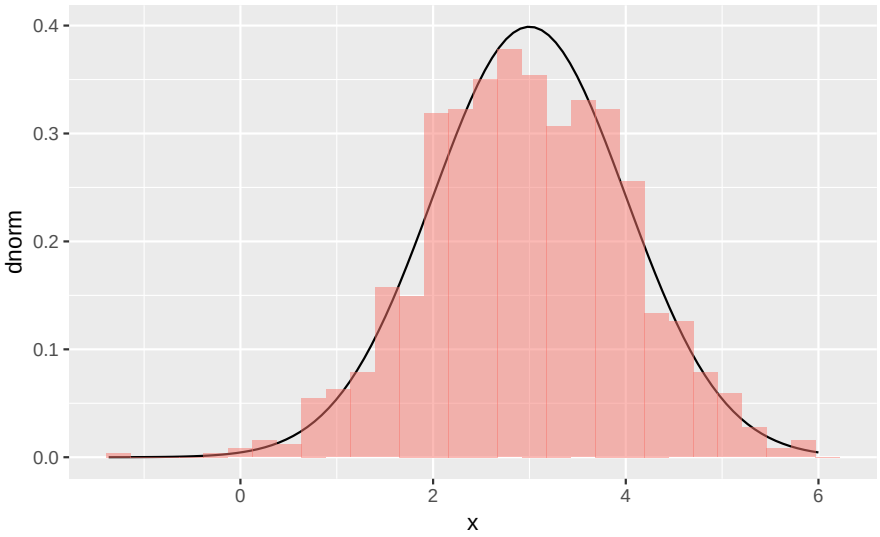


Figure 1.9: Normal random numbers by Box-Muller transformation

1.5.3 Discrete

Proposition 1.3 (Transformation between discrete random variables)**.** *Relation between random variables enables generating target numbers from the others.*

1. $Y_1, \ldots, Y_n \overset{iid}{\sim} Bernoulli(p) \Rightarrow \sum Y_i^2 \sim B(n, p)$
2. $U \sim unif(0, 1) \Rightarrow X_i = \lfloor mU \rfloor + 1$
3. $X =$ *the number of events occurring in 1 unit of time* $\sim Poisson(\lambda)$

Proposition 1.4 (Bernoulli process)**.** *Let* $X_1, X_2, \ldots \overset{iid}{\sim} Bernoulli(p)$.

1. $N =$ *the number of trials until we see a success, i.e.*$X_N = 1 \Rightarrow N \sim Geo(p)$
2. $Y_1, \ldots, Y_r \overset{iid}{\sim} Geo(p) \Rightarrow \sum_{i=1}^{r} Y_i =$ *the number of trials until we see* r *successes* $\sim NegBin(r, p)$

Proposition 1.5 (Count process)**.** *Let* $Y_1, Y_2, \ldots \overset{iid}{\sim} Exp(\lambda)$ *be interarrival times. Then*

$$X = \max\{n : \sum Y_i \leq 1\} = \text{the number of events occurring in 1 unit of time} \sim Poisson(\lambda)$$

1.6 Sums and Mixtures

1.6.1 Convolutions

Definition 1.3 (Convolution)**.** Let $X_1, \ldots, X_n$ be independent and identically distributed and let $S = X_1 + \cdots X_n$. Then the distribution of S is called the n-fold convolution of X and denoted by $F_X^{*(n)}$.

In the last chapter, we have already seen a bunch of random variables that can be generated by summing the other.

Example 1.6 (Chisquare)**.** Let $Z_1, \ldots, Z_n \overset{iid}{\sim} N(0, 1)$. We know from Proposition 1.2 that

$$V = \sum_{i=1}^{n} Z_i \sim \chi^2(n)$$

Building a `n` $\times$ `df` matrix can be a good strategy here. After that, `rowSums` or `colSums` ends the generation work.

```
conv_chisq <- function(n, df) {
  X <-
    matrix(rnorm(n * df), nrow = n, ncol = df)^2
  rowSums(X)
}
```

```
gg_curve(dchisq, from = 0, to = 15, args = list(df = 5)) +
  geom_histogram(
    data = tibble(x = conv_chisq(1000, df = 5)),
    aes(x = x, y = ..density..),
    bins = 30,
    fill = gg_hcl(1),
    alpha = .5
  )
```

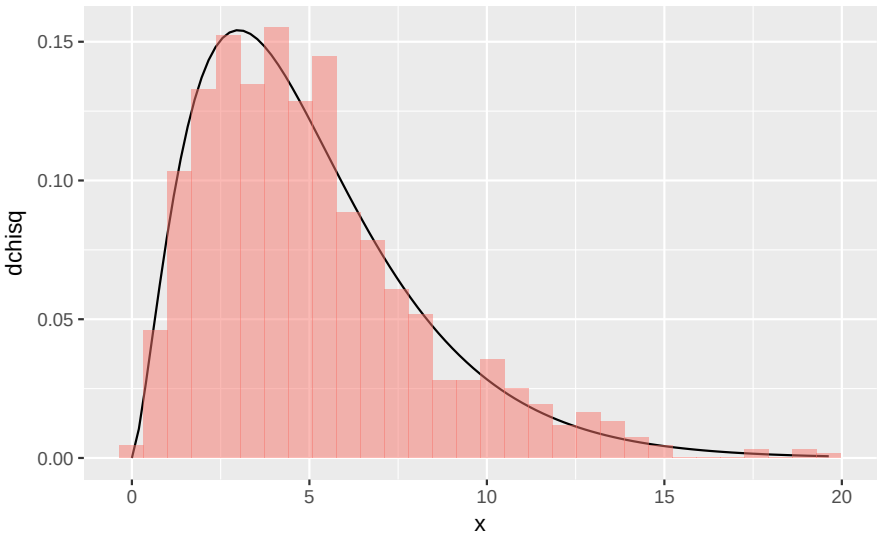


Figure 1.10: χ^2 random numbers from Normal sums

1.6.2 Mixtures

Definition 1.4 (Discrete mixture)**.** A random variable X is a discrete mixture if the distribution of X is a weighted sum

$$F_X(x) = \sum \theta_i F_{X_i}(x)$$

where constants θ_i are called the mixing weights or mixing probabilities.

Definition 1.5 (Continuous mixture)**.** A random variable X is a continuous mixture if the distribution of X is a weighted sum

$$F_X(x) = \int_{\infty}^{\infty} F_{X|Y=y}(x) f_Y(y) dy$$

Example 1.7 (Mixture of several Normal distributions)**.** Generate a random sample of size 1000 from a normal location mixture with components of the mixture $N(0,1)$ and $N(3,1)$, i.e.

$$F_X = p_1 F_{X_1} + (1 - p_1) F_{X_2}$$

For easy combining samples, we use `foreach` library.

```
library(foreach)
```

As in A-R method, Bernoullin splitting would be used.

$$\begin{cases} F_{X_1} & U > p_1 \\ F_{X_2} & \text{otherwise} \end{cases}$$

```
mix_norm <- function(n, p1, mean1, sd1, mean2, sd2) {
  x1 <- rnorm(n, mean = mean1, sd = sd1)
  x2 <- rnorm(n, mean = mean2, sd = sd2)
  k <- as.integer(runif(n) > p1)
```

```
  k * x1 + (1 - k) * x2
}
```

Try various p_1, from 0.1 to 1

```
mixture <-
  foreach(p1 = 0:10 / 10, .combine = bind_rows) %do% {
    tibble(
      value = mix_norm(n = 1000, p1 = p1, mean1 = 0, sd1 = 1, mean2 = 3, sd2 = 1),
      key = rep(p1, 1000)
    )
  }
```

```
mixture %>%
  ggplot(aes(x = value, colour = factor(key))) +
  stat_density(geom = "line", position = "identity") +
  scale_colour_discrete(
    name = expression(p[1]),
    labels = 0:10 / 10
  ) +
  xlab("x")
```

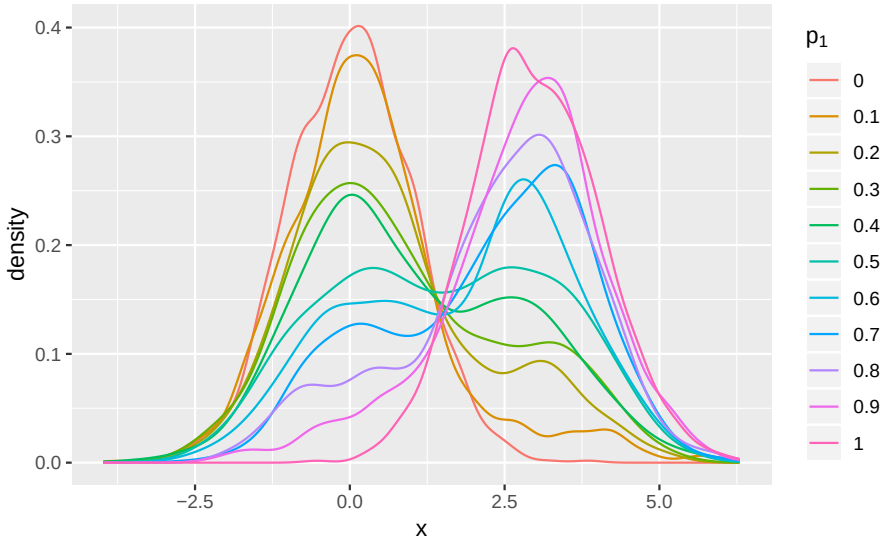


Figure 1.11: Mixture normal random number for each mixing probability

1.7 Multivariate Normal Random Vector

Definition 1.6 (Multivariate normal random vector)**.** A random vector $\mathbf{X} = (X_1, \ldots, X_p)^T$ follows multivariate normal distribution if

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|} \exp\left[-\frac{1}{2}(\mathbf{x}\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}\boldsymbol{\mu}) \right]$$

Remark. Let $\mathbf{Z} \sim MVN(\mathbf{0}, I)$. Then

$$\Sigma^{\frac{1}{2}}\mathbf{Z} + \boldsymbol{\mu} \sim MVN(\boldsymbol{\mu}, \Sigma) \tag{1.1}$$

From this remark, we get to generate *standard normal random vector*.

1.7.1 Spectral decomposition method

Note that covariance matrix is symmetric.

Theorem 1.3 (Spectral decomposition)**.** *Suppose that* Σ *is symmetric. Then*

$$\Sigma = P\Lambda P^T$$

where $(\mathbf{v}_j, \lambda_j)$ *corresponding eigenvector-eigenvalue*

$$\begin{cases} P = \begin{bmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_p \end{bmatrix} \in \mathbb{R}^{p\times p} \text{ orthogonal} \\ \Lambda = diag(\lambda_1, \ldots, \lambda_p) \end{cases}$$

Corollary 1.3. *Suppose that* Σ *is symmetric. Then*

$$\Sigma^{\frac{1}{2}} = P\Lambda^{\frac{1}{2}}P^T$$

where $\Lambda^{\frac{1}{2}} = diag(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})$

`eigen()` performs spectral decomposition. `$values` has eigenvalues and `$vectors` has eigenvectors. We first generate matrix that consists of standard normal random vector:

$$\begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1p} \\ Z_{21} & Z_{22} & \cdots & Z_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{np} \end{bmatrix}$$

Denote that each observation is row. To use Equation (1.1), we should multiply $\Sigma^{\frac{1}{2}}$ behind this matrix, not in front of. $\boldsymbol{\mu}$ matrix should be also made to matrix, in form of

$$\begin{bmatrix} \mu_{11} & \mu_{12} & \cdots & \mu_{1p} \\ \mu_{11} & \mu_{22} & \cdots & \mu_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ \mu_{11} & Z_{n2} & \cdots & \mu_{1p} \end{bmatrix} \in \mathbb{R}^{n\times p}$$

```
rmvn_eigen <- function(n, mu, sig) {
  d <- length(mu)
  ev <- eigen(sig, symmetric = TRUE)
  lambda <- ev$values
  P <- ev$vectors
  sig2 <- P %*% diag(sqrt(lambda)) %*% t(P)
  Z <- matrix(rnorm(n * d), nrow = n, ncol = d)
  X <- Z %*% sig2 + matrix(mu, nrow = n, ncol = d, byrow = TRUE)
  colnames(X) <- paste0("x", 1:d)
  X %>% tbl_df()
}
```

```
# mean vector -------------------------------
mu <- c(0, 1, 2)
# symmetric matrix --------------------------
sig <- matrix(numeric(9), nrow = 3, ncol = 3)
diag(sig) <- rep(1, 3)
sig[lower.tri(sig)] <- c(-.5, .5, -.5) * 2
sig <- (sig + t(sig)) / 2
```

