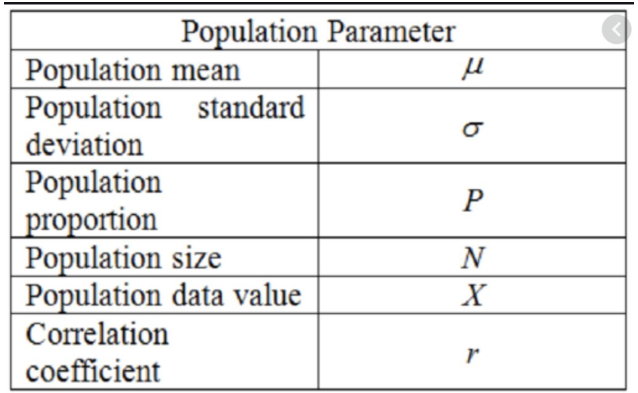
Lab 6: Sampling

Sample distribution

Why we want to sample?

- "A random sample of the data can help us get some point estimates for population" parameters (mean, sd, variance, proportion)."
- a ***sampling distribution*** of our estimate allows us to "learn about the properties of the estimate, such as its distribution".



Population Parameter	
Population mean	μ
Population standard deviation	σ
Population proportion	P
Population size	N
Population data value	X
Correlation coefficient	r

Random number Generation/seed

How does R get Random numbers to sample?

Truly random numbers are expensive

So R uses a “pseudo”random number generator.

This kind of generation is a deterministic (depends on the starting value) process that is almost the same as a true random process

The **seed** is the starting value that determines this sequence.

set.seed()

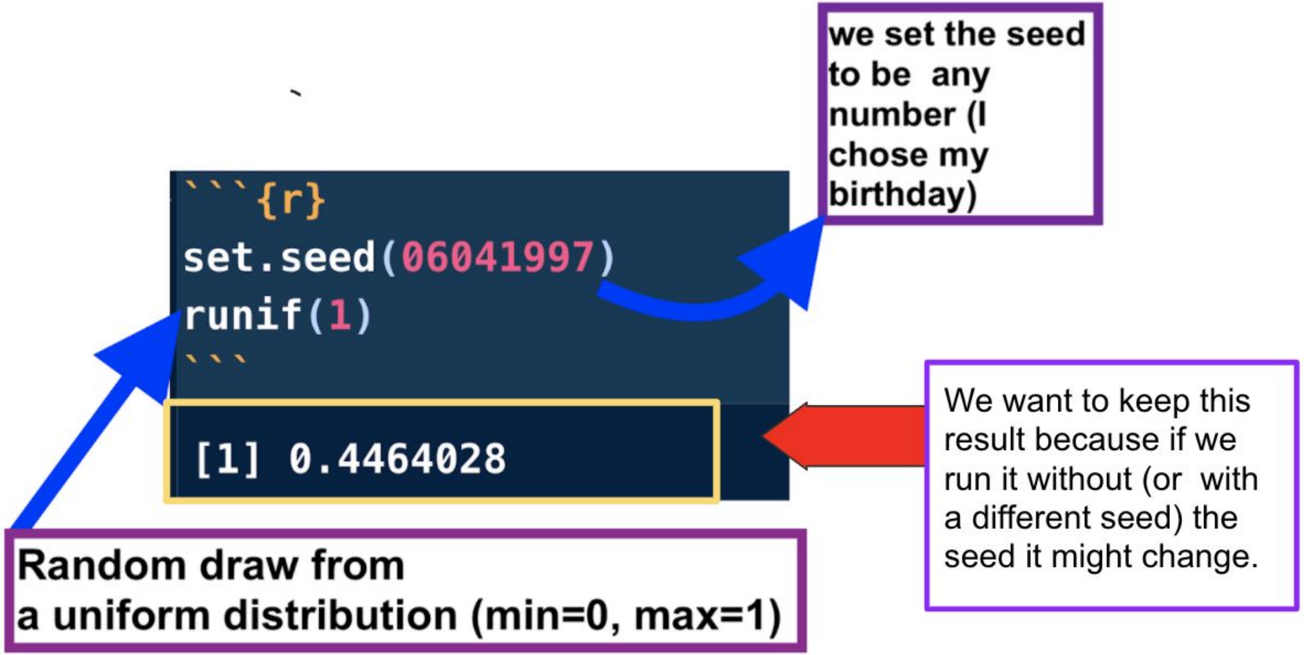
We want some way to “save” our results of a random process.

