Simulations using known populations

Arturo Valdivia

2/21/2023

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.2.1     v stringr 1.4.1
## v readr   2.1.3     v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(moderndive)
```

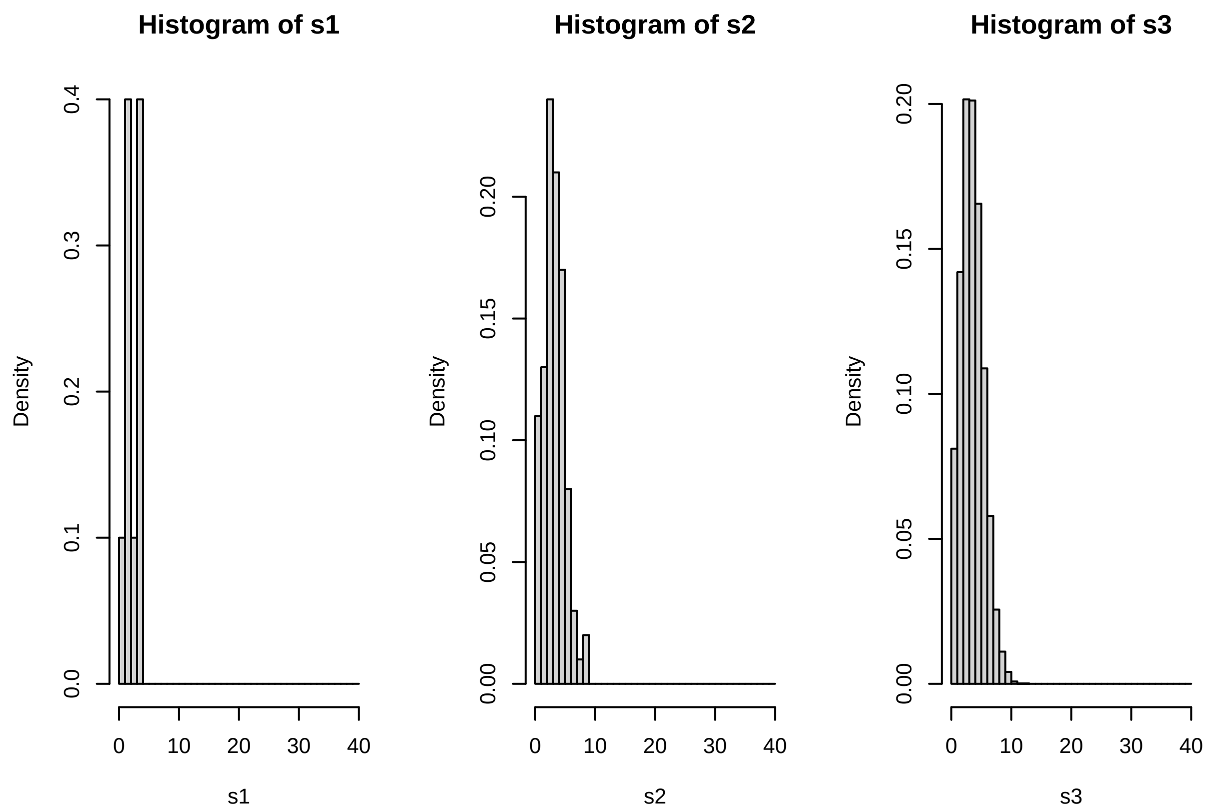
We can approximate the probability distribution with a large number of replications.

Example 1

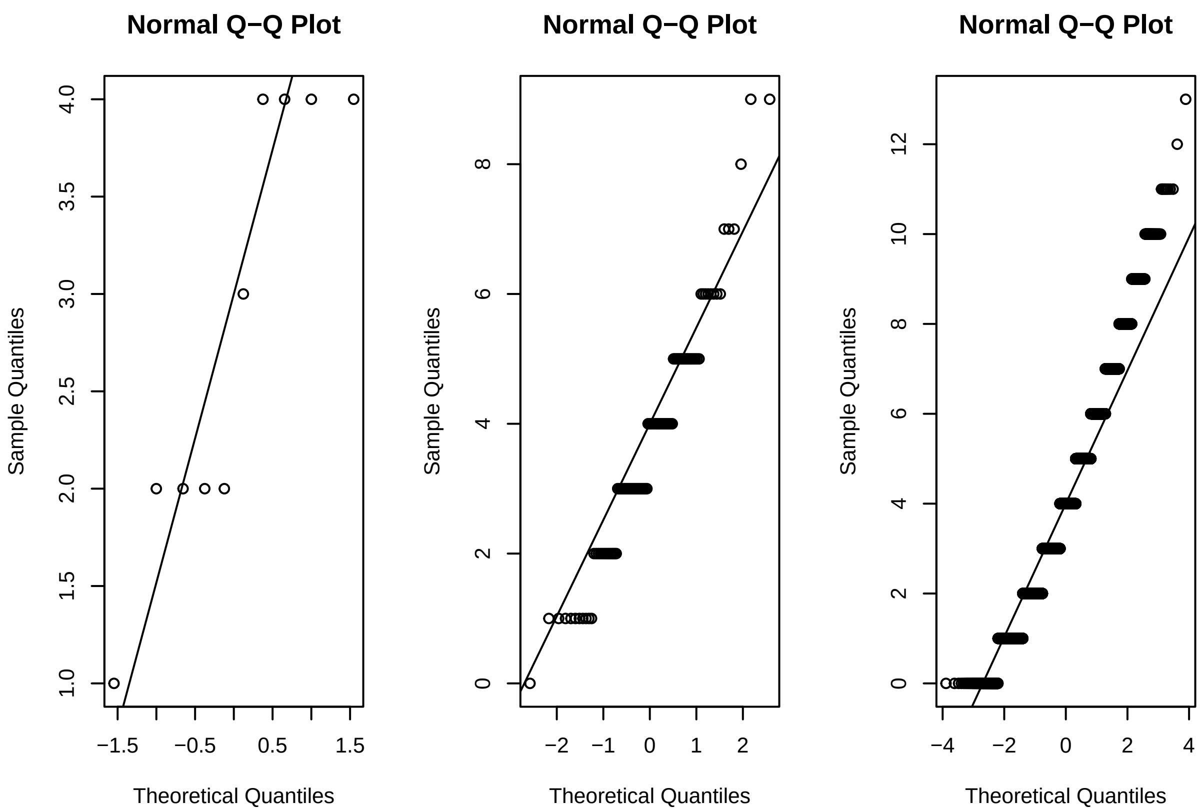
$X \sim binomial(40, 0.1)$ Let's get a sample of N replicates for N=10, 100, and 10000

