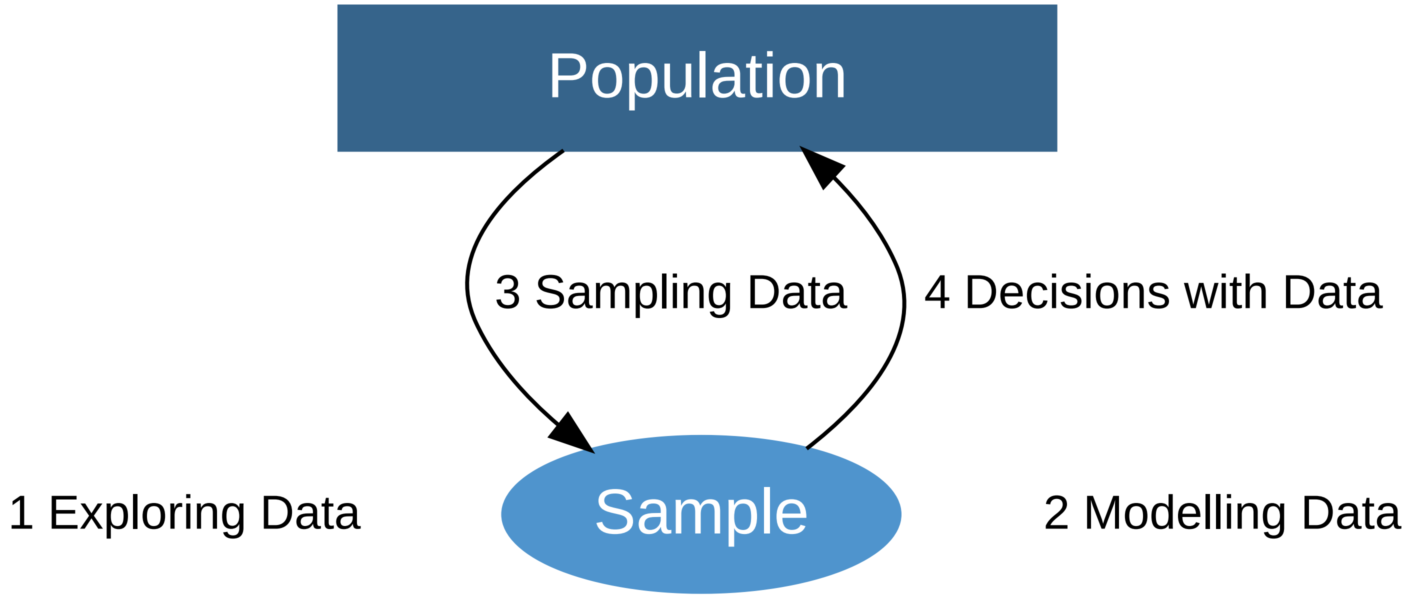
Normal Approximation

Sampling Data | Chance Variability

Unit Overview



Module3 Sampling Data

Understanding Chance

What is chance?

Chance Variability

How can we model chance variability by a box model?

Sample Surveys

How can we model the chance variability in sample surveys?

The Normal Approximation

Data Story | Why is the Normal curve so common?

The Probability Histogram

A crazy result?

The Central Limit Theorem (CLT)

Example

Does the CLT work for products?