Generate

$$\mathbf{X}_i \sim MVN\left((0,1,2), \begin{bmatrix} 1 & -0.5 & 0.5 \\ -0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{bmatrix}\right)$$

```
(mvn3 <- rmvn_eigen(1000, mu = mu, sig = sig))
#> # A tibble: 1,000 x 3
#>         x1       x2    x3
#>      <dbl>    <dbl> <dbl>
#>  1 -0.168   1.41    1.80
#>  2  1.39   -0.00942 2.40
#>  3 -0.710   1.30    1.37
#>  4  0.0314  2.04    1.80
#>  5  0.177   0.568   1.71
#>  6 -0.960   1.23    1.61
#>  7 -1.01    1.28    0.106
#>  8  0.272   0.0842  2.12
#>  9  0.148   1.63    2.53
#> 10 -1.24    1.53    1.28
#> # ... with 990 more rows
```

```
mvn3 %>%
  GGally::ggpairs(
    lower = list(continuous = GGally::wrap(gg_scatter, size = 1))
  )
```

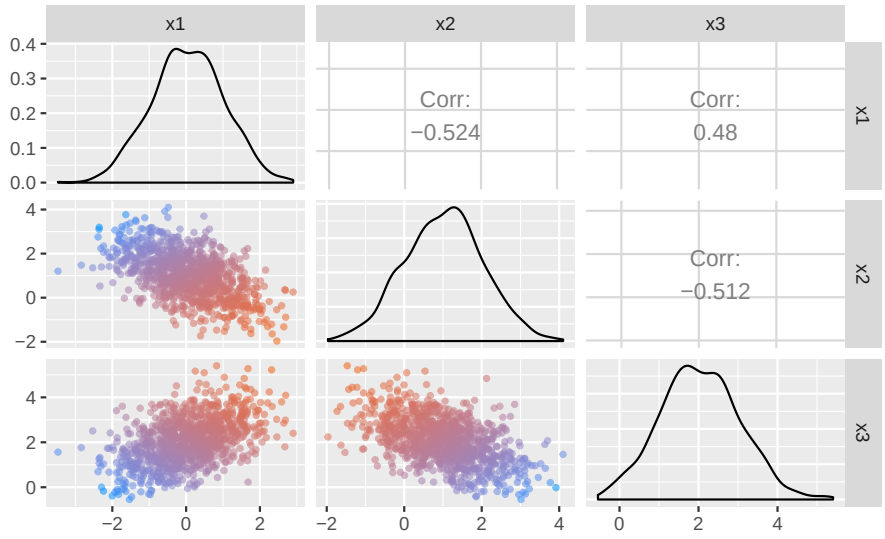


Figure 1.12: Multivariate normal random vector - spectral decomposition method

1.7.2 Singular value decomposition

SVD can be said to be a kind of generalization of spectral decomposition. This method can be used for any matrix, i.e. non-symmetric matrix. For Σ, SVD and spectral decomposition is equivalent. However, SVD does not account for symmetric property, so this method is less efficient compared to spectral decomposition.

```
rmvn_svd <- function(n, mu, sig) {
  d <- length(mu)
  S <- svd(sig)
  sig2 <- S$u %*% diag(sqrt(S$d)) %*% t(S$v)
  Z <- matrix(rnorm(n * d), nrow = n, ncol = d)
  X <- Z %*% sig2 + matrix(mu, nrow = n, ncol = d, byrow = TRUE)
  colnames(X) <- paste0("x", 1:d)
  X %>% tbl_df()
}
```

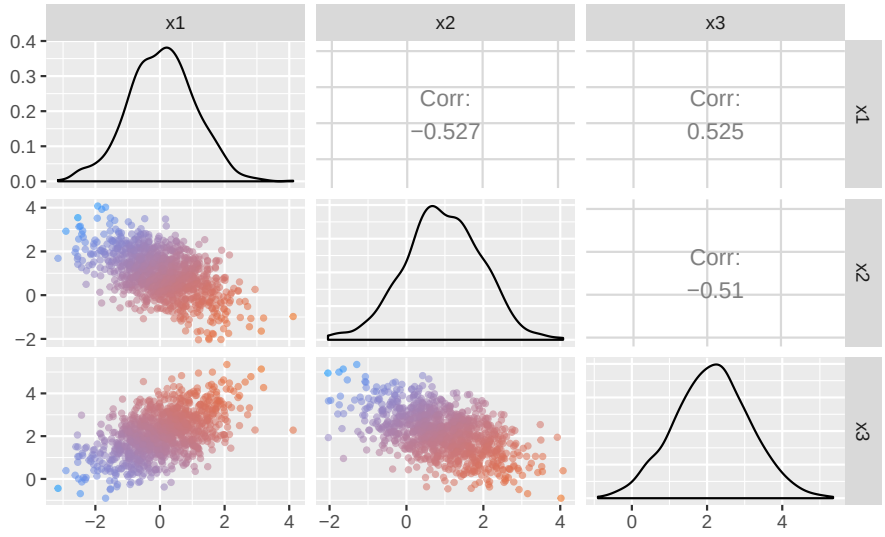


Figure 1.13: Multivariate normal random vector - svd

1.7.3 Choleski decomposition

Theorem 1.4 (Cholesky decomposition)**.** *Suppose that Σ is symmetric and positive definite. Then*

$$\Sigma = Q^T Q$$

where Q is an upper triangular matrix.

Corollary 1.4. *Suppose that Σ is symmetric and positive definite. For cholesky decomposition 1.4, define*

$$\Sigma^{\frac{1}{2}} = Q$$

`chol()` computes cholesky decomposition. In `R`, it gives upper triangular Q. Since some statements cholesky decomposition by $\Sigma = LL^T$ with lower triangular matrix, try not to confuse.

```
rmvn_chol <- function(n, mu, sig) {
  d <- length(mu)
  sig2 <- chol(sig)
  Z <- matrix(rnorm(n * d), nrow = n, ncol = d)
```

```
  X <- Z %*% sig2 + matrix(mu, nrow = n, ncol = d, byrow = TRUE)
  colnames(X) <- paste0("x", 1:d)
  X %>% tbl_df()
}
```

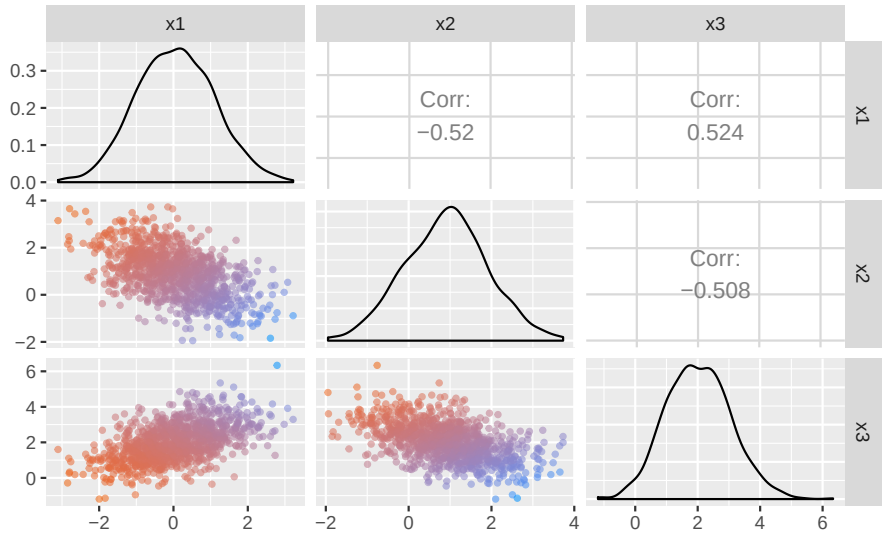


Figure 1.14: Multivariate normal random vector - cholesky decomposition

1.8 Stochastic Processes

Definition 1.7 (Stochastic process)**.** A stochastic process is a collection $\{X(t) : t \in T\}$ of random variables indexed by the set T. The index set T could be discrete or continous.

A State space is called te set of possible values that $X(t)$ can take.

Definition 1.8 (Discrete Time Markov Chain)**.** $\{X_n : n = 0, 1, 2, \ldots\}$ is a Discrete time markov chain on S if and only if

1. S is at most countable
2. Markov property $P(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j \mid X_n = i) = P_{ij}$

If P_{ij} is fixed, then $\{X_n\}$ is called time homogeneous. Otherwise, it is called nonhomogeneous.

Definition 1.9 (Random walk model)**.** Let $\{Y_n : n \in \mathbb{N}\}$ be an IID process on S s.t.

$$P(Y_n = k) = p_k$$

Define

$$S_n := \begin{cases} 0 & n = 0 \\ S_0 + Y_1 + \cdots + Y_n & n \in \mathbb{N} \end{cases}$$

1.8.1 Gambler's ruin model

Definition 1.10 (Gambler's ruin model)**.** Let $\{Y_n : n \in \mathbb{N}\}$ be a process on $\{-1, 1\}$ s.t.

$$P(Y_n = 1) = p, \quad P(Y_n = -1) = 1 - p$$

Define

$$X_n := \begin{cases} a & n = 0 \\ a + Y_1 + \cdots + Y_n & n \in \mathbb{N} \end{cases}$$

Example 1.8 (Gambling with coin)**.** Suppose that A and B each start with a stake of \$10, and bet \$1 on consecutive coin flips. The game ends when either one of the players has all the money. Let S_n be the fortune of player A at time n Then $\{S_n, n \geq 0\}$ is a symmetric random walk with absorbing barriers at 0 and 20. Simulate a realization of the process $\{S_n, n \geq 0\}$ and plot S_n vs the time index from time 0 until a barrier is reached.

Here we have

$$P(Y_n = 1) = P(Y_n = -1) = \frac{1}{2}$$

$$X_n := \begin{cases} 10 & n = 0 \\ 10 + Y_1 + \cdots + Y_n & n \in \mathbb{N} \end{cases}$$

```
gambling <- function(begin = 10, betting = 1, prob = .5) {
  N <- begin * 2
  sa <- begin
  record <- tibble(a = begin, b = begin)
  while(all(record > 0)) {
    sa <- ifelse(runif(1) >= prob, sa + betting, sa - betting)
    record <-
      record %>%
      bind_rows(c(a = sa, b = N - sa))
    if (sa == N) break()
  }
  record %>%
    mutate(idx = 1:n()) %>%
    select(idx, a, b)
}
```

```
gambling(begin = 10, betting = 1, prob = .5) %>%
  gather(-idx, key = "player", value = "fortune") %>%
  ggplot(aes(x = idx, y = fortune, colour = player)) +
  geom_path() +
  geom_point(alpha = .5, size = 1) +
  geom_hline(yintercept = c(0, 20), col = I("grey")) +
  labs(
    x = "Betting",
    y = "Fortune"
  )
```

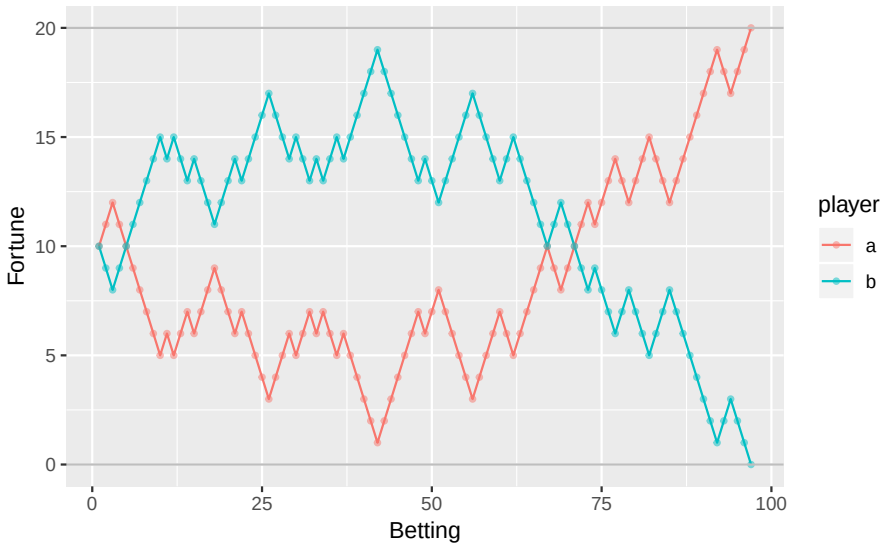


Figure 1.15: Sample path of Gambler's ruin model

In Figure 1.15, we can see the result of process with $p = \frac{1}{2}$. In fact, this process with probability 0.5 is also called *symmetric random walk*.

1.8.2 Homogeneous poisson process

Definition 1.11 (Count process)**.** A stochastic process $\{N(t) : t \geq 0\}$ where $N(t)$ is total number of events that occur by time t is called counting process.

1. $N(t) \geq 0$
2. $N(t) \in \mathbb{Z}$
3. $s \leq t \Rightarrow N(s) \leq N(t)$
4. For $s < t$, $N(t) - N(s) =$ the number of events that occur in $(s, t]$

Poisson process is one of this counting process.