```
set.seed(520)
s1 = rbinom(10, 40, 0.1)
s2 = rbinom(10^2, 40, 0.1)
s3 = rbinom(10^4, 40, 0.1)

op = par(mfrow=c(1,3))
hist(s1, freq = F, breaks = 0:40)
hist(s2, freq = F, breaks = 0:40)
hist(s3, freq = F, breaks = 0:40)
```




```
qqnorm(s1);qqline(s1)
qqnorm(s2);qqline(s2)
qqnorm(s3);qqline(s3)
```




```
par(op)
```

$EX = np = 40 * 0.1$

```
EX = 40*0.1
```

$VarX = np(1-p) = 40 * 0.1 * 0.9$

```
VarX = 40*0.1*0.9

EX = 40*.1
c(EX, mean(s1), mean(s2), mean(s3))
```

```
## [1] 4.000 2.800 3.720 4.004
```

```
VarX = 40*.1*.9
my.var = function(x) mean(x^2) - mean(x)^2
c(VarX, my.var(s1), my.var(s2), my.var(s3))
```

```
## [1] 3.600000 1.160000 3.161600 3.606384
```

```
pbinom(5, 40, 0.1)
```

```
## [1] 0.7937273
```

```r
c(mean(s1 <= 5), mean(s2 <= 5), mean(s3 <= 5))
```

```
## [1] 1.0000 0.8600 0.7915
```

Example 2

Population is a bowl with white and red balls and success is getting a red ball $X \sim Bernoulli(p)$ Let's get N balls with replacement for N=10, 100, and 10000

```r
pop.vec = bowl$color

set.seed(100)
sam1 = sample(pop.vec, 10, T)
sam2 = sample(pop.vec, 10^2, T)
sam3 = sample(pop.vec, 10^4, T)