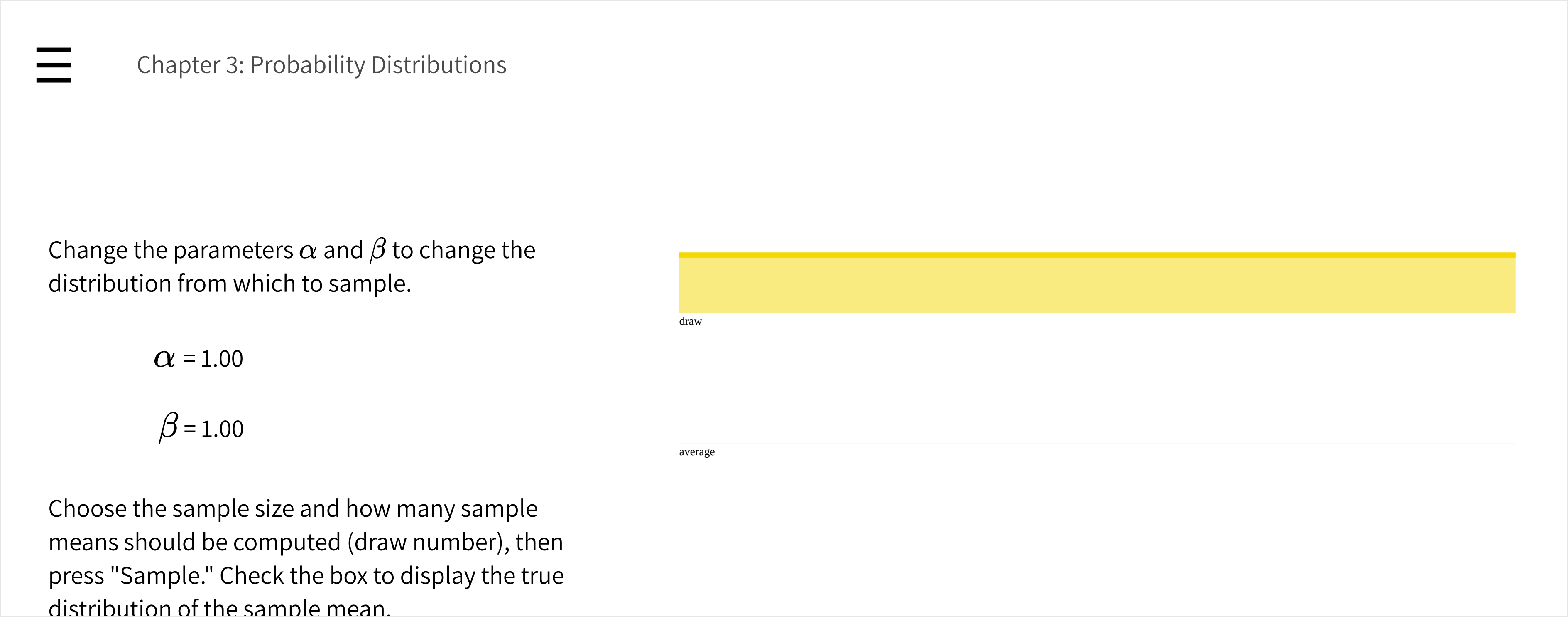
Summary

Data Story

Why is the Normal curve so common?

Why does this work?

Have a play with this app. Why does it work?



Chapter 3: Probability Distributions

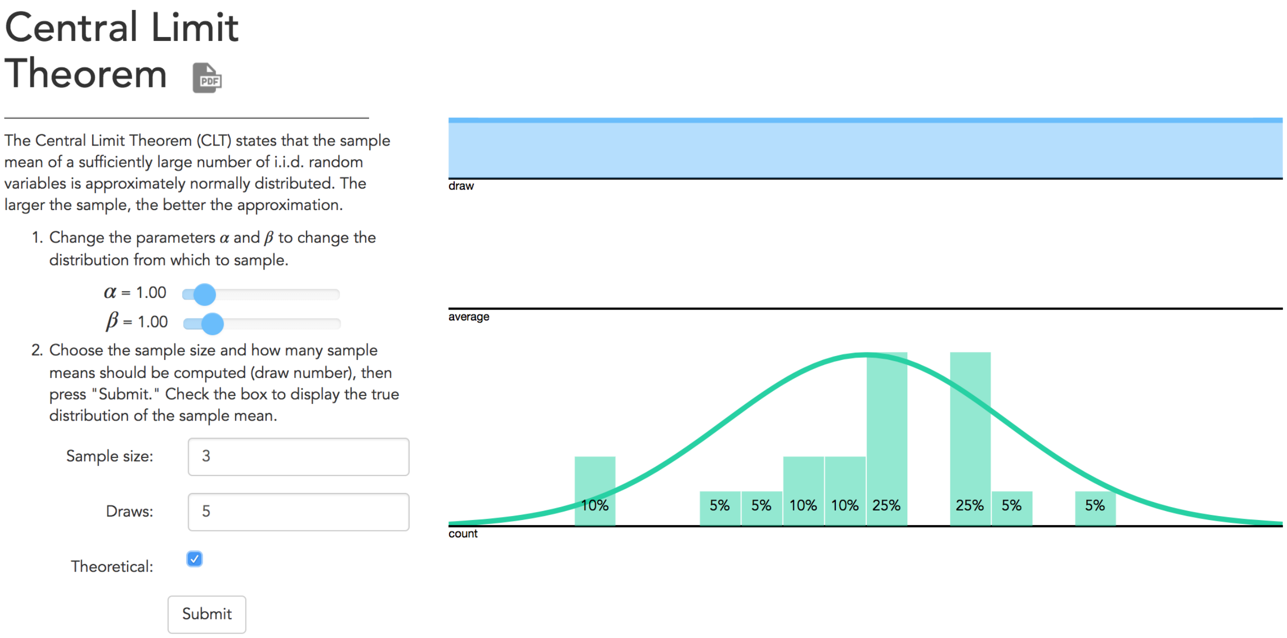
Change the parameters α and β to change the distribution from which to sample.

α = 1.00

β = 1.00

Choose the sample size and how many sample means should be computed (draw number), then press "Sample." Check the box to display the true distribution of the sample mean.

