Lab 1. Introduction to R and Simulation

Objectives

- Introduction to R
- Simulate random events in R
- Show results in plots & simple functions in R

Introduction to R

Go to `http://www.r-project.org/` or `http://cran.r-project.org/` and install R. Once you have successfully set it up, try the following elementary commands.

```
5*9
(5/9)*(94-32)                    # To convert 95 Fahrenheit to Celsius.
x <- c(2.5,4.8,7.9,8.4,9.6,7.3,6.8,7.2,3.8,7.3,6.0,4.5,4.8,4.5,2.9,5.3,3.7,5.2,4.8,3.3,8.6,8.9)
length(x)
sum(x)
sort(x)
summary(x)
mean(x)
median(x)
sd(x)
var(x)
hist(x)
hist(x,prob=T)                   # Displays density, NOT frequency.
lines(density(x),col="red")
boxplot(x)
boxplot(x, horizontal=TRUE)
stem(x)
stem(x,scale=2)
plot(density(x))
```

Ex 1: Age at death of US presidents

Name	Age	Name	Age	Name	Age
Washington	67	Pierce	64	Wilson	67
Adams	90	Buchanan	77	Harding	57
Jefferson	83	Lincoln	56	Coolidge	60
Madison	85	Johnson	66	Hoover	90
Monroe	73	Grant	63	Roosevelt	63
Adams	80	Hayes	70	Truman	88
Jackson	78	Garfield	49	Eisenhower	78
Van Buren	79	Arthur	56	Kennedy	46
Harrison	68	Cleveland	71	Johnson	64
Tyler	71	Harrison	67	Nixon	82
Polk	53	McKinley	58	Ford	93
Taylor	65	Roosevelt	60	Reagan	93
Fillmore	74	Taft	72		

```
age <- c(67,90,83,85,73,80,78,79,68,71,53,65,74,64,77,56,66,63,70,49,56,
   71,67,58,60,72,67,57,60,90,63,88,78,46,64,82)
```

```
summary(age)
stem(age)
stem(age,scale=0.5)
boxplot(age)
par(mfrow=c(1,2))
hist(age, breaks=c(39.5,49.5,59.5,69.5,79.5,89.5,99.5))
plot(density(age))
```

Ex 2: Sample frequency table

Diameter	frequency
12.0	2
12.2	4
12.3	6
12.8	3
13.0	5

The sample mean and sample sd (standard deviation) are:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \qquad \text{and} \qquad s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

or equivalently for a frequency table :

$$\bar{x} = \sum_{i=1}^{k} \frac{f_i \cdot x_i}{n} \qquad \text{and} \qquad s = \sqrt{\sum_{i=1}^{k} \frac{f_i (x_i - \bar{x})^2}{n-1}}$$

```
diameter <- c(rep(12,2),rep(12.2,4),rep(12.3,6),rep(12.8,3),rep(13,5))
summary(diameter)
mean(diameter)
sd(diameter)
var(diameter)
stem(diameter)
dev.new()
boxplot(diameter)
```

Ex 3: Simulating flips of a fair coin and rolls of a fair die

```
sample(1:6, 10, replace=T)
sample(1:6, 10, replace=T)
hist(sample(1:6, 10, replace=T))
table(sample(1:6, 10, replace=T))
```

More proper way of using R is always storing what you have done into an object and use it. For example, roll a fair die 20 times:

```
x <- 1:6
y <- sample(x,20,replace=TRUE)
y
table(y)
hist(y)
```

Another example of flipping a fair coin 100 times.

```
x <- 0:1
(y <- sample(x,100,rep=T))  # Wrapping the whole line with () saves and displays it.
hist(y)
table(y)
```

You can also construct a short `function`:

```
Roll1Die <- function(n) sample(1:6,n,rep=T)
Roll1Die(100)
Flip1Coin <- function(n) sample(0:1,n,rep=T)
Flip1Coin(100)
```

Now, let's simulate the "gender" of 100 new born babies check the *relative frequency* of "girls."

```
x <- 0:1
babies <- sample(x,100,rep=T)
total_girls <- cumsum(babies)
total_babies <- seq(1,100,1)
rel_freq <- total_girls/total_babies
cbind(babies,total_girls,rel_freq)
plot(rel_freq,ylim=c(0,1.0),type="l",lty=1,col="red")
abline(h=0.5, lty=2,col="blue")
```