# EX = p
p1 = mean(sam1 == "red")
p2 = mean(sam2 == "red")
p3 = mean(sam3 == "red")

c(p1, p2, p3)
```

```
## [1] 0.2000 0.4000 0.3703
```

```r
# population
mean(pop.vec == "red")
```

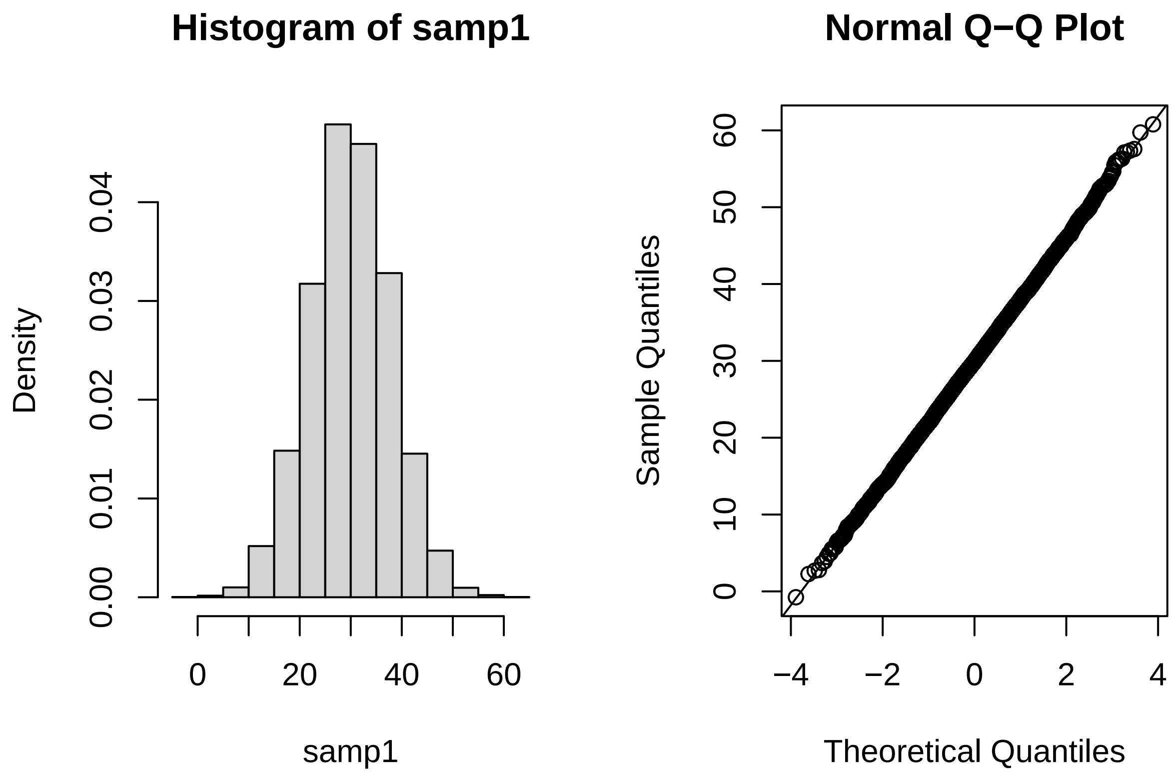
```
## [1] 0.375
```

Example 3

$X \sim Normal(30, 64)$ for N=10000

```r
set.seed(123)
samp1 = rnorm(10^4, 30, 8)

op = par(mfrow=c(1,2))
hist(samp1, freq = F)
qqnorm(samp1);qqline(samp1)
```




```
par(op)
```

$EX = 30$

```
mean(samp1)
```

```
## [1] 29.98103
```

$VarX = 64$

```
my.var = function(x) mean(x^2) - mean(x)^2
my.var(samp1)
```

```
## [1] 63.81923
```

Simulations for sampling distributions (sample mean)

Simulation 1

$X_1, X_2, \ldots, X_{32} \sim binomial(40, 0.1)$. $\bar{X}_n$ is the sample mean. Let's get one sample mean $\bar{x}$:

```
set.seed(100)
mean(rbinom(32, 40, 0.1))
```

```
## [1] 3.90625
```

Let's now get 10^4 sample means

```
xbar.vec = replicate(n = 10^4, expr = mean(rbinom(32, 40, 0.1)))
```

The expected value, $E\bar{X}_{32}$, and

```
mean(xbar.vec)
```

```
## [1] 4.002584
```

The theoretical value of $E\bar{X}_{32} = \mu = np$

```
40*0.1
```

```
## [1] 4
```

The variance, $Var\bar{X}_{32}$

```
my.var(xbar.vec)
```

```
## [1] 0.1151559
```

The theoretical value of $Var\bar{X}_{32} = \sigma^2/32 = np(1-p)/32$