App



Central Limit Theorem

The Central Limit Theorem (CLT) states that the sample mean of a sufficiently large number of i.i.d. random variables is approximately normally distributed. The larger the sample, the better the approximation.

1. Change the parameters α and β to change the distribution from which to sample.

   α = 1.00

   β = 1.00

2. Choose the sample size and how many sample means should be computed (draw number), then press "Submit." Check the box to display the true distribution of the sample mean.

Sample size:	3
Draws:	5
Theoretical:	☑

Submit

The probability histogram

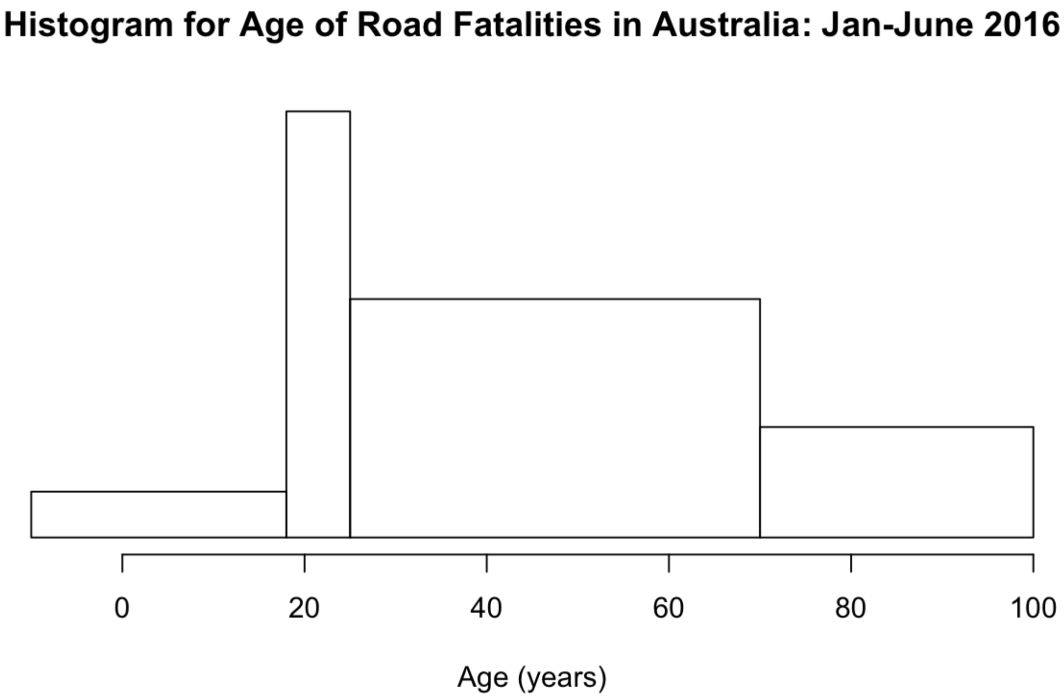
3 types of histograms

There are 3 types of histograms.

Type	Use
Data histogram	Represents data by area
Probability histogram	Represents chance by area
Simulation (empirical) histogram	Converges in shape to the probability histogram

