Chapter 3: Methods for Simulating Data

Statisticians (and other users of data) need to simulate data for many reasons.

→ Goodness of fit test.

For example, I simulate as a way to check whether a model is appropriate. If the observed data are similar to the data I generated, then this is one way to show my model may be a good one.

It is also sometimes useful to simulate data from a distribution when I need to estimate an expected value (approximate an integral). Ch. 5

`R` can already generate data from many (named) distributions:

```
set.seed(400) #reproducibility
```

→ set starting point for pseudo random number generator ⇒ reproduce results.

```
rnorm(10) # 10 observations of a N(0,1) r.v.
```

```
## [1] -1.0365488  0.6152833  1.4729326 -0.6826873 -0.6018386 -1.3526097
## [7]  0.8607387  0.7203705  0.1078532 -0.5745512
```

sd.

```
rnorm(10, 0, 5) # 10 observations of a N(0,5^2) r.v.
```

```
## [1] -4.5092359  0.4464354 -7.9689786 -0.4342956 -5.8546081  2.7596877
## [7] -3.2762745 -2.1184014  2.8218477 -5.0927654
```

$\lambda=1$

```
rexp(10) # 10 observations from an Exp(1) r.v.
```

↑ ,?

```
## [1] 0.67720831 0.04377997 5.38745038 0.48773005 1.18690322 0.92734297
## [7] 0.33936255 0.99803323 0.27831305 0.94257810
```

But what about when we don't have a function to do it?

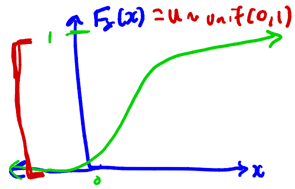
↳ we will need to write our own functions to simulate draws from other distributions.

1 Inverse Transform Method

easiest method when it works

Theorem 1.1 (Probability Integral Transform) If X is a continuous r.v. with cdf F_X, then $U = F_X(X) \sim \text{Uniform}[0, 1]$.

discrete version exists as well.



This leads to to the following method for simulating data.

Inverse Transform Method:

First, generate u from $\text{Uniform}[0, 1]$. Then, $x = F_X^{-1}(u)$ is a realization from F_X.

Note:

F_X^{-1} may not be available in closed form. If that's the case, use something else.

1.1 Algorithm

1. Derive the inverse function F_X^{-1}. To do this, let $F_X(x) = u$ Then solve for x to find $x = F_X^{-1}(u)$.

2. Write a function to compute $x = F_X^{-1}(u)$.
   in R

3. For each realization, simulated value.

   a. generate a random value u from Unif $(0,1)$

   b. Compute $x = F_X^{-1}(u)$
   simulated draw from $F_X(x)$.

Example 1.1 Simulate a random sample of size 1000 from the pdf $f_X(x) = 3x^2, 0 \le x \le 1$.

0. Find the cdf F_X

$$F_X(x) = \int_0^x 3y^2\,dy = y^3\Big|_0^x = \begin{cases} 0 & x<0 \\ x^3 & x \in [0,1]. \\ 1 & x>1. \end{cases}$$

1. Find F_X^{-1}

$u = F_X(x) = x^3 \Rightarrow u^{1/3} = x = F_X^{-1}(u)$.

so $F_X^{-1}(u) = u^{1/3}$ for $0 \le u \le 1$
↑ range of $F_X(x)$.

2.

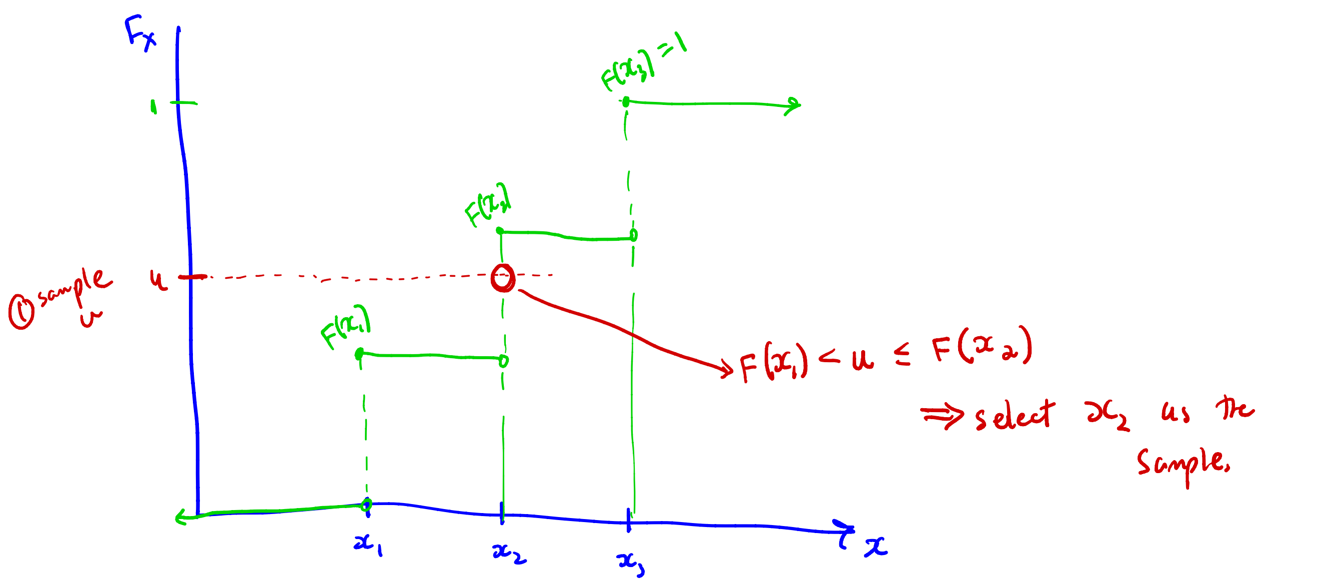
```
# write code for inverse transform example
# f_X(x) = 3x^2, 0 <= x \<= 1
```

a) Write function for F^{-1}
b) sample u values from unif(0,1)] 1000 times.
c) evaluate $x = F_X^{-1}(u)$.