Definition 1.12 (Poisson process)**.** The counting process $\{N(t), t \geq 0\}$ is said to be a Poisson process with rate $\lambda > 0$

1. $N(0) = 0$
2. $N(t) \perp\!\!\!\perp N(t+s) - N(t)$
3. Distribution of $N(t+s) - N(t)$ is the same for all values of t
4. $\lim_{h \to 0} \frac{P(N(h)=1)}{h} = \lambda$
5. $\lim_{h \to 0} \frac{P(N(h) \geq 2)}{h} = 0$

Remark.

$$\{N(t), t \geq 0\} \sim PP(\lambda) \Rightarrow N(t) \sim Poisson(\lambda t)$$

We can generate Poisson process using this relationship. However, it is slow. Thus, we find another way.

Theorem 1.5 (Campbell's Theorem)**.** *Let $\{N(t), t \geq 0\} \sim PP(\lambda)$, let X_t be the interarrival time, and let S_t be the waiting time until t-th event, i.e. $S_t := \sum_{i=1}^{t} X_i$. Then*

$$S_1, S_2, \ldots, S_n \mid N(t) = n \overset{d}{=} (U_{(1)}, U_{(2)}, \ldots, U_{(n)})$$

where $U_i \sim unif(0, t)$.

This Campbell's theorem gives solution to the PP generation.

Algorithm 7: Fast algorithm for Poisson Process

input : end time T

1 Generate $N \sim Poisson(\lambda T)$;

2 For N, generate $U_1, \ldots, U_N \overset{iid}{\sim} unif(0, T)$;

3 Sort $U_1, \ldots, U_N$ in ascending order, i.e. $\{U_{(1)}, \ldots, U_{(N)}\}$;

4 Set $S_1 = U_{(1)}, \ldots, S_N = U_{(N)}$;

output: $\begin{bmatrix} 1 & S_1 \\ 2 & S_2 \\ \cdots & \cdots \\ N & S_n \end{bmatrix}$

$\Rightarrow \{N(S_n)\} \sim PP(\lambda)$ on $[0, T]$

```
rpp <- function(lambda, t0) {
  N <- rpois(1, lambda = lambda * t0)
  tibble(sn = runif(N) * t0) %>%
    arrange(sn) %>% # arrival time
    mutate(pp = 1:n()) # N(sn) ~ PP
}
```

```
rpp(lambda = 1, t0 = 50) %>%
  mutate(
    true_mean = sn, # sn * lambda
    true_sd = sqrt(sn) # sn * lambda
  ) %>%
  ggplot(aes(x = sn)) +
  geom_ribbon(
    aes(ymin = true_mean - true_sd, ymax = true_mean + true_sd),
    fill = "grey70",
    alpha = .5
  ) +
  geom_line(aes(y = true_mean), col = I("white"), size = 2) +
  geom_path(aes(y = pp)) +
  labs(
    x = "t",
    y = expression(N(t))
  )
```

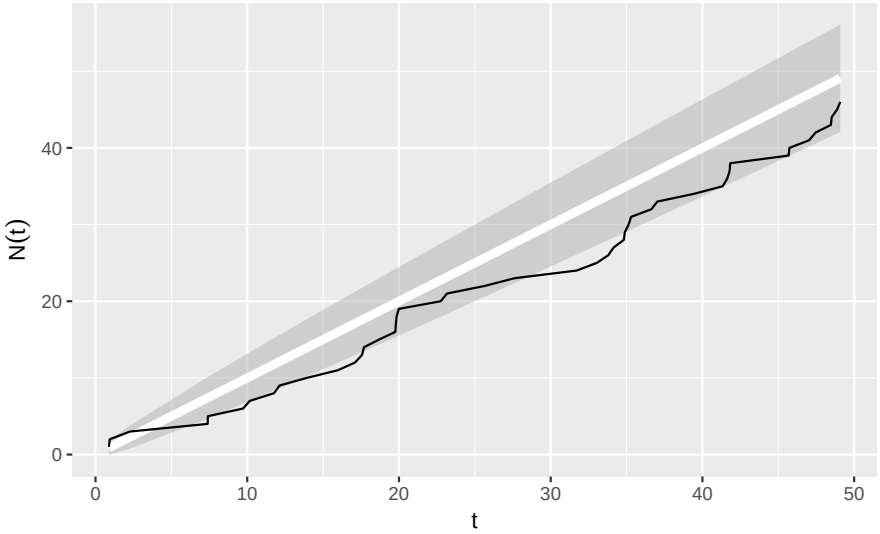


Figure 1.16: Sample path of Poisson process

1.8.3 Nonhomogeneous poisson process

Last section, what we have seen was homogeneous PP whose distribution does not depend on t. On the other hand, this condition can be broken.

Definition 1.13 (Nonhomogeneous Poisson Process)**.** The counting process $\{N(t), t \geq 0\}$ is said to be a Nonhomogeneous Poisson process with rate $\lambda(t) > 0$ if the third condition does not hold.

1. $N(0) = 0$
2. $N(t) \perp\!\!\!\perp N(t+s) - N(t)$
3. $\lim_{h \to 0} \frac{P(N(h)=1)}{h} = \lambda$
4. $\lim_{h \to 0} \frac{P(N(h) \geq 2)}{h} = 0$

$\lambda(t)$ is called intensity at time t.

$m(t) := \int_0^t \lambda(s)ds, t \geq 0$ is called mean-value function.

We can generate this NPP by Bernoulli splitting, which is called *thining approach.*

Lemma 1.1 (Thinning approach)**.** *Choosing λ s.t. $\forall t \leq T \lambda(t) \leq \lambda$. If an event of $PP(\lambda)$ counted with*

$$p(t) = \frac{\lambda(t)}{\lambda}$$

then the process follows $NPP(\lambda(t))$ on $[0, T]$

```
Set t = 0, N = 0;
repeat
    Generate Y ~ Exp(λ);
    Set t ← t + Y;
    Generate U ~ unif(0, 1);
    if U ≤ λ(t)/λ then
        Set N ← N + 1;
        Set S(N) ← t;
until t > T;
```

Chapter 2

Monte Carlo Integration and Variance Reduction

2.1 Monte Carlo Integration

Consider integration problem of a integrable function $g(x)$. We want to compute

$$\theta \equiv \int_a^b g(x)dx$$

For instance, standard normal cdf.

Example 2.1 (Standard normal cdf)**.** Compute values for

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right)dt$$

It might be impossible to compute this integral with hand. So we implement *simulation* concept here, based on the following theorems.

Theorem 2.1 (Weak Law of Large Numbers)**.** *Suppose that* $X_1, \ldots, X_n \overset{iid}{\sim} (\mu, \sigma^2 < \infty)$*. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$$

Let g *be a measurable function. Then*

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{p} g(\mu)$$

Theorem 2.2 (Strong Law of Large Numbers)**.** *Suppose that* $X_1, \ldots, X_n \overset{iid}{\sim} (\mu, \sigma^2 < \infty)$*. Then*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mu$$

Let g *be a measurable function. Then*

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \overset{a.s.}{\to} g(\mu)$$

2.1.1 Simple Monte Carlo estimator

Theorem 2.3 (Monte Carlo Integration)**.** *Consider integration* (2.1)*. This can be approximated via appropriate pdf* $f(x)$ *by*

$$\hat{\theta}_M = \frac{1}{N}\sum_{i=1}^{N} g(X_i)$$

Suppose that we have a distribution $f(x) = I_{sptg}(x)$, i.e. *uniform distribution.* Let $sptg = (a, b)$.

$$\begin{aligned}
\theta &\equiv \int_{sptg} g(x)dx \\
&= \int_a^b g(x)dx \\
&= \int_0^1 g(a + (b-a)t)(b-a)dt \\
&\equiv \int_0^1 h(t)dt \\
&= \int_0^1 h(t) I_{(a,b)}(t)dt \\
&= E[h(U)] \qquad U \sim unif(0,1)
\end{aligned} \tag{2.1}$$

By *the Strong law of large numbers* 2.2,

$$\frac{1}{n}\sum_{i=1}^{n} h(U_i) \overset{a.s.}{\to} E\Big[h(U)\Big] = \theta$$

where $U \sim unif(0,1)$. Thus, what we have to do here are two things.

1. representing g as h.
2. generating lots of U_i

Go back to Example 2.1.

Solution. Case 1: $x > 0$

Since $\Phi(x)$ is symmetry,

$$\Phi(0) = \frac{1}{2}$$

Fix $x > 0$.

$$\begin{aligned}
\int_0^x \exp\Big(-\frac{t^2}{2}\Big)dt &= \int_0^x x \exp\Big(-\frac{t^2}{2}\Big)\frac{I_{(0,x)}(t)}{x}dt \\
&\approx \frac{1}{N}\sum_{i=1}^{N} x \exp\Big(-\frac{U_i^2}{2}\Big)
\end{aligned}$$

with $U_1, \ldots, U_N \overset{iid}{\sim} unif(0, x)$.

Case 2: $x \leq 0$

Recall that $\Phi(x)$ is symmetry.

Hence,

$$\hat{\Phi}(x) = \begin{cases} \frac{1}{\sqrt{2\pi}} \frac{1}{N} \sum_{i=1}^{N} x \exp\left(-\frac{U_i^2}{2}\right) + \frac{1}{2} \equiv \hat{\theta}(x) & x \geq 0 \\ 1 - \hat{\theta}(-x) & x < 0 \end{cases}$$

```
phihat <- function(x, y) {
  yi <- abs(y)
  theta <- mean(yi * exp(-x^2 / 2)) / sqrt(2 * pi) + .5
  ifelse(y >= 0, theta, 1 - theta)
}
```

Then compute $\hat{\Phi}(x)$ for various x values.

```
phi_simul <- foreach(y = seq(.1, 2.5, length.out = 10), .combine = bind_rows) %do% {
  tibble(
    x = y,
    phi = pnorm(y),
    Phihat =
      tibble(x = runif(10000, max = y)) %>%
      summarise(cdf = phihat(x, y = y)) %>%
      pull()
  )
}
```

Table 2.1: Simple MC estimates of Normal cdf for each x

x	pnorm	mc
0.100	0.540	0.540
0.367	0.643	0.643
0.633	0.737	0.737
0.900	0.816	0.816
1.167	0.878	0.878
1.433	0.924	0.923
1.700	0.955	0.958
1.967	0.975	0.976
2.233	0.987	0.987
2.500	0.994	0.990

2.1.2 Hit-or-Miss Monte Carlo

Hit-or-Miss approach is another way to evaluate integrals.

Example 2.2 (Estimation of π)**.** Consider a circle in $\mathbb{R}$ coordinate.

$$x^2 + y^2 = 1$$

Since $y = \sqrt{1 - x^2}$,

$$\int_0^1 \sqrt{1-t^2}dt = \frac{\pi}{4} \tag{2.2}$$

By estimating Equation (2.2), we can estimate π, i.e.

$$\pi = 4\int_0^1 \sqrt{1-t^2}dt$$

Simple MC integration can also be used.

$$\begin{aligned}\int_0^1 \sqrt{1-t^2}dt &= \int_0^1 \sqrt{1-t^2}I_{(0,1)}(t)dt \\ &\approx \frac{1}{N}\sum_{i=1}^{N}\sqrt{1-U_i^2}\end{aligned}$$

```
circ <- function(x) {
  4 * sqrt(1 - x^2)
}
```

```
tibble(x = runif(10000)) %>%
  summarise(mc_pi = mean(circ(x)))
#> # A tibble: 1 x 1
#>   mc_pi
#>   <dbl>
#> 1  3.14
```

On the other way, hit-or-miss MC method applies geometric probability.