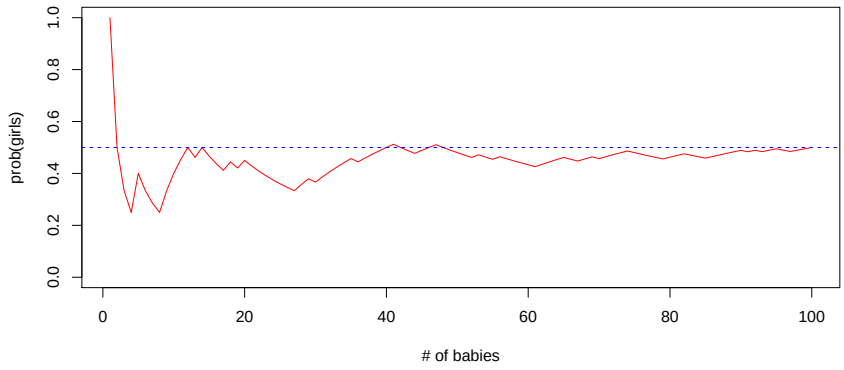

Figure 1: Probability of girls

Ex 4: Simulating more complicated probability

A fair die is rolled six times. If the face numbered k is the outcome on roll k for $k = 1, 2, \ldots, 6$, we say that a "match" has occurred. The experiment is called a **success** if at least one match occurs during the six trials. Otherwise, the experiment is called a failure.

$$P(\text{success}) = 1 - P(\text{all six trials did NOT match}) = 1 - \left(\frac{5}{6}\right)^6 = 0.665102$$

Here is how it's simulated in R (See textbook p. 5).

```
success <- numeric(1000)
x <- matrix(0,1000,6)
for (i in 1:1000) {
```

```
 x[i,] <- sample(1:6, 6, rep=T)
 for (j in 1:6) {
 if(x[i,j]==j) {success[i]=1}
 }
 }
sum(success)
[1] 695
```

Here is more advanced way of doing it.

```
n = 500
p <- numeric(n)
for (k in 1:n) {
  success <- numeric(k)
  x <- matrix(0,k,6)
  for (i in 1:k) {
   x[i,] <- sample(1:6, 6, rep=T)

   for (j in 1:6) {
   if(x[i,j]==j) {success[i]=1}
   }
   }
  p[k] <- (sum(success))/k
  }
plot(p,type="l",ylim=c(0,1),cex=0.7,col="red",axes=F,xlab="# of rolls")
axis(1)
axis(2,las=2)
abline(h=0.665,col="blue")
```

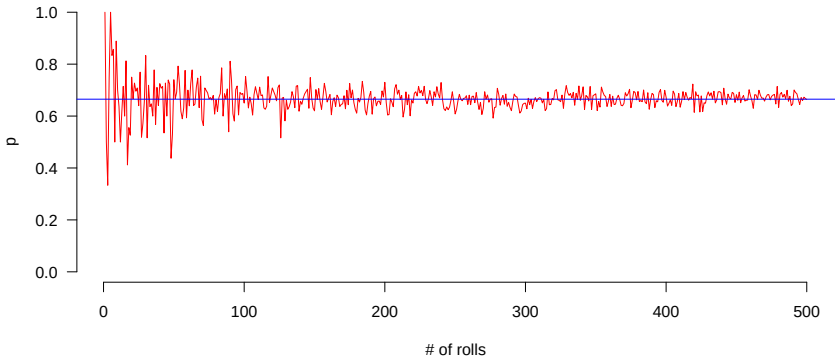

Figure 2: Probability of at least one match

Ex 5: Birthday problem – How many people do you need to have a good chance of seeing *at least* two same birthdays?

What'd be the probability of seeing *at least* two same birthdays in a group of 20 people? It can be shown that the theoretical probability of this is 0.4114, which seems rather high for a small group of *only* 20 people. Here is how it's done.

$$
\begin{aligned}
P(\textit{at least} \text{ two with the same birthdays}) &= 1 - P(\text{everyone with } \textit{different} \text{ birthdays}) \\
&= 1 - \left\{ \left(\frac{365}{365}\right) \cdot \left(\frac{364}{365}\right) \cdots \left(\frac{346}{365}\right) \right\} \\
&= 1 - 0.5886 \\
&= 0.4114
\end{aligned}
$$

Let's simulate this with R. We randomly select 20 days from 365 days (of a year) with replacement.

```
days <- 1:365
bday <- sample(days,20,rep=T)
sort(bday)
table(bday)
```