```
40*0.1*0.9/32
```

```
## [1] 0.1125
```

Simulation 2

The population is a bowl with white and red balls. Let's take a sample of 25 balls, with replacement. The random sample can be represented by: $X_1, X_2, \ldots, X_{50} \sim Bernoulli(p)$ where $\bar{X}_{50}$ is the sample average, or proportion of successes out of 50 and $E\bar{X}_{50} = p$ and $Var\bar{X}_50 = p(1-p)/50$.

Here is 1 proportion obtained from a sample

```
pop.vec = (bowl$color == "red") #population
mean(sample(pop.vec, 50, replace = T))
```

```
## [1] 0.32
```

Let's get the sample proportion for 10^4 replicates

```
samp.vec = replicate(10^4, mean(sample(pop.vec, 50, replace = T)))
```

$E\bar{X}_{50} = p$ and

```
mean(samp.vec) # this is the simulated value of $E \bar X_{50}$
```

```
## [1] 0.374382
```

the theoretical value of $E\bar{X}_{32} = EX = p$

```
p = mean(pop.vec) # This is p
p
```

```
## [1] 0.375
```

$Var\bar{X}_{50} = p(1-p)/50$

```
my.var(samp.vec)
```

```
## [1] 0.004725118
```

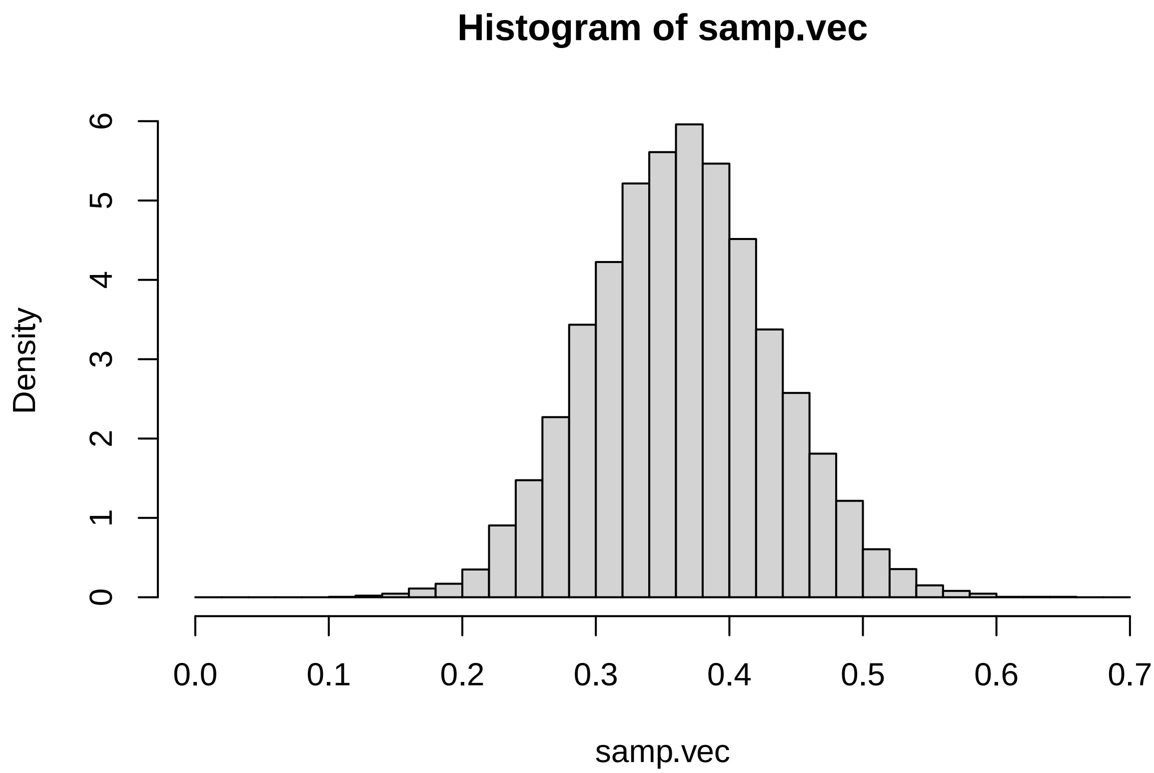
The theoretical value for $Var\bar{X}_{50} = \sigma^2/50 = p(1-p)/50$

```
p*(1-p)/50
```

```
## [1] 0.0046875
```

Here is the graphical representation of the distribution of $\bar{X}_{50}$

```
hist(samp.vec, freq = F, breaks = seq(from=0, to=0.7, by=0.02))
```


Histogram of samp.vec
Density
0
1
2
3
4
5
6
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
samp.vec