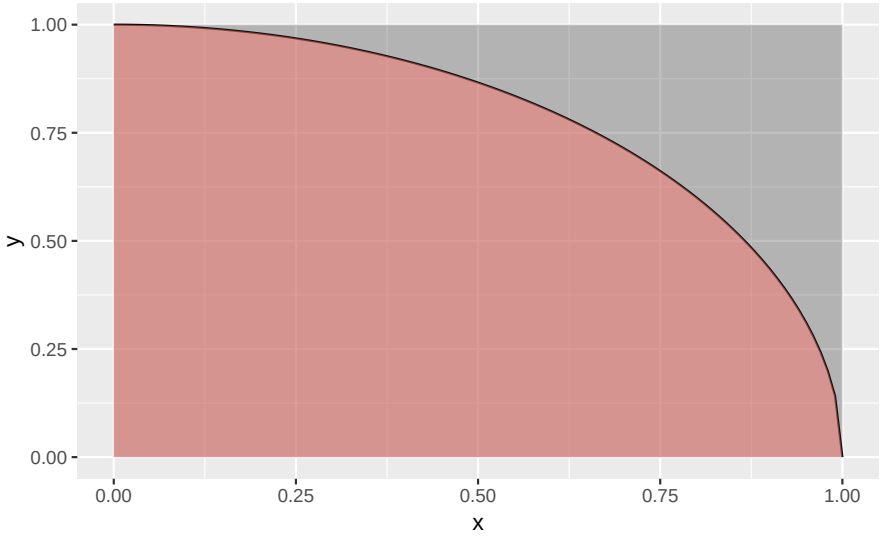


Figure 2.1: Hit-or-Miss

See Figure 2.1. From each coordinate, generate

- $X_i \overset{iid}{\sim} unif(0,1)$

- $Y_i \overset{iid}{\sim} unif(0,1)$

Then the proportion of $Y_i \leq \sqrt{1-X_i^2}$ estimates $\frac{\pi}{4}$.

```
tibble(x = runif(10000), y = runif(10000)) %>%
  summarise(hitormiss = mean(y <= sqrt(1 - x^2)) * 4)
#> # A tibble: 1 x 1
#>   hitormiss
#>       <dbl>
#> 1      3.15
```

2.2 Variance and Efficiency

We have seen two apporoaches doing the same task. Now we want to *evaluate them*. Denote that simple Monte Carlo integration 2.3 is estimating the *expected value of some random variable*. Proportion, which approximates probability is expected value of identity function.

The common statistic that can evaluate estimators expected value might be their variances.

2.2.1 Variance

Note that variance of sample mean is $Var(\overline{g(X)}) = \frac{Var(g(X))}{N}$. This property is one of estimating variance of $\hat{\theta}$.

$$\widehat{Var}(\hat{\theta}) = \frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^{N}(g(X_i) - \overline{g(X_i)})\right) = \frac{1}{N^2}\sum_{i=1}^{N}(g(X_i) - \overline{g(X_i)}) \tag{2.3}$$

For example,

```
tibble(x = runif(10000)) %>%
  summarise(mc_pi = var(circ(x)) / 10000)
#> # A tibble: 1 x 1
#>       mc_pi
#>       <dbl>
#> 1 0.0000795
```

However, this *variance of sample mean* is used in situation when we are in sample limitation situation. We do not have to stick to this. Now, Generating samples as many as we want is possible. So we try another approach: *parametric bootstrap*.

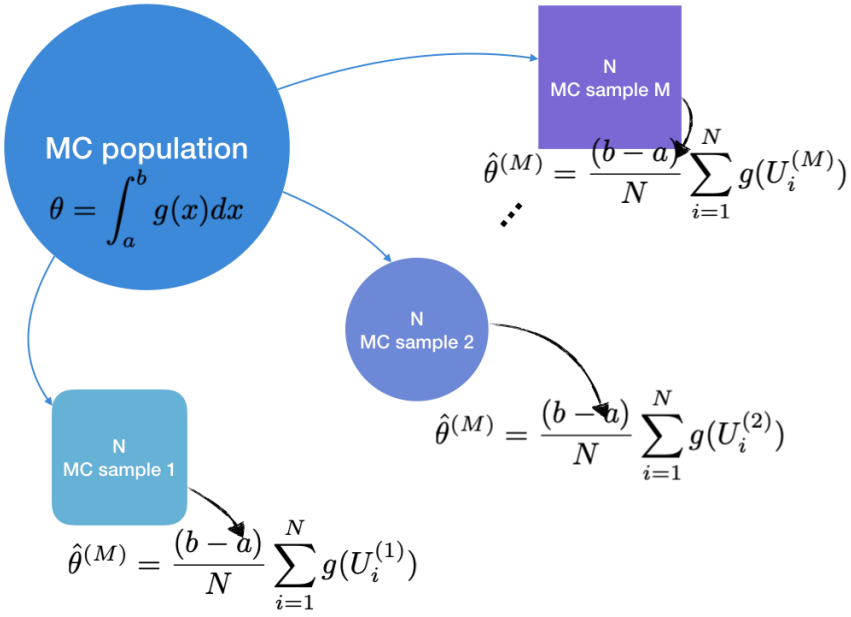


Figure 2.2: Empircal distribution of $\hat{\theta}$

See Figure 2.2. If we estimate $E\Big[g(U \sim unif(a,b))\Big]$, we can get θ. Generate M samples $\{U_1^{(j)}, \ldots, U_N^{(j)}\}, j = 1, \ldots M$ from this $U \sim unif(a,b)$. In each sample, calculate MC estimates $\hat{\theta}^{(j)}$. Now we have M MC estimates $\hat{\theta}$. This gives empirical distribution of $\hat{\theta}$. By *drawing a histogram*, we can see the outline.

Algorithm 8: Variance of $\hat{\theta}$

input : $\theta = \int_a^b g(x)dx$
for $m \leftarrow 1$ **to** M **do**
  Generate $U_1^{(m)}, \ldots, U_N^{(m)} \overset{iid}{\sim} unif(a,b)$;
  Compute $\hat{\theta}^{(j)} = \frac{(b-a)}{N}\sum g(U_i^{(j)})$;
end
$\bar{\hat{\theta}} = \frac{1}{M}\sum \hat{\theta}^{(j)}$;
$\widehat{Var}(\hat{\theta}) = \frac{1}{M-1}\sum(\hat{\theta}^{(j)} - \bar{\hat{\theta}})^2$;
output: $\widehat{Var}(\hat{\theta})$

Since we have to generate large size of data, `data.table` package will be used.

```
library(data.table)
```

Group operation can be used. Additional column (`sam`) would indicate group, and for each group MC operation would be processed. The following is the function generating `data.table` before group operation.

```
mc_data <- function(rand, N = 10000, M = 1000, char = "s", ...) {
  data.table(
    u = rand(n = N * M, ...),
    sam = gl(M, N, labels = paste0("s", 1:M))
  )
}
```

```
pi_mc <-
  mc_data(runif)[,
                 .(mc_pi = mean(circ(u))),
                 keyby = sam]
```

```
pi_mc %>%
  ggplot(aes(x = mc_pi)) +
  geom_histogram(bins = 30, col = gg_hcl(1), alpha = .7) +
  xlab(expression(pi)) +
  geom_vline(xintercept = pi, col = gg_hcl(2)[2])
```

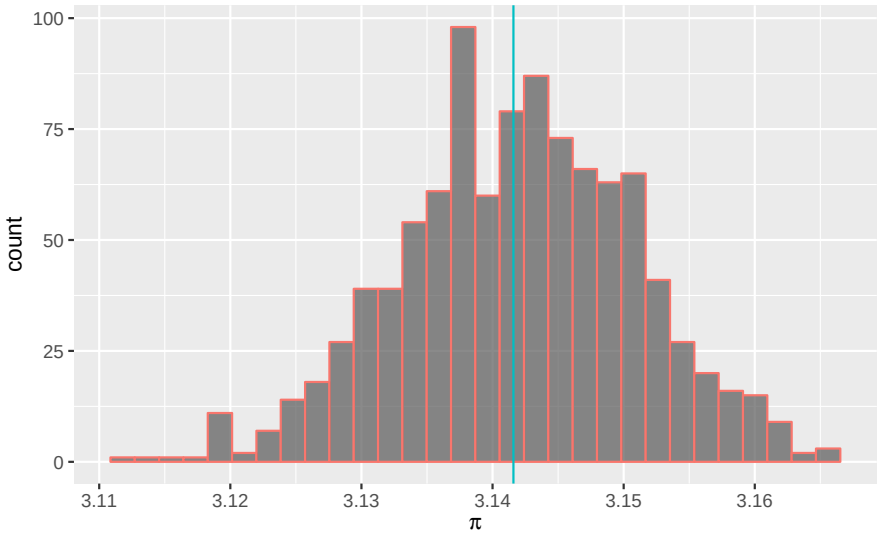


Figure 2.3: Empirical distribution of $\hat{\pi}$ by simple MC

As in Algorighm 9, we can compute the variance as below.

```
(mc_var <-
  pi_mc[,
        .(mc_variance = var(mc_pi))])
#>    mc_variance
#> 1:    8.09e-05
```

On the other hand, we need to generate two sets of random numbers for hit-or-miss MC.

```
pi_hit <-
  mc_data(runif)[
    , u2 := runif(10000 * 1000)
  ][,
    .(hitormiss = mean(u2 <= sqrt(1 - u^2)) * 4),
    keyby = sam]
```

```
pi_mc[pi_hit] %>%
  melt(id.vars = "sam", variable.name = "hat") %>%
  ggplot(aes(x = value, fill = hat)) +
  geom_histogram(bins = 30, alpha = .5, position = "identity") +
  xlab(expression(pi)) +
  geom_vline(xintercept = pi, col = I("red")) +
  scale_fill_discrete(
```

```
    name = "MC",
    labels = c("Simple", "Hit-or-Miss")
  )
```

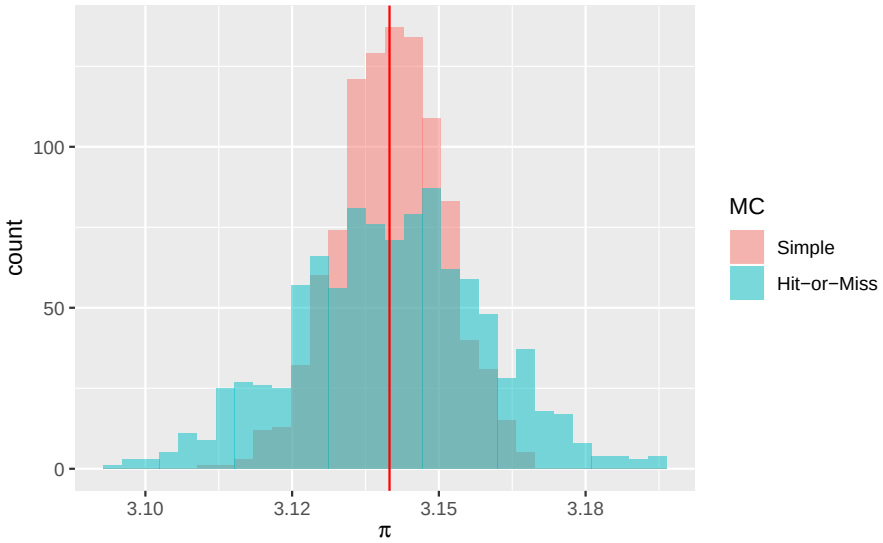


Figure 2.4: Simple MC and Hit-or-Miss MC

```
(hit_var <-
  pi_hit[,
         .(hit_variance = var(hitormiss))])
#>    hit_variance
#> 1:     0.000258
```

2.2.2 Efficiency

See Figure 2.4. It is obvious that Hit-or-Miss estimate produces larger variance than simple MC.

Definition 2.1 (Efficiency)**.** Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators for θ. Then $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if

$$\frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)} < 1$$

In other words, if $\hat{\theta}_1$ has smaller variance than $\hat{\theta}_2$, then $\hat{\theta}_1$ is said to be efficient, which is preferable.

Table 2.2: Simple MC versus Hit-or-Miss

SimpleMC	Hit-or-Miss	SimpleMCefficiency
0	0	TRUE

2.3 Variance Reduction

Consider Equation (2.3) based on $Var(\hat{\theta}) = \frac{\sigma^2}{N}$. This variance can always reduced by adding N. But we want to reduce variance less computationally.

2.3.1 Antithetic Variables

Consider correlated random variables U_1 and U_2. Then we have

$$Var\Big(\frac{U_1+U_2}{2}\Big) = \frac{1}{4}\Big(Var(U_1)+Var(U_2)+2Cov(U_1,U_2)\Big)$$

See the last term $Cov(U_1, U_2)$. If we generate U_{i1} and U_{i2} negatively correlated, we can get reduced variance than previous i.i.d. sample

$$Var\Big(\frac{U_1+U_2}{2}\Big) = \frac{1}{4}\Big(Var(U_1)+Var(U_2)\Big)$$

Lemma 2.1. *U and $1-U$ are identically distributed, but* negatively correlated.

1. $U \sim unif(0,1) \Leftrightarrow 1-U \sim unif(0,1)$
2. $Corr(U, 1-U) = -1$

This is well-known property of uniform distribution. Instead of generating N uniform numbers, try $\frac{N}{2}$ U_i and make corresponding $\frac{N}{2}$ $1-U_i$. This sequence becomes negatively correlated, so we can reduce the variance as mentioned.

When can we replace previous numbers with these *antithetic variables*? We usually plug-in the numbers in some function h to get Monte carlo integration. The thing is, our target is h, not U. $h(U)$ and $h(1-U)$ should *still be negatively correlated*. Hence, h should be *monotonic function*.

Corollary 2.1. *If $g = g(X_1, \ldots, X_n)$ is monotone, then*

$$Y = g(F_X^{-1}(U_1), \ldots, F_X^{-1}(U_n))$$

and

$$Y' = g(F_X^{-1}(1-U_1), \ldots, F_X^{-1}(1-U_n))$$

are negatively correlated.