If you have figured out what's going on, here is how to find the probability of *at least* two with the same birthdays in a group consisting of anywhere from 2 to 100 people.

```
n = 100
days <- 1:365
prob <- numeric(n)
for (k in 2:n) {
bday <-matrix(0,1000,k)
count <- numeric(1000)
  for(i in 1:1000){
  bday[i,] <-sort(sample(days,k,rep =T))
  check <- (k - length(unique(bday[i,])))
  count[i] <- check
  }
prob[k] <- 1-(length(count[count==0]))/1000
}
plot(prob,type="o",pch=16,col="blue",xlab="# of people")
abline(h=0.5,lty=2,col="red")
```

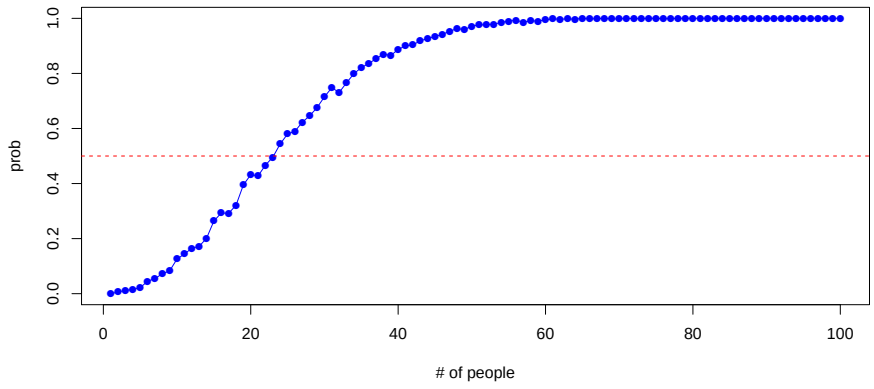

Figure 3: Probability of *at least* two with the same birthdays

Ex 6: Probability distribution table, mean (μ), sd (σ) and simulation

From a given probability distribution like:

x	$p(x)$
x_1	$p(x_1)$
x_2	$p(x_2)$
x_3	$p(x_3)$
$\cdots$	$\cdots$
x_n	$p(x_n)$

The (theoretical) population mean and (theoretical) population sd (standard deviation) are:

$$\mu = \sum_{i=1}^{n} x_i \cdot p(x_i) \qquad \text{and} \qquad \sigma = \sqrt{\sum_{i=1}^{n} (x_i - \mu)^2 \cdot p(x_i)}$$

For example, consider the following distribution of the number of raisins in a certain company's cookies.

# of raisins	prob
0	0.05
1	0.1
2	0.2
3	0.4
4	0.15
5	0.1

Here is how to find the (theoretical) population parameters (i.e., μ and σ) and sample statistics (i.e., $\bar{x}$ and s).

```
x <- 0:5
prob1 <- c(0.05,0.1,0.2,0.4,0.15,0.1)
(mu = sum(x*prob1))
(sigma = sqrt(sum(((x-mu)^2)*prob1)))
sample1 <- sample(x,1000,prob=prob1,rep=T)
mean(sample1)
sd(sample1)
table(sample1)
```

Now let's repeat this 5,000 times and look at teh distribution of the sample means.

```
means <- numeric(5000)
for(i in 1:5000){
   sample1 <- sample(x,1000,prob=prob1,rep=T)
   means[i] <- mean(sample1) }
plot(density(means),main="Distribution of 1,000 means",xlab="")
mean(means); sd(means)
[1] 2.800227
[1] 0.03950997
#Recall mu and sigma
> mu
[1] 2.8
> sigma/(sqrt(1000))
[1] 0.03949684
```

There is a well-known theory in statistics, called the **Central Limit Theorem** and it says that the sample means (i.e., $\bar{x}$'s) will have (i) a normal distribution, (ii) the mean of the sample means $\approx \mu$ and (iii) the sd of the sample means $\approx \frac{\sigma}{\sqrt{n}}$. Please verify if the CLT works with our example shown above.

Gamblers, sport fans, and investors claim there are times when they have a streak of good or bad luck. Simulation can show that luck may be nothing more than chance variation. For example, consider a simple gambling game, in which you and I each bet \$1. You toss a coin. If it comes up tail, you win; if head, you lose. If you play this game for a long time, will you have hot/cold streaks? The outcome of each game can be written as a probability distribution table like:

Win	prob
\$1	0.5
−\$1	0.5

What are the (theoretical) population mean and sd of the outcome? Lets play such games 1,000 times and see what could happen in the end (and during the game).

```
x <- c(1, -1)
prob2 <- c(0.5, 0.5)
eachGamble <- sample(x,1000,prob=prob2,rep=T)
table(eachGamble)
mean(eachGamble)
sd(eachGamble)
so_far <- cumsum(eachGamble)
plot(so_far,type="l",col="red",xlab="# of rounds")
abline(h=0,lty=2,col="blue")
```