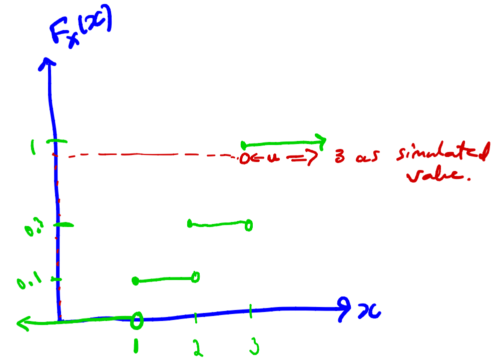
1.2 Discrete RVs

If X is a discrete random variable and $\cdots < x_{i-1} < x_i < \cdots$ are the points of discontinuity (jumps.) of $F_X(x)$, then the inverse transform is $F_X^{-1}(u) = x_i$ (↑ jump point) where $F_X(x_{i-1}) < u \le F_X(x_i)$. This leads to the following algorithm:

1. Generate a r.v. U from Unif$(0, 1)$.
2. Select x_i (↑ jump point) where $F_X(x_{i-1}) < U \le F_X(x_i)$.



Example 1.2 Generate 1000 samples from the following discrete distribution.




```r
x <- 1:3
p <- c(0.1, 0.2, 0.7)
```

pmf

x	1.0	2.0	3.0
f	0.1	0.2	0.7

```r
# write code to sample from discrete dsn
n <- 1000
```

There is a really simple way to do this in R.

using sample()

* remember to allow replacement and specify the prob. vector.

Something we can try when we can't find F^{-1} (in closed form)

2 Acceptance-Reject Method

the distribution we want to sample from.

The goal is to generate realizations from a *target density*, f.

Most cdfs cannot be inverted in closed form.

The Acceptance-Reject (or "Accept-Reject") samples from a distribution that is *similar* to f and then adjusts by only accepting a certain proportion of those samples.

target

and rejecting the rest.

The method is outlined below:

requirements for $g(\cdot)$.

Let g denote another density from which we (1) **know how to sample** and we can (2) **easily calculate** $g(x)$.

Let $e(\cdot)$ denote an *envelope*, having the property $e(x) = cg(x) \geq f(x)$ for all $x \in \mathcal{X} = \{x : f(x) > 0\}$ for a given constant $c \geq 1$.

target

The envelope covers all of f

support of target

(3) This implies that the support of $g(\cdot)$ MUST INCLUDE the support of f!

The Accept-Reject method then follows by sampling $Y \sim g$ and $U \sim \text{Unif}(0, 1)$.

If $U < f(Y)/e(Y)$, accept Y. Set $X = Y$ and consider X to be an element of the target random sample.

Note: $1/c$ is the expected proportion of candidates that are accepted.

We can use this to evaluate the efficiency of our algorithm.

What might be hard/slow?

(slow) - depending on efficiency, we might have to draw a LOT of sample just to keep a few.

(slow) - could be slow to evaluate f

(hard) - choose g & e

2.1 Algorithm

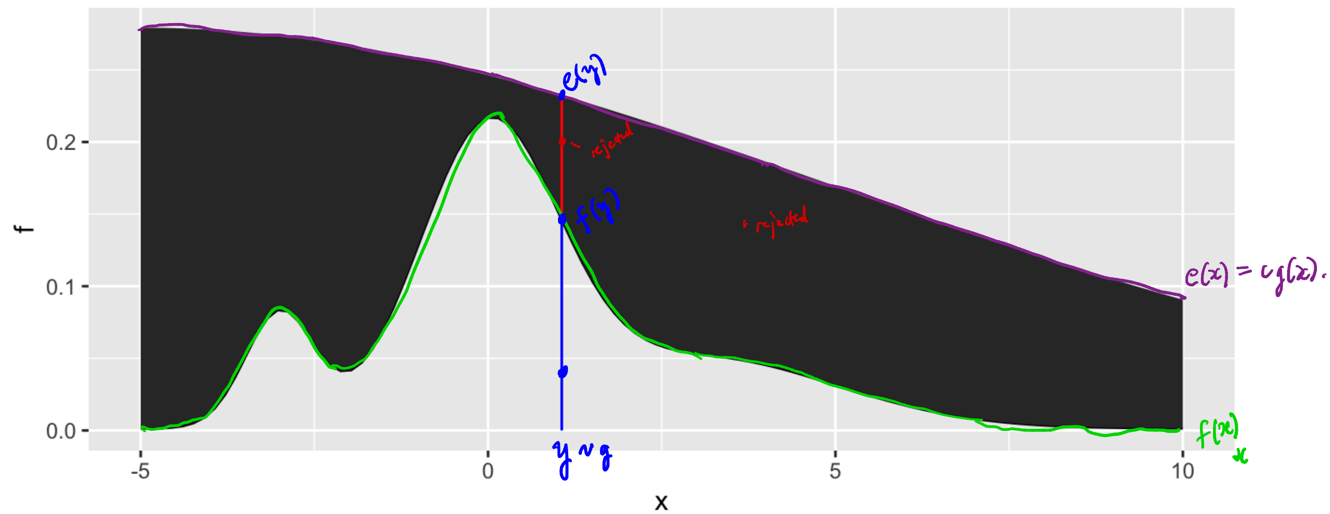
proposal

1. Find a suitable density g and envelope e. (hard step)
2. Sample $Y \sim g$.
3. Sample $U \sim \text{Unif}(0, 1)$.
4. If $U < f(Y)/e(Y)$, accept Y. (slow step*)
5. Repeat from Step 2 until you have generated your desired sample size. (slow step.)

* Requirement: the support of g MUST INCLUDE the support of f *

(BAD) example: If $f \equiv N(0,2)$ and $g \equiv \text{Unif}(-10,10)$.

This would not be appropriate because support of f is $(-\infty, \infty)$ and support of g is $[-10, 10]$.



2.2 Envelopes

Good envelopes have the following properties:

① envelope exceeds target everywhere ← support of g must include support of f
choose c to make this ① happen.

② Easy to sample from g.

③ Generate few rejected draws (save time).

This always works when $\mathcal{X} = [0,1]$.

A simple approach to finding the envelope: in some cases.