Algorithm 9: Variance of $\hat{\theta}$ using antithetic variables

input : h : *monotonic*

for $m \leftarrow 1$ **to** M **do**

- Generate $U_{1,1}^{(m)}, \ldots, U_{\frac{N}{2},1}^{(m)} \overset{iid}{\sim} unif(0,1)$;
- Set $U_{i,2}^{(m)} := 1 - U_{i,1}^{(m)} \overset{iid}{\sim} unif(0,1)$;
- $\{U_i^{(m)}\}_1^N = \{U_{1,1}^{(m)}, \ldots, U_{\frac{N}{2},2}^{(m)}\}$;
- $\hat{\theta}^{(j)} = \frac{1}{N}\sum h(U_i^{(j)})$;

end

$\bar{\hat{\theta}} = \frac{1}{M}\sum \hat{\theta}^{(j)}$;

$\widehat{Var}(\hat{\theta}) = \frac{1}{M-1}\sum(\hat{\theta}^{(j)} - \bar{\hat{\theta}})^2$;

output: $\widehat{Var}(\hat{\theta})$

Check again Example 2.1. We have try to calculate

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{t^2}{2}\Big)dt$$

using simple monte carlo. To make the support $(0, 1)$, let $y = \frac{t}{x}$ be a change of variable. Then

$$\begin{aligned}\int_0^x \exp\left(-\frac{t^2}{2}\right)dt &= \int_0^1 x \exp\left(-\frac{(xy)^2}{2}\right)dy \\ &\approx \frac{1}{N}\sum_{i=1}^{N} x \exp\left(-\frac{(xU_i)^2}{2}\right)\end{aligned}$$

```
phiunif <- function(x, y) {
  yi <- abs(y)
  theta <- mean(yi * exp(-(yi * x)^2 / 2)) / sqrt(2 * pi) + .5
  ifelse(y >= 0, theta, 1 - theta)
}
```

Consider $\Phi(2)$.

```
phi2 <-
  mc_data(runif)[,
                 .(p2 = phiunif(u, y = 2)),
                 keyby = sam]
```

Now apply antithetic variables.

```
phi2_anti <-
  mc_data(runif, N = 10000 / 2)[,
                                u2 := 1 - u] %>%
  melt(id.vars = "sam", value.name = "U") %>%
  .[,
    .(anti_p2 = phiunif(U, y = 2)),
    keyby = sam]
```

```
phi2[phi2_anti] %>%
  melt(id.vars = "sam", variable.name = "hat") %>%
  ggplot(aes(x = value, fill = hat)) +
  geom_histogram(bins = 30, alpha = .5, position = "identity") +
  xlab(expression(pi)) +
  geom_vline(xintercept = pnorm(2), col = I("red")) +
  scale_fill_discrete(
    name = "MC",
    labels = c("Simple", "Antithetic")
  )
```

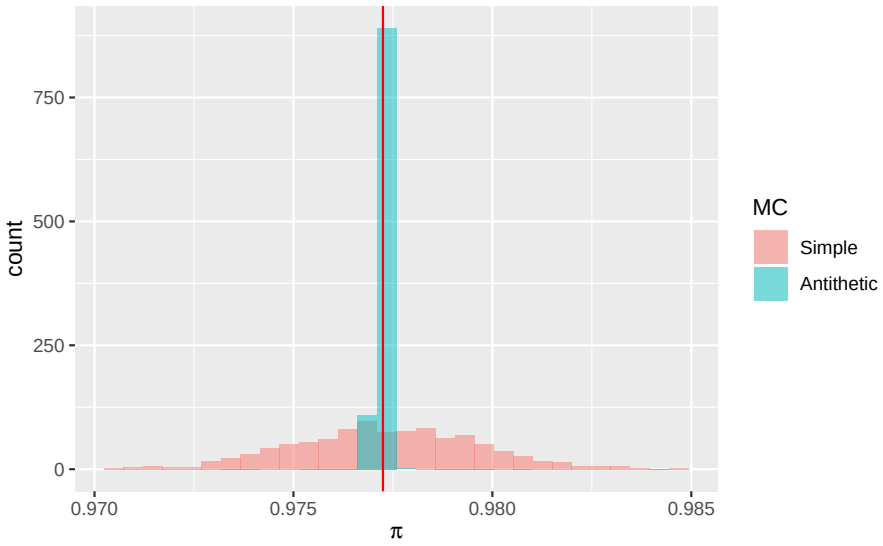


Figure 2.5: Use of antithetic variables

Obviously, variance has been reduced.

```
phi2[phi2_anti] %>%
  melt(id.vars = "sam", variable.name = "hat") %>%
  .[,
    .(variance = var(value)),
    by = hat]
#>         hat variance
#> 1:       p2  5.2e-06
#> 2: anti_p2  1.5e-08
```

2.3.2 Control Variates

Recall that we are trying to estimate $\theta = EX$ here in MC integration. Consider other output random variable. Suppose that $\mu_Y \equiv E(Y)$ is known. Then

$$X + c(Y - \mu_Y)$$

is an unbiased estimator for θ for any $c \in \mathbb{R}$.

2.4 Importance Sampling

Chapter 3

Monte Carlo Methods in Inference

3.1 Parametric Bootstrap

In this setting, we know distribution of X. We can freely generate from this distribution.

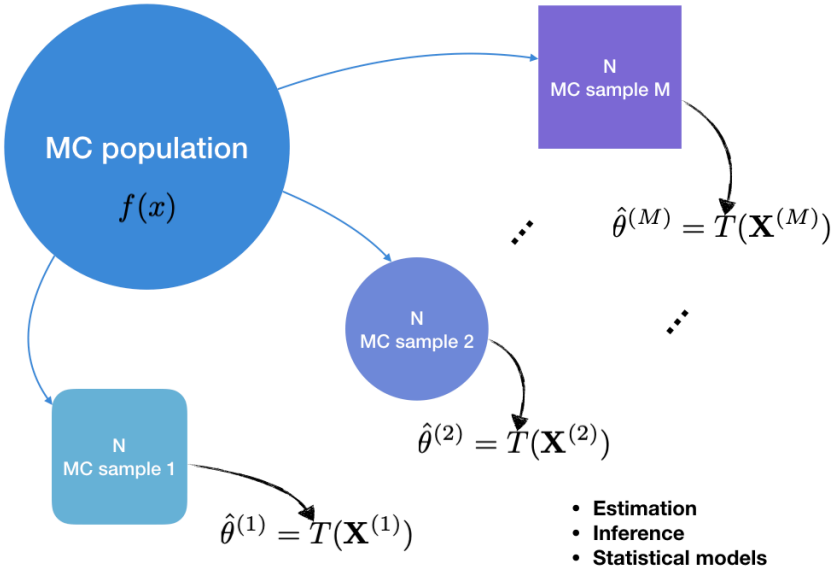


Figure 3.1: Parametric bootstrap

See Figure 3.1. From the "true" distribution, we can generate multiple samples. From each sample estimator can be computed. Then we can check these multiple estimates. Multiple estimates are close to motivation of estimator, so it helps exploring statistical inference with simple steps.

```
mc_data <- function(rand, N = 10000, M = 1000, char = "s", ...) {
  data.table(
    x = rand(n = N * M, ...),
    sam = gl(M, N, labels = paste0("s", 1:M))
  )
}
```

3.2 Monte Carlo Methods for Estimation

Example 3.1 (Any quantity of interest)**.** Suppose that $X_1, X_2 \overset{iid}{\sim} N(0,1)$. We want to estimate

$$\theta = E|X_1 - X_2|$$

3.2.1 Empirical distribution

Algorithm 10: Empirical distribution of $\hat{\theta}$

input : distribution f

for $m \leftarrow 1$ **to** M **do**

  Generate $(X_1^{(m)}, X_2^{(m)}) \overset{iid}{\sim} N(0,1)$;

  Compute $\hat{\theta}^{(m)} = |X_1^{(m)} - X_2^{(m)}|$;

end

Draw a histogram;

output: $\bar{\hat{\theta}} = \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}_m^{(m)}, \{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(M)}\}$

```
basicmc <-
  mc_data(rnorm, N = 2)[,
                        xname := gl(2, 1, length = 2000, labels = c("x1", "x2"))] %>%
  dcast(sam ~ xname, value.var = "x") %>%
  .[,
    .(that = mean(abs(x1 - x2))),
    by = sam]
```

```
basicmc[,
        .(est = mean(that))]
#>    est
#> 1: 1.1
```

```
basicmc %>%
  ggplot(aes(x = that)) +
  geom_histogram(bins = 30, col = gg_hcl(1), alpha = .7) +
  xlab(expression(theta))
```

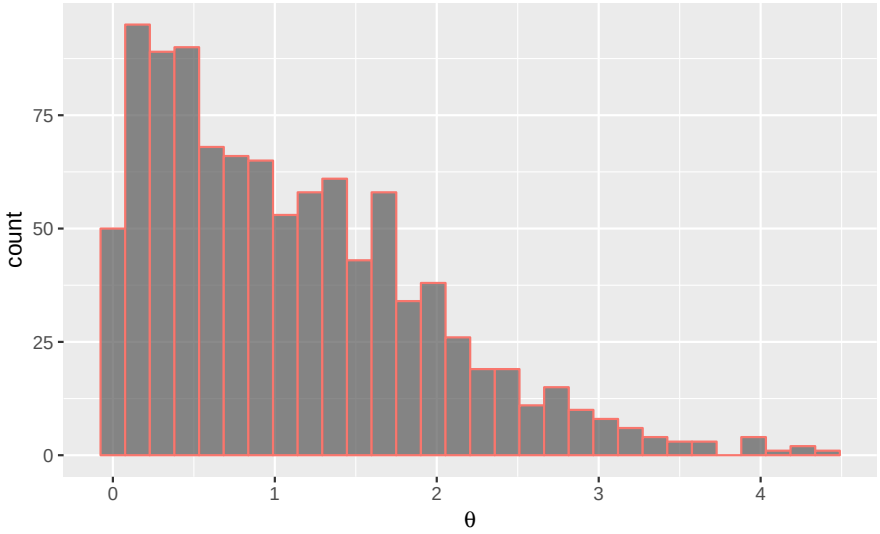


Figure 3.2: Empirical distribution of $\hat{\theta}$ for $|X_1 - X_2|$

3.2.2 Standard error

In Algorithm 11, we can get standard error by just calculating standard deviation of

$$\{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(M)}\}$$

Algorithm 11: Standard error of $\hat{\theta}$

input : distribution f

for $m \leftarrow 1$ **to** M **do**

  Generate $(X_1^{(m)}, X_2^{(m)}) \overset{iid}{\sim} N(0,1)$;

  Compute $\hat{\theta}^{(m)} = |X_1^{(m)} - X_2^{(m)}|$;

end

$\bar{\hat{\theta}} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m^{(m)}$;

$\widehat{SE}(\hat{\theta}) = \sqrt{\frac{1}{M-1} \sum_{m=1}^{M} (\hat{\theta}^{(m)} - \bar{\hat{\theta}})}$;

output: $\widehat{SE}(\hat{\theta})$

```
basicmc[,
        .(se = sd(that))]
#>        se
#> 1: 0.844
```

3.2.3 Mean squared error

MSE is used when comparing several estimators.

Definition 3.1 (Mean squared error)**.**

$$MSE(\hat{\theta}) := E(\hat{\theta} - \theta)^2$$

To know MSE, however, we should compute expectation. Some of them might be complicated even though we know true distribution. As the last chapter, we can apply Monte carlo method.

Example 3.2 (MSE of a trimmed mean)**.** Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} N(2,1)$. Consider three estimators for $\mu = 2$.