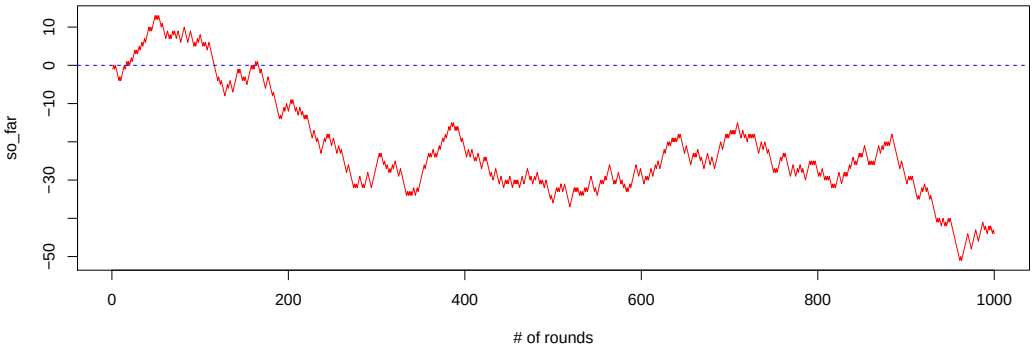

Figure 4: 1,000 rounds of *even* chance games.

You can not only see the entire game at a glance, but also see the winning and losing *streaks*. This is often called the *Gambler's Ruin* scenario. In population ecology, this scenario is called the Critter Extinction. You start with an initial population of some relatively small size (say $n = 20$). A toss is represented by a vital ecological change in which either a birth or death occurs. We can simulate if the critters go extinct and how many pivotal moments there are.

Ex 7: Challenge! Calculating π

Consider a unit circle with radius 1 and a square that circumscribes it (with each side = 2) – see below. Generate 1,000 random numbers and check how many of them are falling within the circle. We can estimate π from the proportion of points being inside the circle. That is,

$$prop(\text{inside the unit circle}) = \left(\frac{\pi}{4}\right)$$

That is,

$$\pi = 4 \cdot prop$$

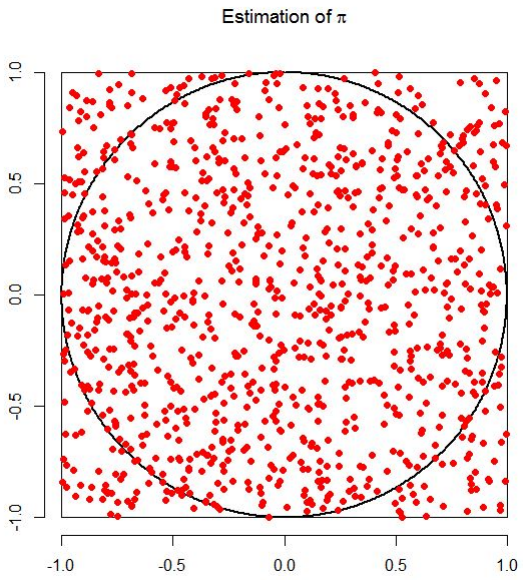

Figure 5: Count how many points are falling within the unit circle.

Here is how to do it in R.

```
n <- 1000
x <- runif(n, -1, 1)
y <- runif(n, -1, 1)
distance <- x^2 + y^2
inside_circle <- sum(distance <= 1)
(my_pi <- 4*(inside_circle/n))
corner_x <- c(-1,-1, 1, 1,-1)
corner_y <- c(-1, 1, 1,-1,-1)
angle <- seq(0,2*pi,length=1000)
par(pin=c(5,5))
plot(corner_x,corner_y,type="l",xlab="",ylab="",axes=F)
axis(1)
axis(2)
lines(cos(angle),sin(angle),lwd=2)
points(x,y,pch=16,col=2)
title(expression(paste("Estimation of ",pi)))
```

Selected Problems

1. A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a 12-hour shift. For a random sample of 50 patients, the following information was obtained. Let X = the number of times a patient rings the nurse during a 12-hour shift. For this exercise, $x = 0, 1, 2, 3, 4, 5$. $P(x)$ = the probability that X takes on value x.

X	0	1	2	3	4	5
$P(X = x)$	$\frac{4}{50}$	$\frac{8}{50}$	$\frac{16}{50}$	$\frac{14}{50}$	$\frac{6}{50}$	$\frac{2}{50}$

Why is this a discrete probability distribution function (two reasons)? Hint: Are they acceptable probabilities? What is their sum?