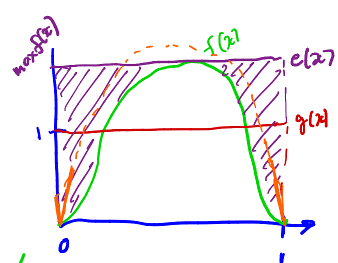
Say the support of f is $\mathcal{X} = \{x : 0 \le x \le 1\}$

Find $\max_x(f(x))$ and let $c = \max_x(f(x))$

Let $g(x) \equiv Unif(0,1) = \begin{cases} 1 & \text{if } x \in [0,1] \\ 0 & \text{otherwise} \end{cases}$

support of g matches the support of f ✓

$e(x) = cg(x)$



This is often not efficient if you know more about the shape of f you can select a better envelope.

Plotting is your friend here!

Example 2.1 We want to generate a random variable with pdf $f(x) = 60x^3(1-x)^2$, $0 \le x \le 1$. This is a Beta$(4,3)$ distribution.

↳ could just use rbeta() in R.

Can we invert $F(x)$ analytically?

No. Let's use accept-reject!

Let $g \sim \text{Unif}(0,1)$.

If not, find the maximum of $f(x)$. $= c$

$$f'(x) = 60\left(3x^2(1-x)^2 - 2x^3(1-x)\right)$$
$$= 60x^2(1-x)\left[3(1-x) - 2x\right]$$
$$= 60x^2(1-x)(3-5x) = 0 \text{ when } x=0,\ x=1, \text{ or } \boxed{x = \tfrac{3}{5}}$$

$f(0) = f(1) = 0.$

$$c = \max_x f(x) = f\left(\tfrac{3}{5}\right) = 2.0736$$

density for beta

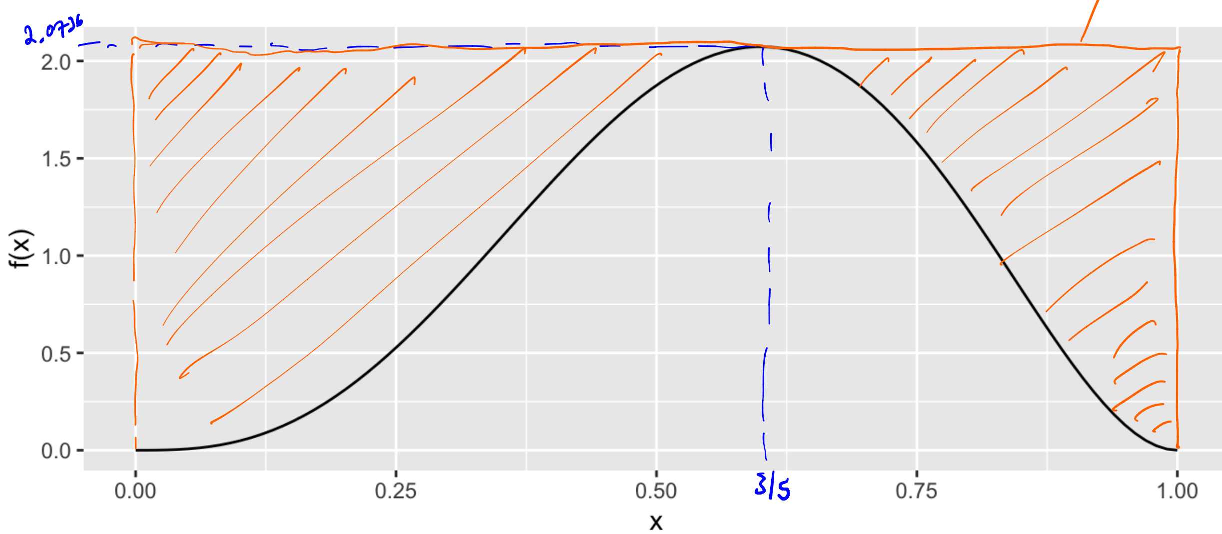
```
# pdf function, could use dbeta() instead
f <- function(x) {
  60*x^3*(1-x)^2
}

# plot pdf
x <- seq(0, 1, length.out = 100)
ggplot() +
  geom_line(aes(x, f(x)))
```

make x values

f evaluated at x values

line.

$$e(x) = \begin{cases} 2.0736 & x \in [0,1] \\ 0 & \text{o.w.} \end{cases}$$



```r
envelope <- function(x) {
  ## create the envelope function
}
```

c · unif pdf
= c · 1
= $f(3/5)$

```r
# Accept reject algorithm
n <- 1000 # number of samples wanted
accepted <- 0 # number of accepted samples
samples <- rep(NA, n) # store the samples here
```

we don't know how many iterations it will take => for loop not helpful.
run loop until we have accepted enough (n).

```r
while(accepted < n) {
  # sample y from g
```

unif(0,1)

y <- runif(1).

```r
  # sample u from uniform(0,1)
  u <- runif(1)

  if(u < f(y)/envelope(y)) {
    # accept
    accepted <- accepted + 1
    samples[accepted] <- y
  }
}
```

increment accepted so loop ends eventually.

store samples.

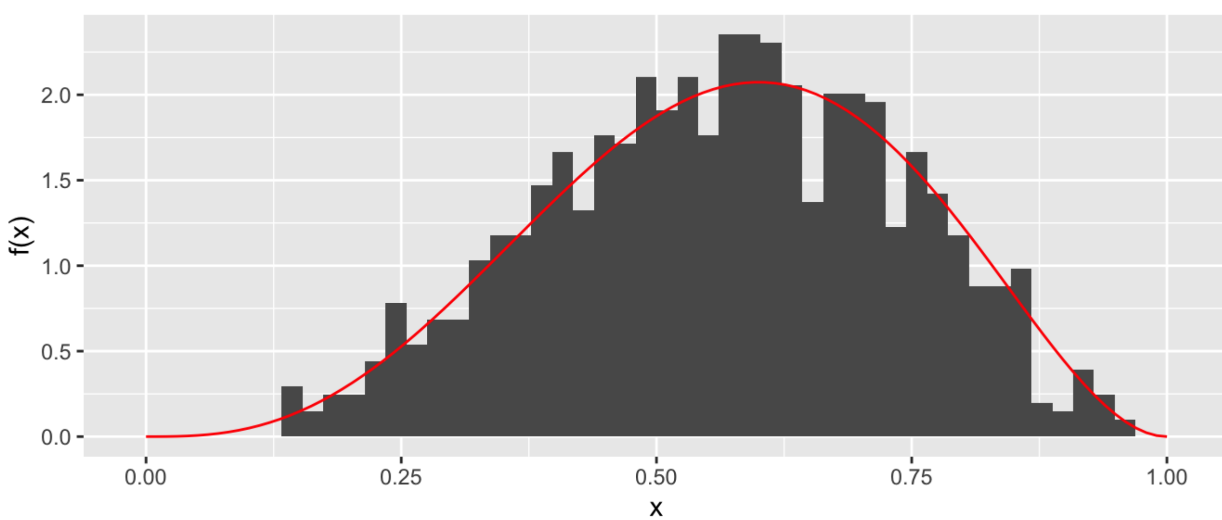
necessary so that histogram is on the same scale as the density function instead of raw counts.