1. mean $\overline{X}$
2. median $\tilde{X}$
3. kth trimmed mean $\overline{X}_{[-k]}$

Algorithm 12: MSE of mean, median, and kth trimmed mean

input : distribution f

for $m \leftarrow 1$ **to** M **do**

- Generate $(X_1^{(m)}, \ldots, X_N^{(m)}) \overset{iid}{\sim} N(2,1)$;
- Sort $(X_1^{(m)}, \ldots, X_N^{(m)})$ in increasing order, i.e. $(X_{(1)}^{(m)}, \ldots, X_{(N)}^{(m)})$;
- Mean $\overline{X}^{(m)} = \frac{1}{N} \sum_{i=1}^{N} X_i^{(m)}$;
- Median $\tilde{X}^{(m)} = \begin{cases} X_{\frac{N}{2}+1}^{(m)} & N \text{ odd} \\ \frac{X_{\frac{N}{2}}^{(m)} + X_{\frac{N}{2}+1}^{(m)}}{2} & N \text{ even} \end{cases}$;
- kth trimmed mean $\overline{X}_{[-k]}^{(m)} = \frac{1}{N-2k} \sum_{i=k+1}^{n-k} X_{(i)}^{(m)}$

end

$\widehat{MSE}(\overline{X}) = \frac{1}{M} \sum_{m=1}^{M} (\overline{X}^{(m)} - 2)^2$;

$\widehat{MSE}(\tilde{X}) = \frac{1}{M} \sum_{m=1}^{M} (\tilde{X}^{(m)} - 2)^2$;

$\widehat{MSE}(\overline{X}_{[-k]}) = \frac{1}{M} \sum_{m=1}^{M} (\overline{X}_{[-k]}^{(m)} - 2)^2$;

output: $\widehat{MSE}(\overline{X})$, $\widehat{MSE}(\tilde{X})$, and $\widehat{MSE}(\overline{X}_{[-k]})$

```
trim <- function(x, k = 1) {
  n <- length(x)
  x <- sort(x)
  sum(x[(k + 1):(n - k)]) / (n - 2 * k)
}
#---------------------------------------
mu_list <- function(x, k) {
  list(mean = mean(x), median = median(x), trim = trim(x, k))
}
```

Try $k = 1$.

```
(trim_mc <-
  mc_data(rnorm, mean = 2, sd = 1)[,
                                   unlist(lapply(.SD, mu_list, k = 1)) %>% as.list,
                                   by = sam])
#>          sam x.mean x.median x.trim
#>    1:     s1   2.02     2.02   2.02
#>    2:     s2   2.00     2.00   2.00
#>    3:     s3   2.00     2.01   2.00
#>    4:     s4   1.99     1.98   1.99
#>    5:     s5   2.00     1.99   2.00
#>   ---
#>  996:   s996   2.02     2.02   2.02
```

```
#>  997:  s997   2.00     1.99   2.00
#>  998:  s998   1.99     1.99   1.99
#>  999:  s999   1.99     1.99   1.99
#> 1000: s1000   2.00     2.01   2.00
```

```
trim_mc %>%
  melt(id.vars = "sam", variable.name = "hat") %>%
  ggplot(aes(x = value, fill = hat)) +
  geom_histogram(bins = 30, alpha = .3, position = "identity") +
  xlab(expression(mu)) +
  geom_vline(xintercept = 2, col = I("red")) +
  scale_fill_discrete(
    name = "Estimates",
    labels = c("Mean", "Median", "Trimmed")
  )
```

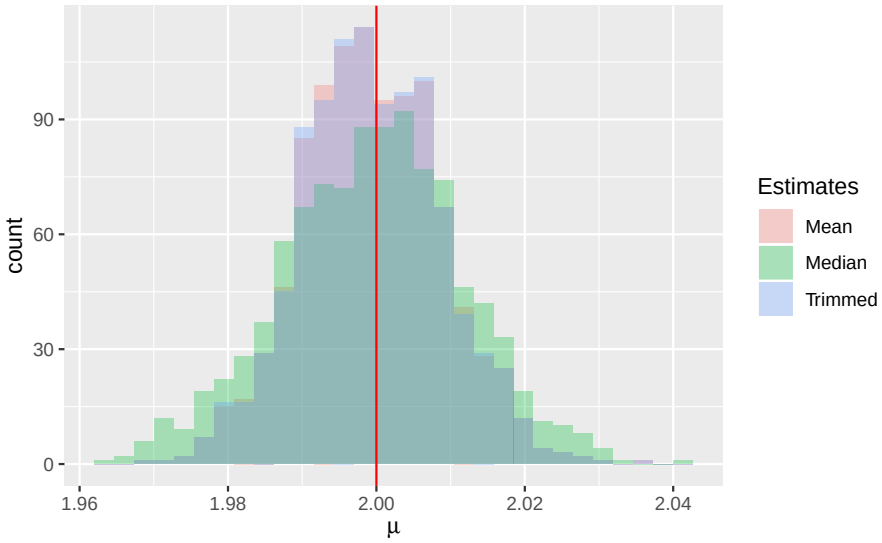


Figure 3.3: Empirical distribution of each estimator for $\mu = 2$

Here, median shows the largest standard error.

```
trim_mc[,
        lapply(.SD, sd),
        .SDcols = -"sam"]
#>     x.mean x.median  x.trim
#> 1: 0.00942   0.0123 0.00942
```

Now try various k for trimmed mean.

```
mse_list <- function(x, k) {
  list(mse = mean((x - 2)^2), se = sd(x))
}
#----------------------------------------
trim_mse <-
  mc_data(rnorm, mean = 2, sd = 1)[,
                                   lapply(.SD, function(x) {
                                     sapply(0:9, function(k) {
                                       trim(x = x, k = k)
```

```
                                    })
                                  }) %>%
                                    unlist() %>%
                                    as.list(),
                                  by = sam][,
                                            lapply(.SD, mse_list) %>%
                                              unlist() %>%
                                              as.list(),
                                            .SDcols = -"sam"]
```

```
trim_mse %>%
  transpose() %>%
  .[,
    `:=`(
      k = rep(0:9, each = 2),
      hat = gl(2, k = 1, length = 2 * 10, labels = c("mse", "se"))
    )] %>%
  dcast(k ~ hat, value.var = "V1")
#>     k      mse      se
#>  1: 0 9.83e-05 0.00992
#>  2: 1 9.83e-05 0.00992
#>  3: 2 9.83e-05 0.00992
#>  4: 3 9.83e-05 0.00992
#>  5: 4 9.82e-05 0.00992
#>  6: 5 9.83e-05 0.00992
#>  7: 6 9.83e-05 0.00992
#>  8: 7 9.83e-05 0.00992
#>  9: 8 9.83e-05 0.00992
#> 10: 9 9.83e-05 0.00992
```

3.3 Confidence interval

Remember the meaning of 95% confidence interval. *If we have 100 samples and construct confidence interval in each sample, 95 intervals would include true parameter.* In this Monte Carlo setting, we know true population distribution, so we can generate multiple samples. Thus, we can reproduce this confidence interval situation.

3.3.1 Empirical confidence interval

See one of histograms of Figure 3.3. Estimates are sorted. Calculating the upper and lower quantiles would give values close to confidence interval. See Figure 3.2. While the former show symmetric distribution, this is not. 0.25 and 0.975 quantile might be inappropriate. In this case, we should pick the *shortest interval*

with 95%. Best critical region leads to the shortest length of CI given α, so we are finding this one.

Algorithm 13: Empirical confidence interval by Monte Carlo method

input : distribution f
for $m \leftarrow 1$ **to** M **do**
  Generate $X_1^{(m)}, \ldots, X_n^{(m)} \overset{iid}{\sim} f$;
  Compute $\hat{\theta}^{(m)} = \hat{\theta}(\mathbf{X^{(m)}})$;
end
if *Distribution of* $\{\hat{\theta}^{(m)}\}_1^M$ *symmetric* **then**
  Sort $\{\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(M)}\}$ in decreasing order, i.e. $\{\hat{\theta}_{(1)}^{(1)}, \ldots, \hat{\theta}_{(M)}^{(M)}\}$;
  Compute $LB = \frac{\alpha}{2}$ sample quantile and $UB = 1 - \frac{\alpha}{2}$ sample quantile;
else
  foreach $lb < 0.05$ *with* $ub - lb = 1 - \alpha$ **do**
    Candidate interval (lb, ub);
    calculate length $l_i = ub - lb$;
  end
  (LB, UB): pick up the interval with the smallest length l_i;
end
output: (LB, UB)

3.3.2 Empirical confidence level

On the contrary, we can estiamte confidence level given confidence interval.

Example 3.3 (Confidence interval for variance)**.** If $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then

$$T = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

Thus, $100(1-\alpha)\%$ confidence interval is given by

$$(0, \frac{(n-1)S^2}{\chi_\alpha^2(n-1)})$$

For each MC sample, compute confidence interval. Just check if *known true parameter* is in the interval. Its proportion becomes the confidence level. It is simpler that estimate confidence interval itself.

Algorithm 14: Empirical confidence level by Monte Carlo method

input : distribution f with parameter θ
for $m \leftarrow 1$ **to** M **do**
  Generate $X_1^{(m)}, \ldots, X_n^{(m)} \overset{iid}{\sim} f$;
  Compute the confidence interval C_m;
  Compute $Y_j = I(\theta \in C_m)$, i.e. whether θ is in the CI;
end
Empirical confidence level $\overline{Y} = \sum_{m=1}^{M} Y_m$;
output: $\overline{Y}$

Let $\mu = 0$, $\sigma = 2$, $N = 20$, and let $M = 1000$.

```
ci_var <- function(x, variance, alpha) {
  n <- length(x)
```

```
  s2 <- var(x)
  (n - 1) * s2 / qchisq(alpha, df = n - 1) > variance
}
#----------------------------
ci_lev <-
  mc_data(rnorm, N = 20, M = 1000, mean = 0, sd = 2)[,
                                                     .(hat = mean(ci_var(x, variance = 4, alpha = .05)))
                                                     by = sam]
```

```
ci_lev[,
       .N,
       by = hat][,
                 proportion := N / sum(N)] %>%
  ggplot(aes(x = hat, y = proportion, fill = factor(hat))) +
  geom_bar(stat = "identity") +
  scale_fill_discrete(
    name = "CI",
    labels = c("out", "in")
  ) +
  xlab(expression(y))
```

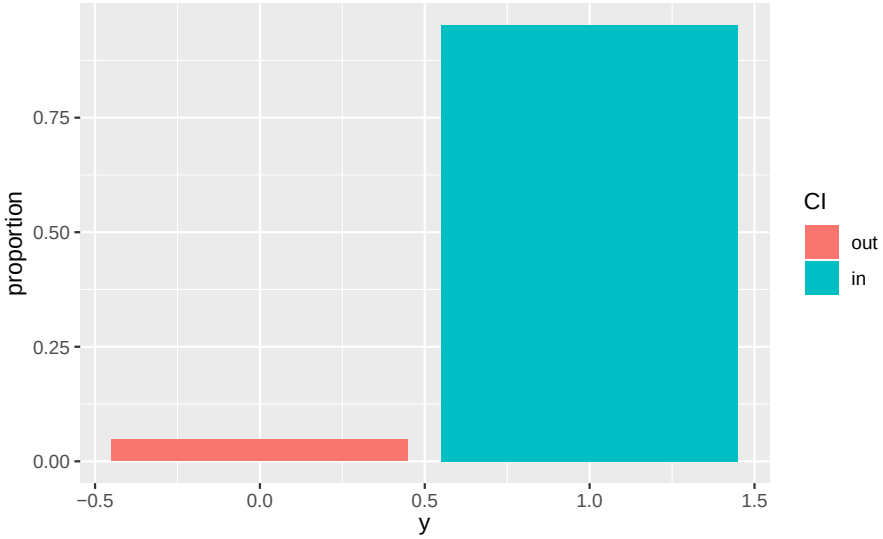


Figure 3.4: Proportion of σ^2 in confidence intervals

This leads to empirical confidence level, i.e. *sample proportion*. Just follow the last step 6 of Algorithm 15.

```
(ci_lev <-
  ci_lev[,
         .(level = mean(hat))])
#>    level
#> 1: 0.952
```

It is very close to 0.95. One of advantages of simulation study is we can assume various situation. For example, *violation of Gausiannity*.

Example 3.4 (Violation of Normal distribution assumption)**.** Refer to Example 3.3. This has assumed that $X_i \overset{iid}{\sim} N(\mu = 2, \sigma^2 = 4)$. What if not? For instance,

$$X_1, \ldots, X_n \overset{iid}{\sim} \chi^2(df = 2)$$

Just change random numbers.

```
ci_lev2 <-
  mc_data(rchisq, N = 20, M = 1000, df = 2)[,
                                            .(hat = mean(ci_var(x, variance = 4, alpha = .05))),
                                            by = sam][,
                                                      .(non_normal = mean(hat))]
```

Table 3.1: Empirical confidence level for each population

Normal	Chisq
0.952	0.763

From Table 3.1, we found that *non-normality lowers confidence level* from 0.952 to 0.763.

3.4 Hypothesis tests

Using MC method, we have done point estimation and interval estimation. Now consider *hypothesis testing*.

$$H_0 : \theta \in \Theta_0 \qquad \text{vs} \qquad H_1 : \theta \in \Theta_1$$

where $\{\Theta_0, \Theta_1\}$ is a partition of the parameter space Θ. First of all, we have *test statistic*

$$T(\mathbf{X}) \overset{H_0}{\sim} f$$

and f is called *null distribution*. Given observed data, we compute this test statistic T_0. Where T_0 is located in the null distribution f decides whether we reject or accept H_0. If T_0 is very far from the middle, we can say that the realized data set is very rare event under H_0. In this case, we reject H_0. Otherwise, accept it. This is why we compute the tail probability, p-value.