Data histogram

- The **data histogram** data represents the amount of data by area: `hist()`



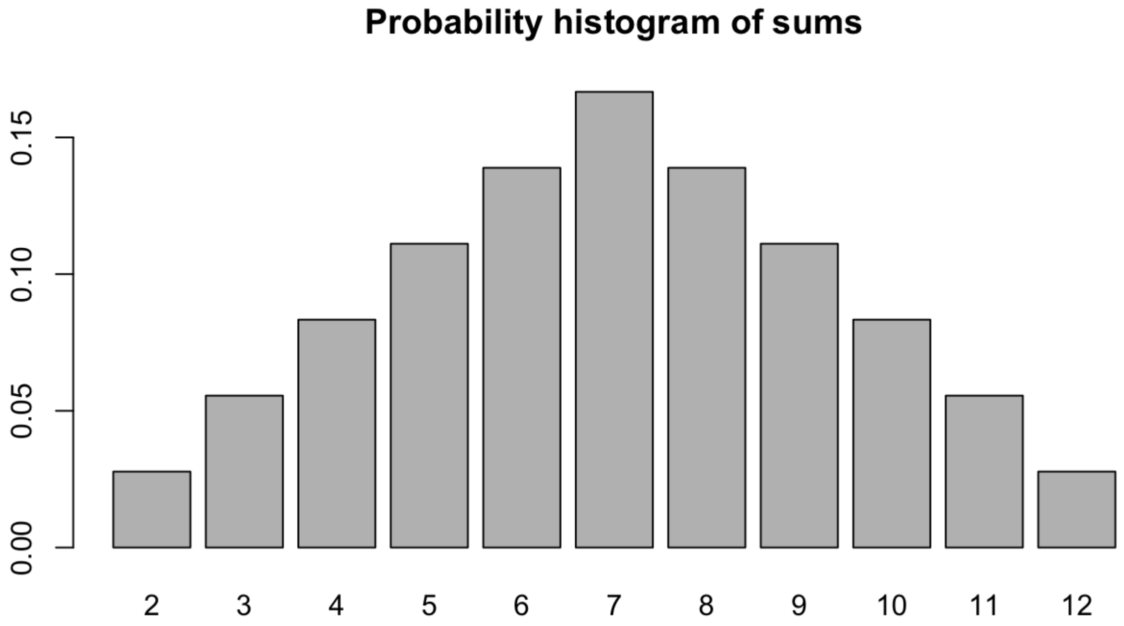
Probability histogram

- The **probability histogram** represents **chance** by area.

Example: Toss a pair of dice and calculate the sum.

sum	2	3	4	5	6	7	8	9	10	11	12
chance	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

```r
sum=c(2:12)
chance=c(1,2,3,4,5,6,5,4,3,2,1)/36
t=data.frame(sum,chance)
barplot(t$chance,names.arg=t$sum,main="Probability histogram of sums")
```



Simulation histogram

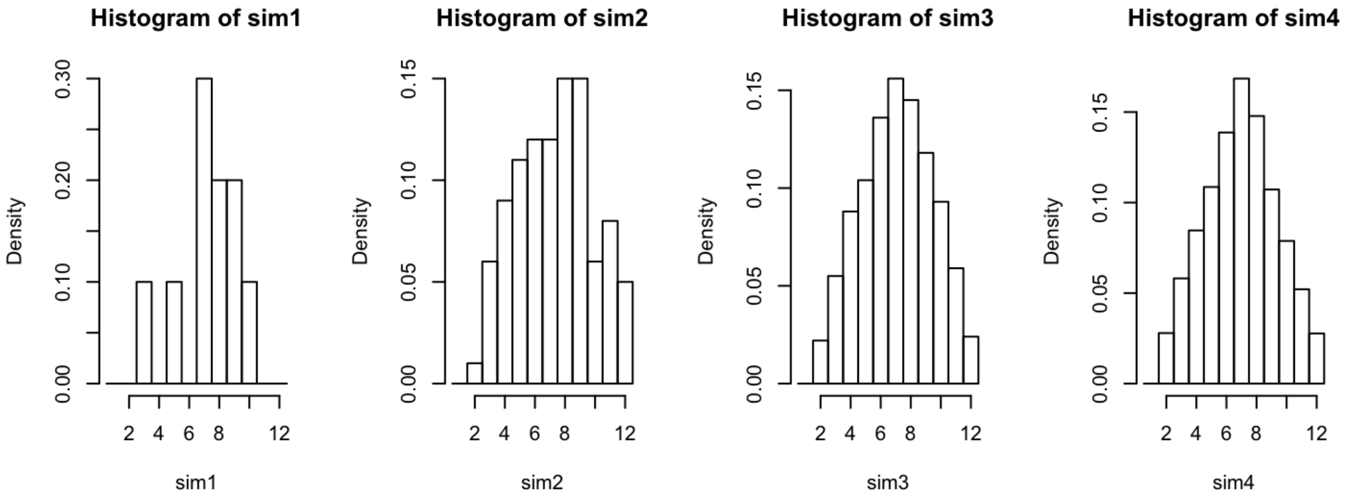
- The **simulation histogram** represents **chance** by area, for a simulation of a chance process.

Example: Toss a pair of dice and calculate the sum.

Here we study the simulation histogram for the sum of 2 draws from the same box (“dice”) with different number of replicates.

```
set.seed(10)
dice=c(1:6)
sim1 = sample(dice,replace=T,10)+sample(dice,replace=T,10)
sim2 = sample(dice,replace=T,100)+sample(dice,replace=T,100)
sim3 = sample(dice,replace=T,1000)+sample(dice,replace=T,1000)
sim4 = sample(dice,replace=T,10000)+sample(dice,replace=T,10000)
```

```
par(mfrow=c(1,4))
#breaks=c(2:12)
breaks=c(0.5:12.5)
hist(sim1,br=breaks,freq=F)
hist(sim2,br=breaks,freq=F)
hist(sim3,br=breaks,freq=F)
hist(sim4,br=breaks,freq=F)
```



Convergence of simulation histograms

For repeated simulations of a chance process resulting in a sum, the **simulation** histogram of the observed values **converges** to the **probability** histogram.