samples from f

* important for next homework.

```r
ggplot() +
  geom_histogram(aes(samples, y = ..density..), bins = 50, ) +
  geom_line(aes(x, f(x)), colour = "red") +
  xlab("x") + ylab("f(x)")
```

theoretical pdf



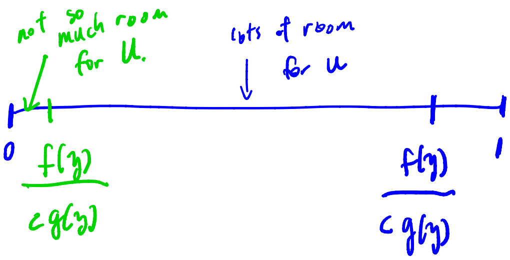
2.3 Why does this work?

Recall that we require

$$e(y) = cg(y) \geq f(y) \qquad \forall y \in \{y : f(y) > 0\}.$$

Thus,

$$0 \leq \frac{f(y)}{e(y)} = \frac{f(y)}{cg(y)} \leq 1$$



The larger the ratio $\frac{f(y)}{cg(y)}$, the more the random variable Y looks like a random variable distributed with pdf f and the more likely Y is to be accepted.

2.4 Additional Resources

should be a copy w/ library.

See p.g. 69-70 of Rizzo for a proof of the validity of the method.

How to choose envelope (if support is not $[0,1]$):

consulting a table dist's (wikipedia)

① start w/ support of f. $\Rightarrow$ list of potential g's either w/ same support or larger.

making plots.

② plot f to get a sense of the shape. Try to pick a g from my list w/ similar shape.

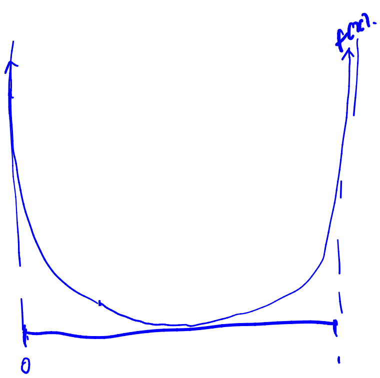
③ pick c s.t. $cg(x) \geq f(x)$ $\forall x$.

↳ picking a bunch of c's, plotting $c \cdot g(x)$ vs. $f(x)$

evaluating $c \cdot g(x)$ vs. $f(x)$ at a wide range of x's.

Choose the smallest c I can that makes $c \cdot g(x) \geq f(x)$ $\forall x$.

SORRY I lied to you



we can't evaluate max $f(x)$

3 Transformation Methods

We have already used one transformation method – **Inverse transform method** – but there are many other transformations we can apply to random variables.

1. If $Z \sim N(0,1)$, then $V = Z^2 \sim \chi^2_1$
2. If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ are independent, then $F = \frac{U/m}{V/n} \sim F_{m,n}$
3. If $Z \sim N(0,1)$ and $V \sim \chi^2_n$ are independendent, then $T = \frac{Z}{\sqrt{V/n}} \sim t_n$
4. If $U \sim \text{Gamma}(r, \lambda)$ and $V \sim \text{Gamma}(s, \lambda)$ are independent, then $X = \frac{U}{U+V} \sim \text{Beta}(r, s)$.
5. If $X \sim F$, then $F^{-1}(x) \sim \text{Unif}(0,1)$. (PIT, leads to inverse transform method).

$X \to g(X)$.

Definition 3.1 A *transformation* is any function of one or more random variables.

Sometimes we want to transform random variables if observed data don't fit a model that might otherwise be appropriate. Sometimes we want to perform inference about a new statistic.

Example 3.1 If $X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$. What is the distribution of $\sum_{i=1}^n X_i$?

Can derive $\sum X_i \sim \text{Binomial}(n,p)$.

Example 3.2 If $X \sim N(0,1)$, what is the distribution of $X + 5$?

$X+5 \sim N(5,1)$.

Example 3.3 For $X_1, \ldots, X_n$ iid random variables, what is the distribution of the median of $X_1, \ldots, X_n$? What is the distribution of the order statistics? $X_{[i]}$?

This is more complex...

There are many approaches to deriving the pdf of a transformed variable. could then use pdf in accept-reject... (maybe).

1. Change of variable

If (1) g is monotone, then for (2) cts X and $Y = g(X)$

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|, & y \in \mathcal{Y} \\ 0 & \text{o.w.} \end{cases}$$

2. Moment generating functions

$M_X(t) = E(e^{tX})$

$M_{g(X)}(t) = E(e^{t g(X)})$

3. Convolution Theorem

$Z = X + Y$

etc.

But the theory isn't always available. What can we do?

Use other computational tools to simulate from the transformed distribution:

3.1 Algorithm

Let $X_1, \dots, X_p$ be a set of independent random variables with pdfs $f_{X_1}, \dots, f_{X_p}$, respectively, and let $g(X_1, \dots, X_p)$ be some transformation we are interested in simulating from.

1. Simulate $X_1 \sim f_{X_1}, \dots, X_p \sim \boxed{f_{X_p}}$.
   either straightforward (named-distribution) inverse transform method, accept-reject, etc.
2. Compute $G = g(X_1, \dots, X_p)$. This is one draw from $g(X_1, \dots, X_p)$.
   target distribution.
3. Repeat Steps 1-2 many times to simulate from the target distribution.