2. A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of 50 mothers, the following information was obtained. Let X = the number of times per week a newborn baby's crying wakes its mother after midnight. For this example, $x = 0, 1, 2, 3, 4, 5$.
   $P(x)$ = probability that X takes on a value x.

X	0	1	2	3	4	5
$P(X = x)$	$\frac{2}{50}$	$\frac{11}{50}$	$\frac{23}{50}$	$\frac{9}{50}$	$\frac{4}{50}$	$\frac{1}{50}$

   (a) Find the expected value of the number of times a newborn baby's crying wakes its mother after midnight. The expected value is the expected number of times per week a newborn baby's crying wakes its mother after midnight.
   (b) Calculate the standard deviation of the variable as well.
   (c) Simulate 500 values of X using R and find the sample mean $\left(\bar{X}\right)$ and the sample sd (s) of the simulated numbers. Are they close to μ and σ?

3. A certain university has 14 statistics classes scheduled for its Summer 2015 term. One class has space available for 30 students, eight classes have space for 60 students, one class has space for 70 students, and four classes have space for 100 students.
   (a) What is the average class size assuming each class is filled to capacity?
   (b) Space is available for 980 students. Suppose that each class is filled to capacity and select a statistics student at random. Let the random variable X equal the size of the student's class. Construct the probability distribution of X. Also, define the pmf of X.
   (c) Find the mean (μ) of X.
   (d) Find the sd (σ) of X.
   (e) Simulate 500 values of X using R and find the sample mean $\left(\bar{X}\right)$ and the sample sd (s) of the simulated numbers. Are they close to μ and σ?

4. A "friend" offers you the following "deal." For a \$10 fee, you may pick an envelope from a box containing 100 seemingly identical envelopes. However, each envelope contains a coupon for a free gift.
   - Ten of the coupons are for a free gift worth \$6.
   - Eighty of the coupons are for a free gift worth \$8.
   - Six of the coupons are for a free gift worth \$12.
   - Four of the coupons are for a free gift worth \$40.

   Based upon the financial gain or loss over the long run, should you play the game?
   (a) Let the random variable X equal your "net" profit. Construct the probability distribution of X. Also, define the pmf of X.
   (b) Choose one.
      i. Yes, I expect to come out ahead in money.
      ii. No, I expect to come out behind in money.
      iii. It doesn't matter. I expect to break even.
   (c) Simulate 50 values of X using R and find the sample mean $\left(\bar{X}\right)$ and the sample sd (s) of the simulated numbers. Show the trend of X in graph.

5. The chance of an IRS audit for a tax return with over $25,000 in income is about 2% per year. We are interested in the expected number of audits a person with that income has in a 20-year period. Assume each year is independent.
   (a) In words, define the random variable X.
   (b) List the values that X may take on.
   (c) Give the distribution of X. $X \sim$ _____ (_____, _____)
   (d) How many audits are expected in a 20-year period?
   (e) Find the probability that a person is not audited at all.
   (f) Find the probability that a person is audited more than twice.

6. There are two similar games played for Chinese New Year and Vietnamese New Year. In the Chinese version, fair dice with numbers 1, 2, 3, 4, 5, and 6 are used, along with a board with those numbers. In the Vietnamese version, fair dice with pictures of a gourd, fish, rooster, crab, crayfish, and deer are used. The board has those six objects on it, also. We will play with bets being $1. The player places a bet on a number or object. The "house" rolls three dice. If none of the dice show the number or object that was bet, the house keeps the $1 bet. If one of the dice shows the number or object bet (and the other two do not show it), the player gets back his or her $1 bet, plus $1 profit. If two of the dice show the number or object bet (and the third die does not show it), the player gets back his or her $1 bet, plus $2 profit. If all three dice show the number or object bet, the player gets back his or her $1 bet, plus $3 profit. Let X = number of matches and Y = profit per game.
   (a) List the values that X may take on.
   (b) Give the distribution of X. $X \sim$ _____ (_____, _____)
   (c) List the values that Y may take on. Then, construct one pmf table that includes both X and Y and their probabilities.
   (d) Calculate the average expected matches over the long run of playing this game for the player.
   (e) Calculate the average expected earnings over the long run of playing this game for the player.
   (f) Determine who has the advantage, the player or the house. By how much?
   (g) Simulate 1000 values of X and Y using R and find the sample mean $\left(\bar{X}\right)$ and the sample sd (s) of the random variables. Show the trend of X and Y in graph.