Expected value and standard error of probability histogram

For a probability histogram, on the horizontal axis,

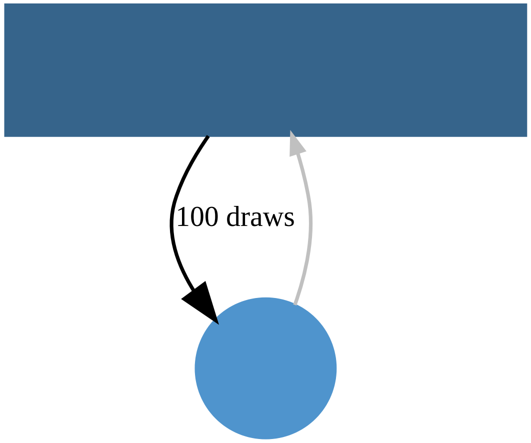
- the **EV** measures the **centre**
- the **SE** measures the **spread**.

A crazy result?

Simulation from different boxes

Consider 4 very different box models:

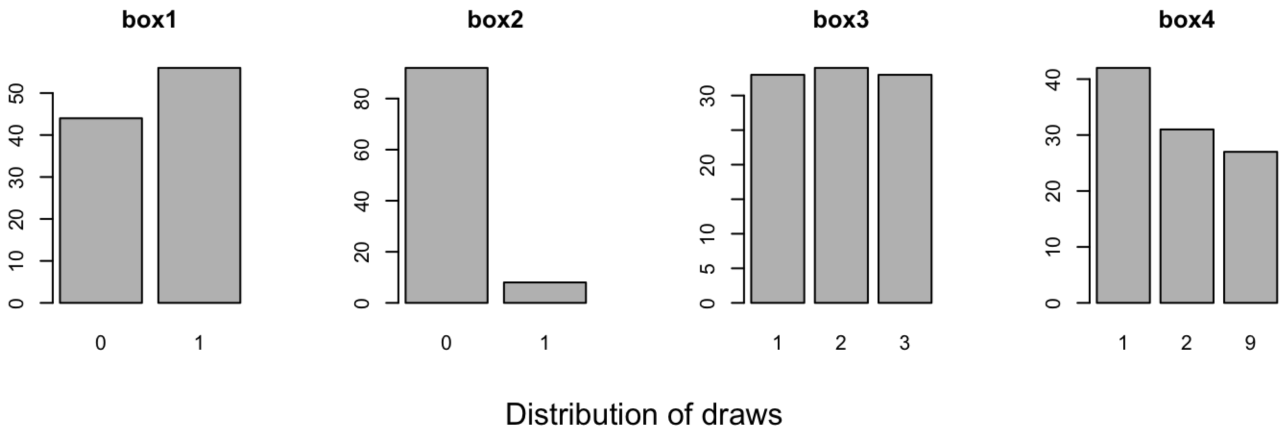
- 0,1
- 9x0, 1
- 1,2,3
- 1,2,9



1. Sample 100 times from each box, and plot the results

```
box1=c(0,1)
box2=c(0,0,0,0,0,0,0,0,0,1)
box3=c(1,2,3)
box4=c(1,2,9)
s1=sample(box1,replace=T,100)
s2=sample(box2,replace=T,100)
s3=sample(box3,replace=T,100)
s4=sample(box4,replace=T,100)
```

```r
par(mfrow=c(1,4))
barplot(table(s1),main="box1")
barplot(table(s2),main="box2")
barplot(table(s3),main="box3")
barplot(table(s4),main="box4")
mtext("Distribution of draws", side =1, line = -1, outer = TRUE)
```

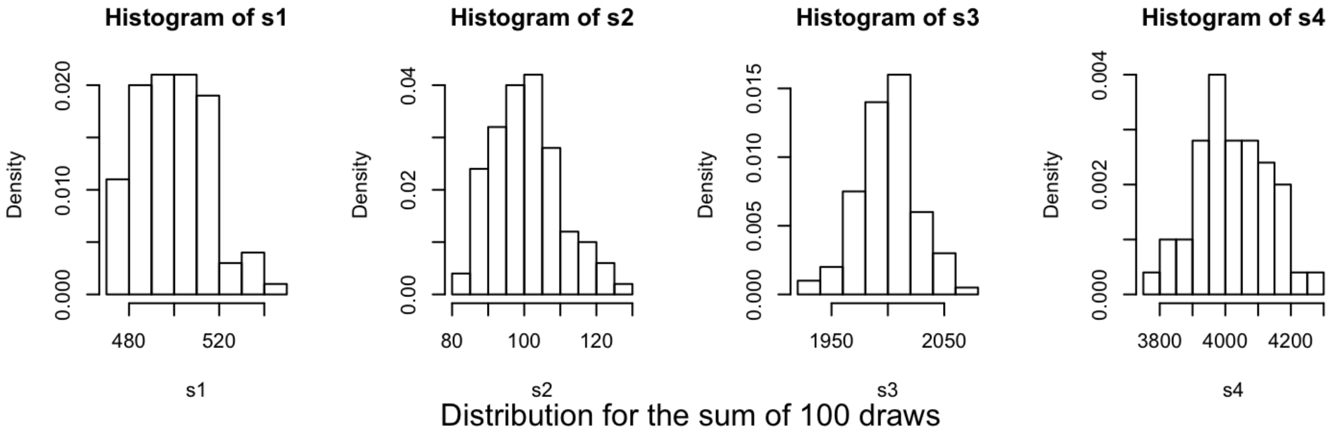


Notice the difference between the composition of the boxes.