Example 3.4 It is possible to show for $X_1, \dots, X_p \overset{iid}{\sim} N(0,1)$, $Z = \sum_{i=1}^p X_i^2 \sim \chi^2_p$. Imagine that we cannot use the `rchisq` function. How would you simulate Z?

degrees of freedom controls shape of dsn.

1. Simulate p variables from N(0,1).
2. Compute $\sum X_i^2$
3. Repeat 1-2 many times.

```
library(tidyverse)

# function for squared r.v.s
squares <- function(x) x^2
```

\# draws; # of r.v.'s (d.f. of χ^2)

```
sample_z <- function(n, p) {
  # store the samples
  samples <- data.frame(matrix(rnorm(n*p), nrow = n))
```

iid n draws from p iid N(0,1).

accept-reject if necessary...

```
  samples %>%
    mutate_all("squares") %>% # square the rvs
    rowSums() # sum over rows
}
```

there are many ways to do this

```
# get samples
n <- 1000 # number of samples

# apply our function over different degrees of freedom
samples <- data.frame(chisq_2 = sample_z(n, 2),
                      chisq_5 = sample_z(n, 5),
                      chisq_10 = sample_z(n, 10),
```

p=2

p=100

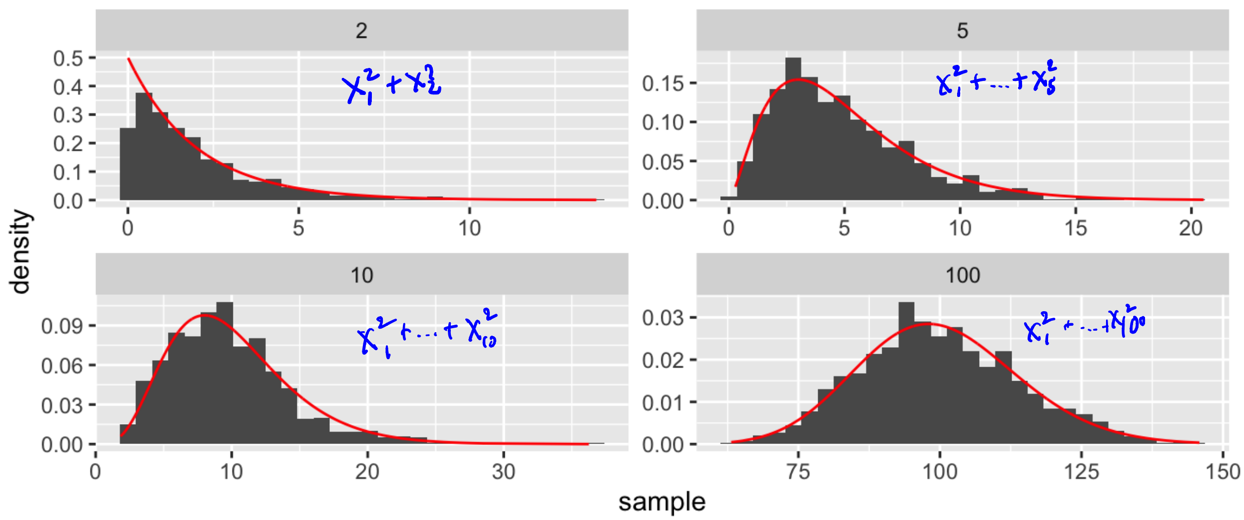
```
chisq_100 = sample_z(n, 100))

# plot results
samples %>%
  gather(distribution, sample, everything()) %>% # make easier to
  plot w/ facets
  separate(distribution, into = c("dsn_name", "df")) %>% # get the df
  mutate(df = as.numeric(df)) %>% # make numeric
  mutate(pdf = dchisq(sample, df)) %>% # add density function values
  ggplot() + # plot
  geom_histogram(aes(sample, y = ..density..)) + # samples
  geom_line(aes(sample, pdf), colour = "red") + # true pdf
  facet_wrap(~df, scales = "free")
```

true

← remember to plot on the same scale!

in red

different scales for different dfs.



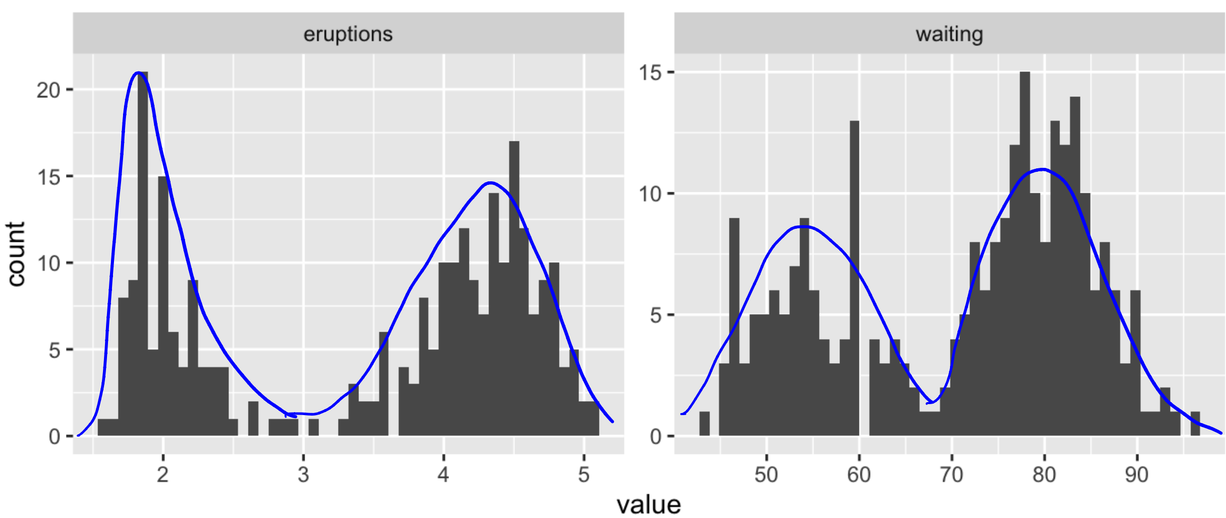
4 Mixture Distributions

The `faithful` dataset in `R` contains data on eruptions of Old Faithful (Geyser in Yellowstone National Park).

```
head(faithful)
```

```
##   eruptions waiting
## 1     3.600      79
## 2     1.800      54
## 3     3.333      74
## 4     2.283      62
## 5     4.533      85
## 6     2.883      55
```

```
faithful %>%
  gather(variable, value) %>%
  ggplot() +
  geom_histogram(aes(value), bins = 50) +
  facet_wrap(~variable, scales = "free")
```



What is the shape of these distributions?