3.4.1 Empirical p-value

Example 3.5. Suppose that $X_1, \ldots, X_{10} \overset{iid}{\sim} Exp(\lambda = 1)$, which are observed as follows

$$`rxexp`$$

Let $\theta = E(X) = \frac{1}{\lambda}$.

$$H_0 : \theta = 0.5 \qquad \text{vs} \qquad H_1 : \theta > 0.5$$

Test using $T = \frac{\overline{X} - \theta_0}{S/\sqrt{n}}$ statistic.

Before looking at p-value, briefly look at *empirical null distribution* of test statistic.

```
mc_data(rexp, rate = 2)[,
                        .(tstat = t.test(x, mu = .5)$statistic),
                        by = sam] %>%
  ggplot(aes(x = tstat)) +
  geom_histogram(bins = 30, col = gg_hcl(1), alpha = .7) +
```

```
geom_vline(xintercept = t.test(xexp, mu = .5)$statistic, col = I("red")) + # xexp: observed data
geom_vline(xintercept = -t.test(xexp, mu = .5)$statistic, col = I("red")) +
xlab("T")
```

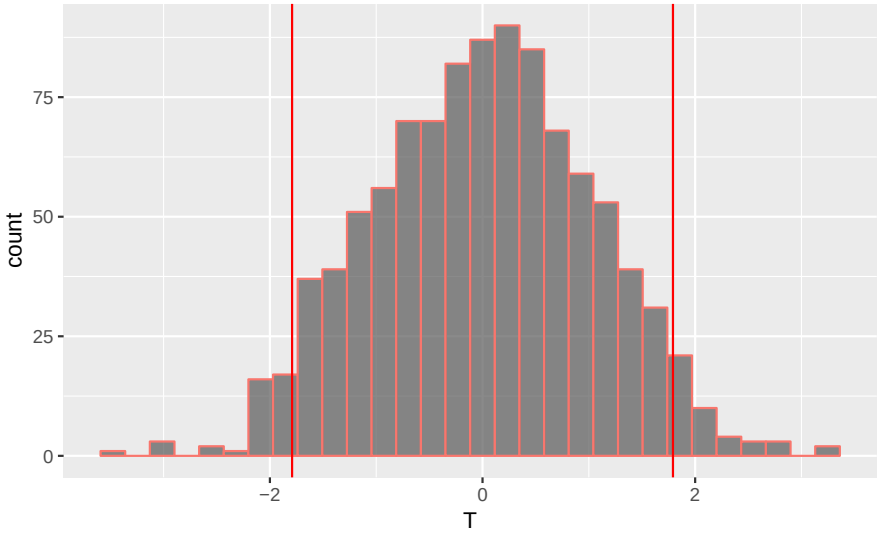


Figure 3.5: Emprirical Null Distribution

By proceeding the similar way, we can get empirical distribution of test statistics. Some are out of observed T_0, some are not. Motivation is that we just count these. Proportion of these would estimate p-value. Recap what p-value is.

Definition 3.2 (p-value)**.** Let T be test statistic and let T_0 be observed test statistic given data. Then p-value is

$$p-value := \begin{cases} P(|T| \geq T_0 \mid H_0) & \text{both sided} \\ P(T \geq T_0 \mid H_0) & \text{one sided} \\ P(T \leq T_0 \mid H_0) & \text{one sided} \end{cases}$$

Denote that p-value is probability. So in MC setting, we can estimate this by computing *sample mean of identity function.*

Lemma 3.1 (Empirical p-value)**.** *Let T_0 be observed test statistic and let $\{T_1, \ldots, T_M\}$ be test statistic computed in each MC sample.*

$$\textit{Empirical p-value} = \begin{cases} \frac{\left|\{T_j : (T_j > |T_0|) \text{ or } (T_j < -|T_0|)\}\right|}{M} & \textit{both-sided} \\ \frac{\left|\{T_j : (T_j > T_0)\}\right|}{M} \text{ or } \frac{\left|\{T_j : (T_j < T_0)\}\right|}{M} & \textit{one-sided} \end{cases}$$

Algorithm 15: Empirical p-value by Monte Carlo method

input : Given observed data, compute T_0

for $m \leftarrow 1$ **to** M **do**

  Generate $X_1^{(m)}, \ldots, X_n^{(m)} \overset{H_0}{\sim} f$;

  Compute $T_m(\mathbf{X}^{(m)})$;

end

Empirical p-value $\hat{p} = \begin{cases} \dfrac{\left|\{T_j : (T_j > |T_0|) \text{ or } (T_j < -|T_0|)\}\right|}{M} & \text{both-sided} \\ \dfrac{\left|\{T_j : (T_j > T_0)\}\right|}{M} \text{ or } \dfrac{\left|\{T_j : (T_j < T_0)\}\right|}{M} & \text{one-sided} \end{cases}$;

output: $\hat{p}$

Go back to Example 3.5. Only left is computing 5 of Algorighm 16. (Denote that `xexp` in the code is vector object of observed data).

```
(tt_exp <-
  mc_data(rexp, rate = 2)[,
                          .(tstat = t.test(x, mu = .5)$statistic),
                          by = sam][,
                                    .(pval = mean(tstat > abs(t.test(xexp, mu = .5)$statistic)))])
#>      pval
#> 1: 0.046
```

It is smaller that `0.05`, so we reject H_0.

3.4.2 Comparing several tests

MC method would be used in comparing tests rather than conducting test itself. By generating random number, we can evaluate tests.

$$H_0 : \theta \in \Theta_0 \qquad \text{vs} \qquad H_1 : \theta \in \Theta_1$$

As mentioned earlier, $\{\Theta_0, \Theta_1\}$ is a partition of the parameter space Θ. For this test, we can perform several tests. Test method 1, test method 2, et cetera. All these methods produce error, but these errors might be different. So we try to compare this.

what is true	accept H_0	reject H_0
H_0	correct decision	*Type I error*
H_1	*Type II Error*	correct decision

In most tests, we aims to reject H_0. By rejecting it, we can evidently say that H_0 is not true. In this sense, we treat type I error more importantly that type II error in general. Test strategy becomes to control type I error probability first and then lower type II error probabilty.

Definition 3.3 (Power function)**.** Let $\theta \in \Theta$ be a parameter of a test.

$$\beta(\theta) := P(\text{reject } H_0 \mid \theta)$$

With this power function, each type I error and type II error probability is given.

Lemma 3.2 (typeerr)**.** *1.* $P(\textit{Type I error}) = \beta(\theta_0), \quad \theta_0 \in \Theta_0$

2. $\textit{Power } \beta(\theta_1) = 1 - P(\textit{Type II error}), \quad \theta_1 \in \Theta_1$

Following our test strategy, fixing $P(\text{Type I error})$ and maximizing $\beta(\theta_1)$, we construct following test.

Definition 3.4 (Size α Test)**.** A test with $\beta(\theta)$ is called size α test if and only if

$$\alpha := \sup_{\theta \in \Theta_0} \beta(\theta), \quad 0 \leq \alpha \leq 1$$

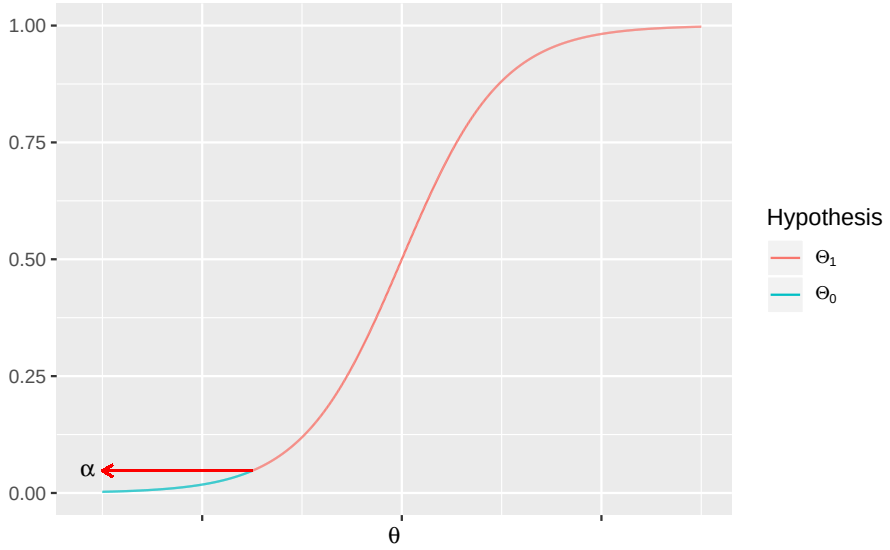


Figure 3.6: Size α Test

Then how to compare tests? Look at the following example. Three columns of the middle part are type I error rate.

test methods	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	Power
Test 1	0.09	0.04	0.01	0.7
Test 2	0.11	0.06	0.01	0.65
Test 3	0.15	0.07	0.02	0.9

Here, we will choose **Test 1**.

1. Type I error rate $\approx \alpha$
   - before looking at power, this should be satisfied.
   - So Test 3 is excluded
2. Larger power
   - Thus, we select Test 1.

3.4.3 Empirical type-I error rate

Recall Lemma 3.2. As in p-value, we just compute sample proportion for each type I error rate and power under null and alternative distribution.

Lemma 3.3. *Consider* $H_0 : \theta \in \Theta_0$ *vs* $H_1 : \theta \in \Theta_1$.

Define $I(\mathbf{X})$ *by*

$$I(\mathbf{X}) = \begin{cases} 1 & H_0 \text{ is rejected} \mid H_0 \\ 0 & \text{otherwise} \end{cases}$$

For each MC sample, compute this statistic $I_m = I(\mathbf{X}^m)$*. Then empirical type I error rate can be computed as*

$$\frac{1}{M}\sum_{m=1}^{M} I_m$$

Algorithm 16: Empirical type I error rate by Monte Carlo method

input : $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$

for $m \leftarrow 1$ **to** M **do**

  Generate $X_1^{(m)}, \ldots, X_n^{(m)} \overset{H_0}{\sim} f$;

  Compute $T_m(\mathbf{X}^{(m)})$;

  Compute $I_m = \begin{cases} 1 & H_0 \text{ is rejected} \mid H_0 \\ 0 & \text{otherwise} \end{cases}$;

end

Empirical Type I error rate $\hat{\alpha} = \frac{1}{M}\sum_{m=1}^{M} I_m$;

output: compare $\hat{\alpha}$ with α

Example 3.6 (Testing normal mean)**.** Suppose that $X_1, \ldots, X_{20} \overset{iid}{\sim} N(\mu, \sigma^2 = 100)$. Test

$$H_0 : \mu = 500 \qquad \text{vs} \qquad H_1 : \mu > 500$$

1. Z-test: $Z = \frac{\overline{X}-500}{\sigma^2/\sqrt{20}} \overset{H_0}{\sim} N(0,1)$
2. t-test: $T = \frac{\overline{X}-500}{S/\sqrt{20}} \overset{H_0}{\sim} t(20-1)$

```
test_list <- function(x, mu, sig, a = .05) {
  n <- length(x)
  xbar <- mean(x) - mu
  list(
    z = xbar / (sig / sqrt(n)) > qnorm(a, lower.tail = FALSE),
    t = xbar / (sd(x) / sqrt(n)) > qt(a, df = n - 1, lower.tail = FALSE)
  )
}
#--------------------
err_mc <-
  mc_data(rnorm, N = 20, mean = 500, sd = 10)[,
                                              lapply(.SD, test_list, mu = 500, sig = 10) %>%
                                                unlist() %>%
                                                as.list(),
                                              by = sam][,
                                                        lapply(.SD, mean),
                                                        .SDcols = -"sam"]
```

Both test have Type I error close to α, but Z-test seems bit better.

Table 3.4: Empirical Type I error for Z and T

Z-test	T-test
0.048	0.051

3.4.4 Empirical power

Next step is power. See Figure 3.6. Power is different in that this is computed in *alternative distribution, not null distribution.*

$$\beta(\theta_1) = P(\text{reject } H_0 \mid \theta_1 \in \Theta_1)$$

Lemma 3.4. *Consider* $H_0 : \theta \in \Theta_0$ *vs* $H_1 : \theta \in \Theta_1$.

Define $I(\mathbf{X})$ *by*

$$I(\mathbf{X}) = \begin{cases} 1 & H_0 \textit{ is rejected} \mid H_1 \\ 0 & \textit{otherwise} \end{cases}$$

For each MC sample, compute this statistic $I_m = I(\mathbf{X}^m)$. *Then empirical power can be computed as*

$$\frac{1}{M}\sum_{m=1}^{M} I_m$$

Process will be same but we test under H_1. However, this makes a lot difference due to structure of each hypothesis. In many cases, H_0 is simple, i.e. $\mu = 500$. In 2 of Algorithm 17, we can consider only $N(\mu = 500, 100)$. Since Θ_0 and Θ_1 form partition, alternative hypothesis usually is not simple. In this example, $\mu > 500$. We cannot specify one distribution for alternatrive. How to deal with this?