2. Sum up 1000 draws from each box, repeat 100 times, and plot the results

```
s1=replicate(100,sum(sample(box1,replace=T,1000)))
s2=replicate(100,sum(sample(box2,replace=T,1000)))
s3=replicate(100,sum(sample(box3,replace=T,1000)))
s4=replicate(100,sum(sample(box4,replace=T,1000)))
```

```
par(mfrow=c(1,4))
hist(s1, prob=T)
hist(s2, prob=T)
hist(s3, prob=T)
hist(s4, prob=T)
mtext("Distribution for the sum of 100 draws", side =1, line = -0.9, outer = TRUE, cex=1)
```



Notice the simulation histograms of the sums all seem to have a **normal** shape. Why?

The Central Limit Theorem

The Central Limit Theorem

The Central Limit Theorem (CLT)

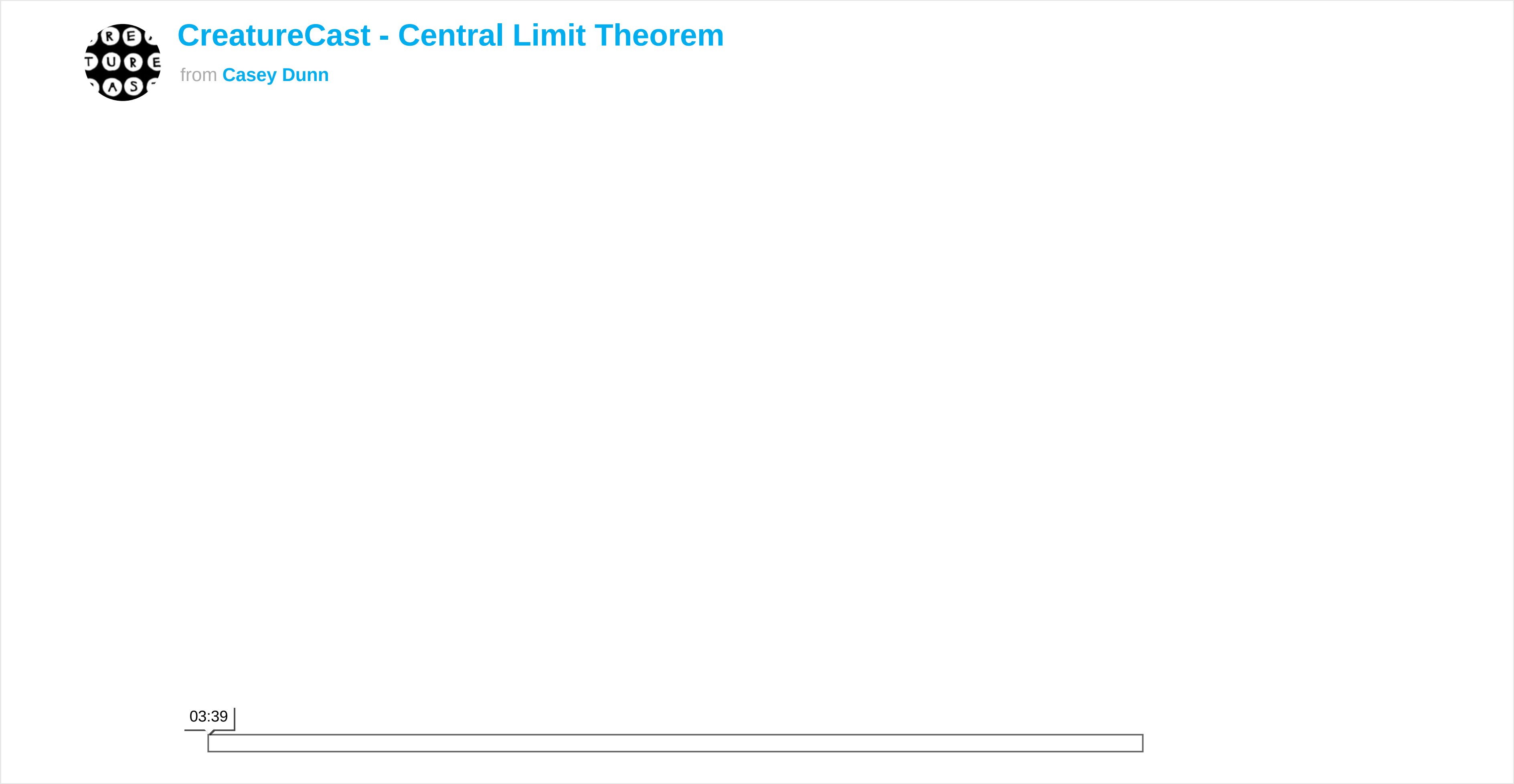
When drawing at random with replacement from a box, if the sample size for the sum (or average) is sufficiently large, then

- the **probability histogram** for the sum (or average) will closely follow the **normal curve**
- even if the contents of the box does not.