Bimodal, i.e. two modes.

distribution is weighted sum of distributions

Definition 4.1 A random variable Y is a discrete mixture if the distribution of Y is a weighted sum $F_Y(y) = \sum \theta_i F_{X_i}(y)$ for some sequence of random variables $X_1, X_2, \ldots$ and $\theta_i > 0$ such that $\sum \theta_i = 1$.

For 2 r.v.s,

$\theta + (1-\theta) = 1$ ✓

$$f_Y(y) = \theta \boxed{f_{X_1}(y)} + (1-\theta)\boxed{f_{X_2}(y)}$$

two different distributions

How do we simulate from this distribution?

There are 2 sources of variability.

(1) → which distribution to draw from (f_{X_1} or f_{X_2}):

$$Z \sim \text{Bernoulli}(\theta) \rightarrow \begin{cases} Z = 1 & X \sim f_{X_1} \\ Z = 0 & X \sim f_{X_2} \end{cases}$$

Algorithm:

(1) draw $Z \sim \text{Bernoulli}(\theta)$

(2) if $Z = 1$, draw $X \sim f_{X_1}$

if $Z = 0$, draw $X \sim f_{X_2}$

repeat many times.

Example 4.1

```
x <- seq(-5, 25, length.out = 100)

mixture <- function(x, means, sd) {
  # x is the vector of points to evaluate the function at
  # means is a vector, sd is a single number
  f <- rep(0, length(x))
  for(mean in means) {
    f <- f + dnorm(x, mean, sd)/length(means) # why do I divide?
  }
  f
}

# look at mixtures of N(mu, 4) for different values of mu
data.frame(x,
           f1 = mixture(x, c(5, 10, 15), 2),
           f2 = mixture(x, c(5, 6, 7), 2),
           f3 = mixture(x, c(5, 10, 20), 2),
           f4 = mixture(x, c(1, 10, 20), 2)) %>%
  gather(mixture, value, -x) %>%
  ggplot() +
  geom_line(aes(x, value)) +
  facet_wrap(.~mixture, scales = "free_y")
```

vector of 3 means

common sd.

density of $N(\mu_i, \sigma^2)$

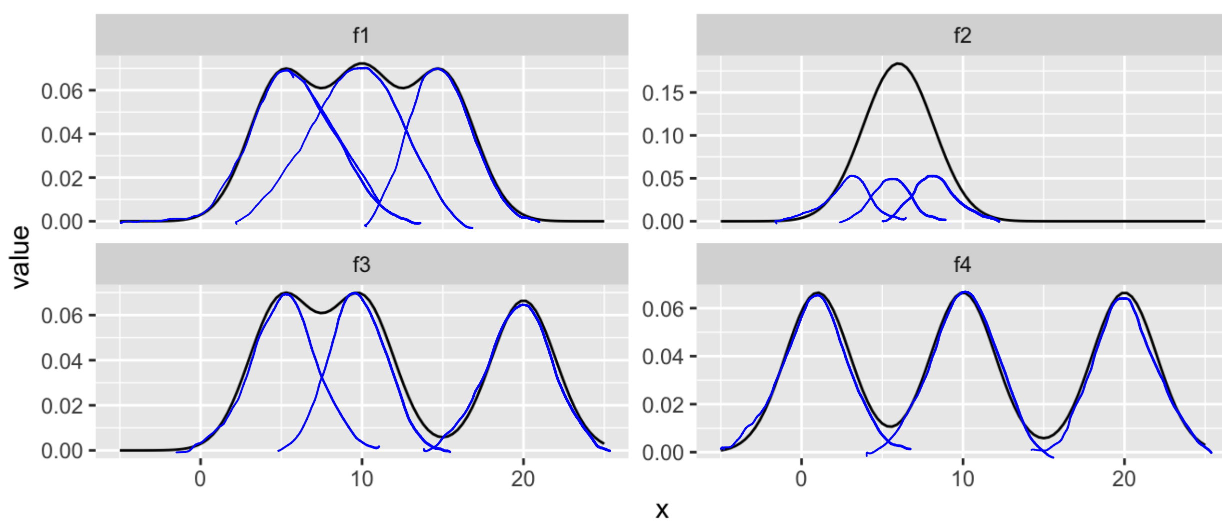
3?

mu

equally weighting each distribution.

$\theta_1 = \theta_2 = \theta_3 = \frac{1}{3}$.

(we don't have to equally weight need $\sum \theta_i = 1$).



4.1 Mixtures vs. Sums

Note that mixture distributions are *not* the same as the distribution of a sum of r.v.s.

mixtures are weighted sums of distributions

NOT distributions of weighted sums of random variables!

Example 4.2 Let $X_1 \sim N(0,1)$ and $X_2 \sim N(4,1)$, independent.

weighted sum of r.v.s called a "convolution"

$S = \frac{1}{2}(X_1 + X_2)$

$E(S) = E\left(\frac{1}{2}(X_1 + X_2)\right)$
$= \frac{1}{2}EX_1 + \frac{1}{2}EX_2 = \frac{1}{2}(0+4) = 2.$

$Var(S) = Var\left(\frac{1}{2}(X_1+X_2)\right) \overset{indep}{=} \frac{1}{4}(Var X_1 + Var X_2) = \frac{1}{4}(1+1) = \frac{1}{2}$

Can show in fact $S = \frac{1}{2}(X_1+X_2) \sim N\left(2, \frac{1}{2}\right)$ unimodal, bell shaped.