Trying many points for $\mu_1 \in \Theta_1 = \{\mu : \mu > 500\}$ might be possible. Our goal is finding larger power. So *find test with larger power for all points in* Θ_1.

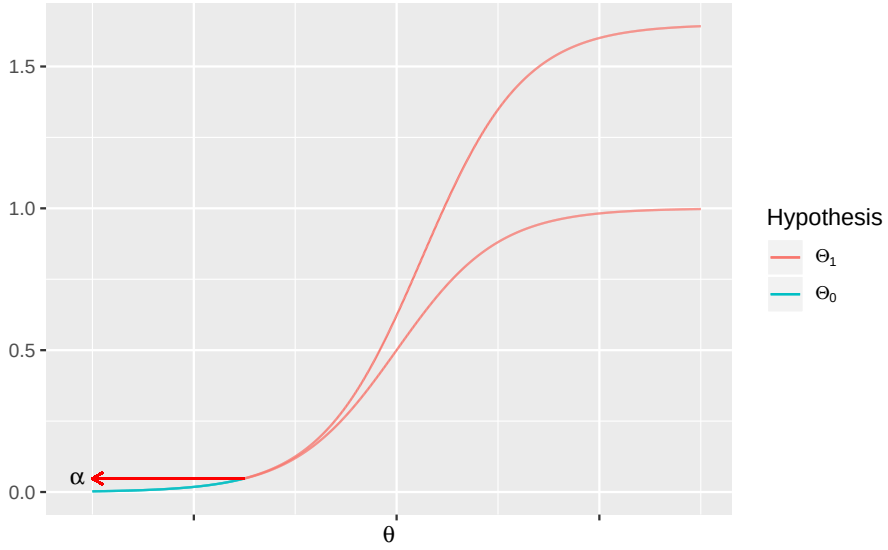


Figure 3.7: Comparing power between two test methods

See Figure 3.7. One test method has higher $\beta(\theta)$ function curve in Θ_1. This test is *powerful than the other*. We would choose this test in this step. So what we have to do is choose some points $\theta_1 \in \Theta_1$, and draw the power curve.

Algorithm 17: Empirical power by Monte Carlo method

input : $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$

foreach $\theta_1 \in \Theta_1$ **do**

  for $m \leftarrow 1$ **to** M **do**

    Generate $X_1^{(m)}, \ldots, X_n^{(m)} \overset{H_0}{\sim} f$;

    Compute $T_m(\mathbf{X}^{(m)})$;

    Compute $I_m = \begin{cases} 1 & H_0 \text{ is rejected} \mid H_1 \\ 0 & \text{otherwise} \end{cases}$;

  end

  Empirical power $\hat{\beta} = \frac{1}{M} \sum_{m=1}^{M} I_m$;

end

Draw a power curve $\hat{\beta}$ against θ_1 **output:** curve and $\{\hat{\beta}\}$

In fact, we can try every $\theta \in \Theta$ and *draw entire power curve*. Refer to Example 3.6. `foreach` library would be additionaly used.

```
library(foreach)
```

```
(pw_mc <-
  foreach(mu1 = seq(450, 650, by = 10), .combine = rbind) %do% {
    mc_data(rnorm, N = 20, mean = mu1, sd = 10)[,
                                                h1 := mu1]
  })
#>            x   sam  h1
#>      1: 448    s1 450
#>      2: 452    s1 450
#>      3: 453    s1 450
#>      4: 450    s1 450
#>      5: 449    s1 450
#>     ---
#> 419996: 646 s1000 650
#> 419997: 647 s1000 650
#> 419998: 660 s1000 650
#> 419999: 652 s1000 650
#> 420000: 652 s1000 650
```

One column is added from previous process. This is group for H_1. So we should specify `by = .(h1, sam)`.

```
pw_mc <-
  pw_mc[,
        lapply(.SD, test_list, mu = 500, sig = 10) %>%
          unlist() %>%
          as.list(),
        by = .(h1, sam)][,
                         lapply(.SD, mean),
                         by = h1, .SDcols = -"sam"]
```

```
pw_mc %>%
  melt(id.vars = "h1", variable.name = "test") %>%
```

```
  ggplot(aes(x = h1, y = value, colour = test)) +
  geom_path() +
  geom_point() +
  scale_colour_discrete(
    name = "Test",
    labels = c("Z", "T")
  ) +
  labs(
    x = expression(mu),
    y = expression(beta)
  )
```

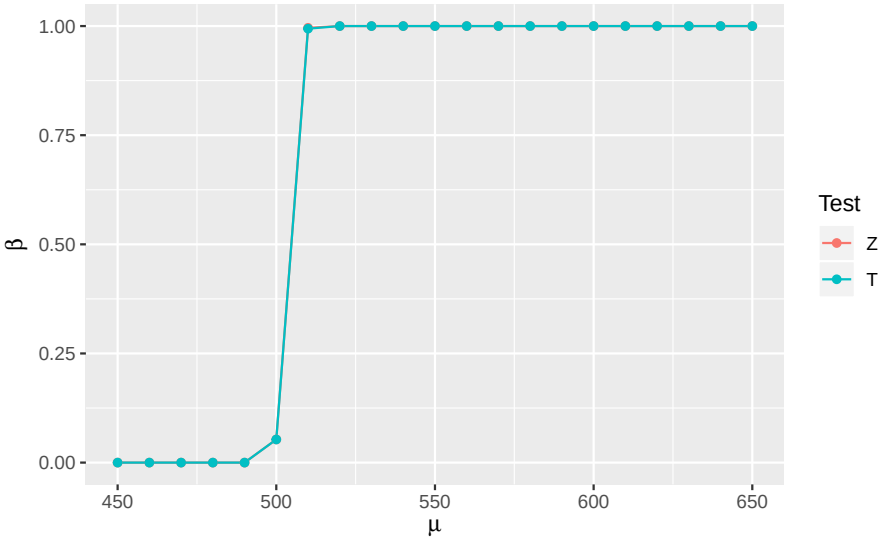


Figure 3.8: Empirical power curve of each z-test and t-test

Recall that we are estimating power. Instead of `mean()`, we can use `sd()`. This would give us *standard error* of our estimator for power. Since it is sample proportion,

$$\widehat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{M}}$$

Consider T-test.

```
pw_mc2 <-
  foreach(mu1 = seq(450, 650, by = 10), .combine = rbind) %do% {
    mc_data(rnorm, N = 20, mean = mu1, sd = 10)[,
                                                h1 := mu1]
  }
#------------------
pw_mc2 <-
  pw_mc2[,
         .(te = t.test(x, alternative = "greater", mu = 500)$p.value <= .05),
         by = .(h1, sam)][,
                          .(te = mean(te)),
                          by = h1][,
                                   se := sqrt(te * (1 - te) / 1000)]
```

```
pw_mc2 %>%
  ggplot(aes(x = h1, y = te)) +
  geom_ribbon(aes(ymin = te - se, ymax = te + se), col = gg_hcl(1)) +
  geom_path(alpha = .7) +
  geom_point() +
  labs(
    x = expression(mu),
    y = expression(beta)
  )
```

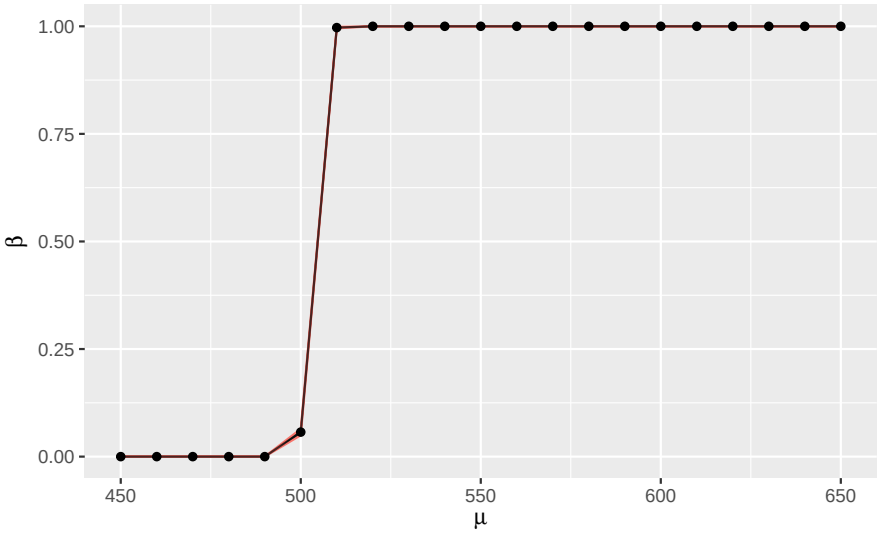


Figure 3.9: Empirical power curve $\hat{p} \pm \widehat{SE}(\hat{p})$ for t-test

3.4.5 Count Five test for equal variance

Commonly, F-test is used for equality of two population variances. McGrath and Yeh (2005) suggests nonparametric testing without Normal assumption, so called *Count Five*. Instead, this method requires some conditions.

1. same mean
2. same sample size

Algorithm 18: Count Five test

input : $X_1, \ldots, X_{n_x} \perp\!\!\!\perp Y_1, \ldots, Y_{n_y}$
$H_0 : \sigma_X^2 = \sigma_Y^2$

Compute $C_X = \left|\{i : |X_i - \overline{X}| > \max_j |Y_j - \overline{Y}|\}\right|$;
if $C_X \geq 5$ **then**
  return reject H_0;
else
  return accept H_0;
end

```
gauss <-
  tibble(
    x1 = rnorm(20, mean = 0, sd = 1),
    x2 = rnorm(20, mean = 0, sd = 1.5)
  )
```

```
gauss %>%
  gather(key = "variable", value = "value") %>%
  ggplot(aes(x = variable, y = value, fill = variable)) +
  geom_boxplot() +
  geom_point(alpha = .5)
```

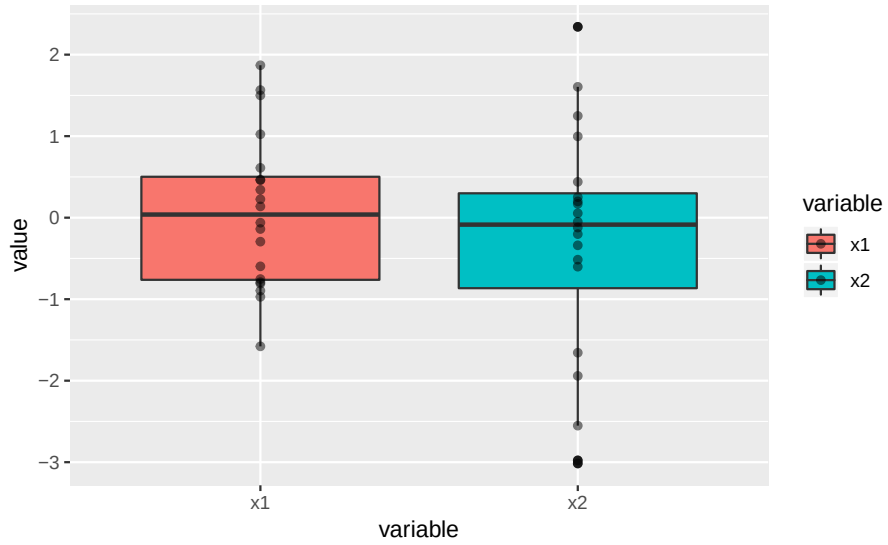


Figure 3.10: Side-by-side boxplot

We would perform *Count Five* test for multiple simulated data sets such as in Figure 3.10.

$$X_1^{(m)}, \ldots, X_{20}^{(m)} \sim N(0,1) \perp\!\!\!\perp Y_1^{(m)}, \ldots, Y_{20}^{(m)} \sim N(0, 1.5)$$

```
count5test <- function(x, y) {
  X <- x - mean(x)
  Y <- x - mean(y)
  outx <- sum(X > max(Y)) + sum(X < min(Y))
  outy <- sum(Y > max(X)) + sum(Y < min(X))
  max(c(outx, outy)) > 5
}
```

Apply MC method to get *empirical type I error*.

```
mc_data(rnorm, N = 20, M = 1000)[,
                                 x2 := rnorm(20 * 1000)][,
                                                         .(chat = count5test(x = x, y = x2)),
                                                         by = sam][,
                                                                   .(chat = mean(chat))]
#>      chat
#> 1: 0.026
```

3.5 Statistical Methods

3.6 Bootstrap

Bibliography

McGrath, R. N. and Yeh, A. B. (2005). A Quick, Compact, Two-Sample Dispersion Test. *The American Statistician*, 59(1):47–53.