More generally: The distribution (behaviour of the chances) for the sum (or average) will closely follow the normal curve.

Conditions for the CLT

- The number of draws must be reasonably large (especially if the histogram of the box differs from the normal curve).
- How large? This depends on the shape of the histogram.
- A common convention is the number of draws larger than 30 (assuming a basically symmetric distribution with no obvious outliers).



CreatureCast - Central Limit Theorem

from Casey Dunn

player.vimeo.com/video/75089338

Example

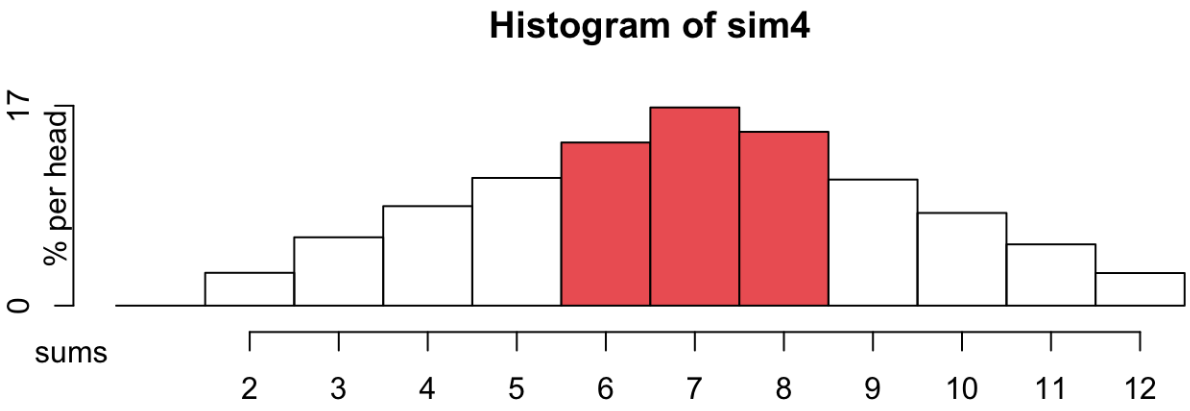
Example

Based on Sim4.Toss a pair of dice 10000 times and calculate the sum. What is the probability of getting a sum between 6 and 8?

Method1: Approximate the area from the data histogram (approximating the probability histogram)

```
length(sim4[sim4>=6 & sim4<=8])/length(sim4)
```

```
## [1] 0.455
```



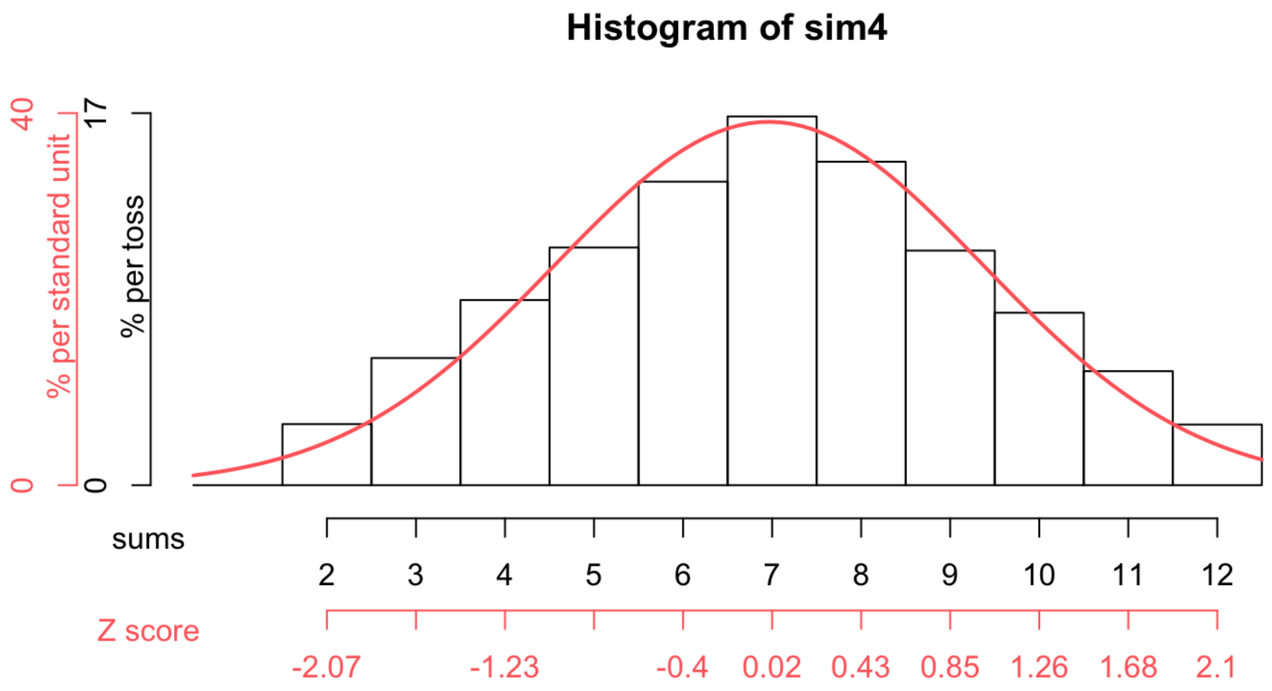
Method2: Use the empirial mean and SD of the sample (sim4) to model the Sum of the sample.

```
mean(sim4)
```

```
## [1] 6.9639
```

```
sd(sim4)
```

```
## [1] 2.40266
```