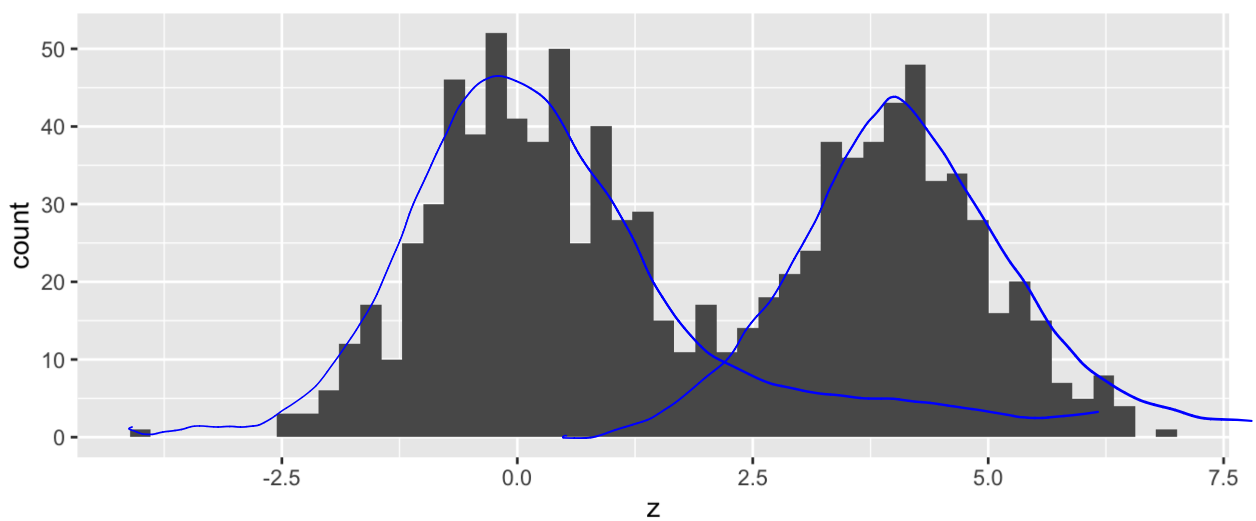
mixture density.

Z such that $f_Z(z) = \underset{\theta}{0.5} f_{X_1}(z) + 0.5 f_{X_2}(z)$.

```
① n <- 1000   # draws
② u <- rbinom(n, 1, 0.5)   θ   choose which dsn
                  u=1     match up with the mixture I want.
③ z <- u*rnorm(n) + (1 - u)*rnorm(n, 4, 1)
       N(0,1)              N(4,1).
④ ggplot() +
    geom_histogram(aes(z), bins = 50)
```

This is NOT a $N(2, \frac{1}{2})$.

draws from density of Z



What about $f_Z(z) = 0.7 f_{X_1}(z) + 0.3 f_{X_2}(z)$?

change line ② in above code.

u <- rbinom(n, 1, 0.7)

4.2 Models for Count Data (refresher)

Recall that the Poisson(λ) distribution is useful for modeling count data.

$$f(x) = \frac{\lambda^x \exp\{-\lambda\}}{x!}, \quad x = 0, 1, 2, \dots$$

Where X = number of events occuring in a fixed period of time or space. $X \sim$ Poisson(λ).

When the mean λ is low, then the data consists of mostly low values (i.e. $0, 1, 2$, etc.) and less frequently higher values.

As the mean count increases, the skewness goes away and the distribution becomes approximately normal.

With the Poisson distribution,

$$E[X] = VarX = \lambda.$$

restricts the shape of the dsn!

Example 4.3

- # homes sold per day by a real estate company
- # of calls coming per minute into a hotel reservation call center
- # of meows in a 2 minute cat video on youtube.

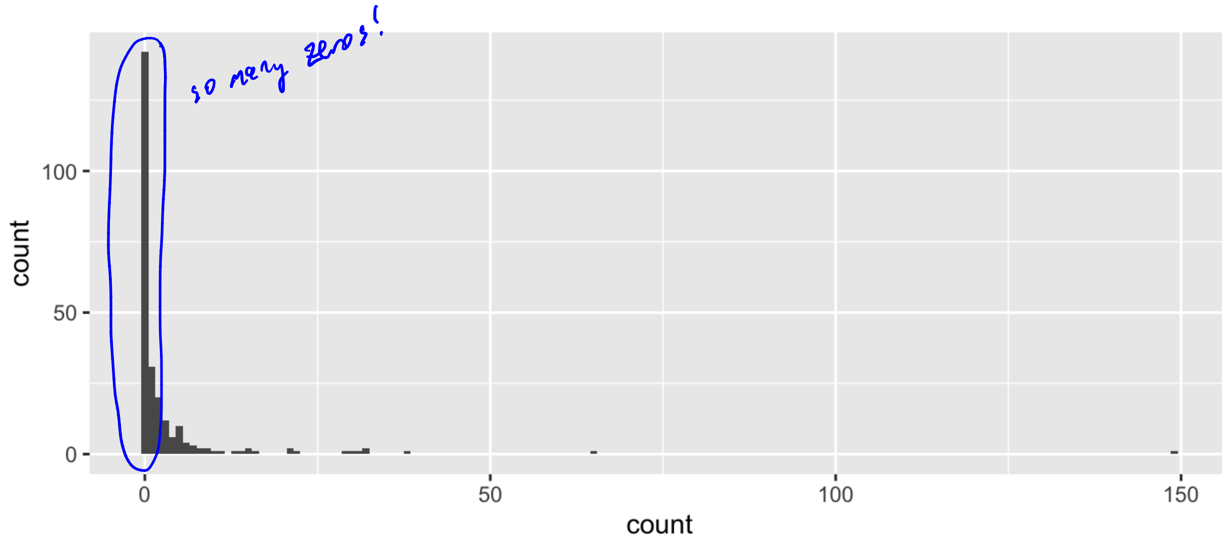
Example 4.4 The Colorado division of Parks and Wildlife has hired you to analyze their data on the number of fish caught in Horsetooth resevoir by visitors. Each visitor was asked - How long did you stay? - How many fish did you catch? - Other questions: How many people in your group, were children in your group, etc.

Some visiters do not fish, but there is not data on if a visitor fished or not. Some visitors who did fish did not catch any fish.