We want to use the set.seed() function to set a seed.

The set.seed() function sets the starting number used to generate a sequence of random numbers.

This means that you will get the same result if you run the same process again (if you keep the seed set at the same number).

Example



```{r}
set.seed(06041997)
runif(1)
```

```
[1] 0.4464028
```

we set the seed to be any number (I chose my birthday)

We want to keep this result because if we run it without (or with a different seed) the seed it might change.

Random draw from
a uniform distribution (min=0, max=1)

sample()

x <- vector we would like to sample from

N <- how many values do we want in the sample

sample(x, N)

We use the sample function because it is easier to work with a smaller sample than the whole data set

The output is a new vector of size N

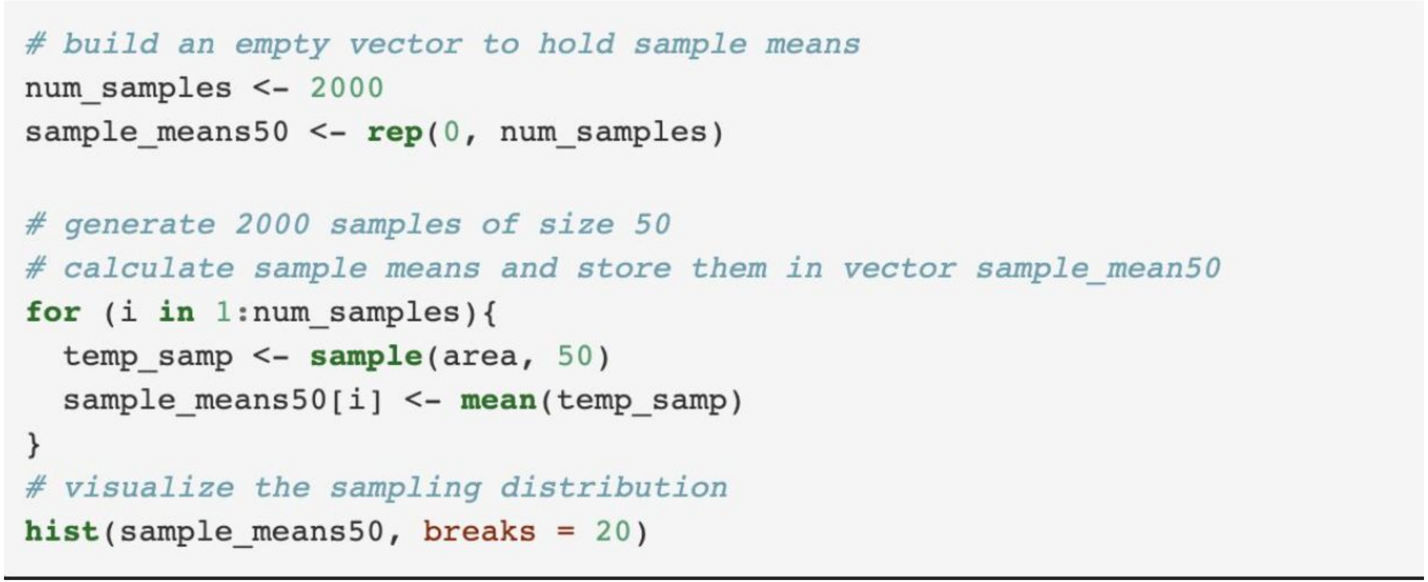
Sampling With replacement / without replacement?

sampling with replacement means that if I’m drawing from a complete deck of 52 playing cards and I first draw the queen of spades, I put it back into the deck again (so the deck still has 52 cards) so I can still have a chance of drawing the queen of spades.

sampling without replacement means that I keep the card from the first draw and for the second draw I’m picking from an incomplete deck of 51 cards .

For Loop

We want to a create a loop so we do not need to code for a task over and over again (i.e. taking many samples)



```r
# build an empty vector to hold sample means
num_samples <- 2000
sample_means50 <- rep(0, num_samples)

# generate 2000 samples of size 50
# calculate sample means and store them in vector sample_mean50
for (i in 1:num_samples){
  temp_samp <- sample(area, 50)
  sample_means50[i] <- mean(temp_samp)
}
# visualize the sampling distribution
hist(sample_means50, breaks = 20)
```