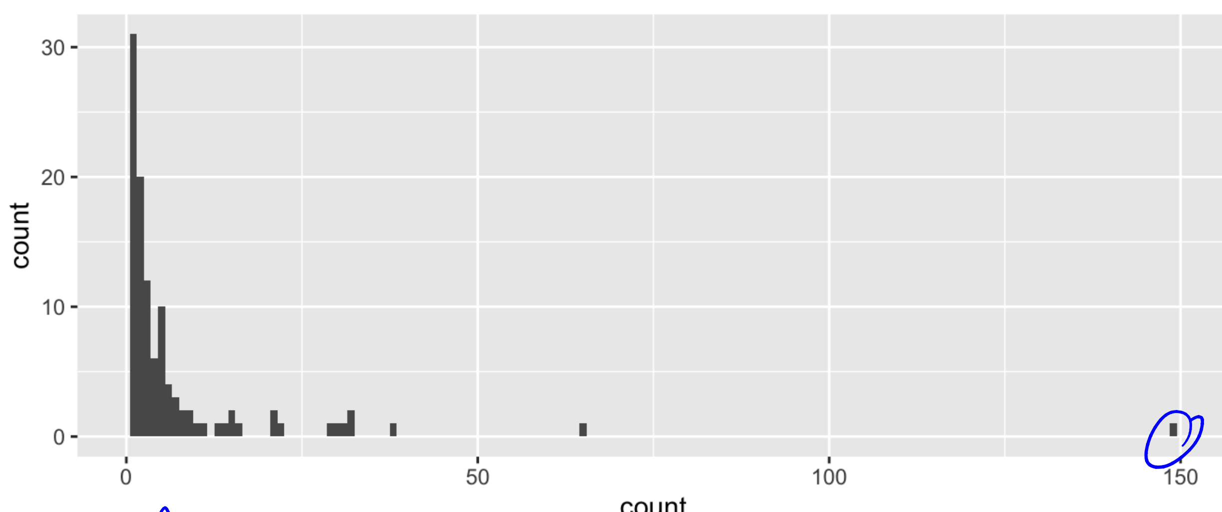
Note, this is modified from https://stats.idre.ucla.edu/r/dae/zip/.

```
fish <- read_csv("https://stats.idre.ucla.edu/stat/data/fish.csv")
```

```r
# with zeroes
ggplot(fish) + geom_histogram(aes(count), binwidth = 1)
```




```r
# without zeroes
fish %>%
  filter(count > 0) %>%
  ggplot() +
  geom_histogram(aes(count), binwidth = 1)
```



this may look more like Poisson dsn (some outliers).

A *zero-inflated* model assumes that the zero observations have two different origins – structural and sampling zeroes.

↳ when non-zero is impossible ↳ a zero is possible and occurs by random chance.

Example 4.5

Outcome of a study = # cows with foot and mouth disease (FMD) per regions in Turkey.

↳ structural zero – there are no cows in the region

↳ sampling zero – cows in region, but not FMD.

key point: you don't know whether region has no cows or no disease.

A zero-inflated model is a **mixture model** because the distribution is a weighted average of the sampling model (i.e. Poisson) and a point-mass at 0.

distribution for structural zeros.

For $Y \sim ZIP(\lambda)$,

$$Y \sim \begin{cases} 0 & \text{with probability } \pi \\ \text{Poisson}(\lambda) & \text{with probability } 1-\pi \end{cases}$$

So that,

$$Y = \begin{cases} 0 & \text{w.p. } \pi + (1-\pi)e(-\lambda) \\ k & \text{w.p. } (1-\pi)\dfrac{\lambda^k e(-\lambda)}{k!} \quad k = 1, 2, \ldots \end{cases}$$

To simulate from this distribution,

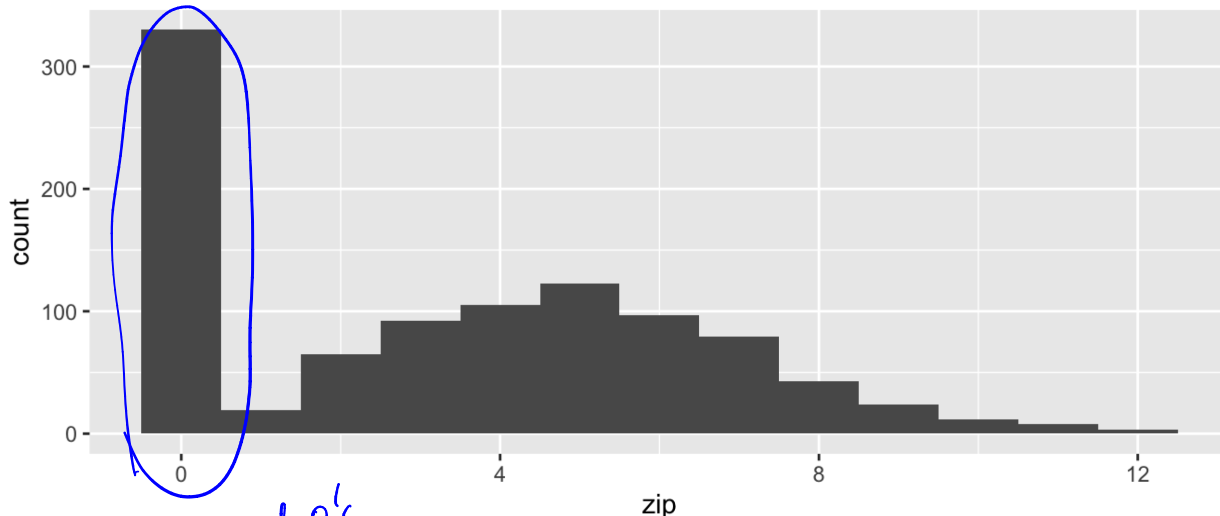
$Z \sim \text{Bern}(\pi)$

if $Z = 0$ $\quad Y \sim \text{Poisson}(\lambda)$

if $Z = 1$ $\quad Y = 0$.

```
n <- 1000
lambda <- 5
pi <- 0.3

u <- rbinom(n, 1, pi)
zip <- u*0 + (1-u)*rpois(n, lambda)
```

```r
# zero inflated model
ggplot() + geom_histogram(aes(zip), binwidth = 1)
```



structural 0's.

```r
# Poisson(5)
ggplot() + geom_histogram(aes(rpois(n, lambda)), binwidth = 1)
```