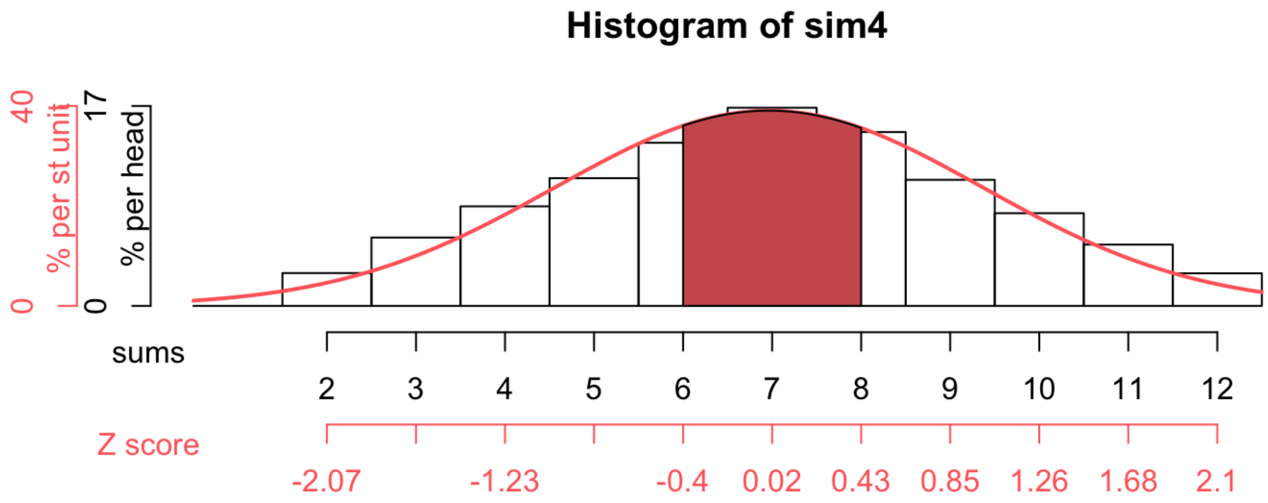


- Black y-axis: Percentage per toss
- Red y-axis: Percentage per standard units = SE * Percentage per toss

The number of standard units is between ± 0.41.

```
pnorm(0.41)-pnorm(-0.41)
```

```
## [1] 0.3181941
```



Method3: Work out the exact results for the EV and SE, to model the Sum of the sample

- The box represents a dice (1,2,3,4,5,6).
- The mean of the box is 3.5 and the SD of the box = $\text{RMS} = \sqrt{\frac{(1-3.5)^2+...+(6.3.5)^2}{6}} = 1.708$.
- For the sum of 2 tosses of the dice: $\text{EV} = 2 \times 3.5 = 7$ and $\text{SE} = \sqrt{2} \times 1.708 = 2.415$.
- Hence, the number of standard units (red) from 7 to 8 is $\frac{8-7}{2.415} = 0.414$.
- For the vertical axis (black): % per head is $\frac{6}{36} = 17$% (see p11).
- For the vertical axis (red): % per standard unit is % per head $\times$ $\text{SE} = 17$%.

Continuity Correction (for edges)

Notice that the Normal curve is missing part of the area calculated by the data histogram. To remedy this we adjust by 0.5 either side.

Lower threshold = 6 -> 5.5

- This has standard unit = -0.6 (approx)

Upper threshold = 8 -> 8.5

- This has standard unit = 0.6 (approx)

```r
pnorm(0.6)-pnorm(-0.6)
```

```
## [1] 0.4514938
```

The Continuity Correction (cc)

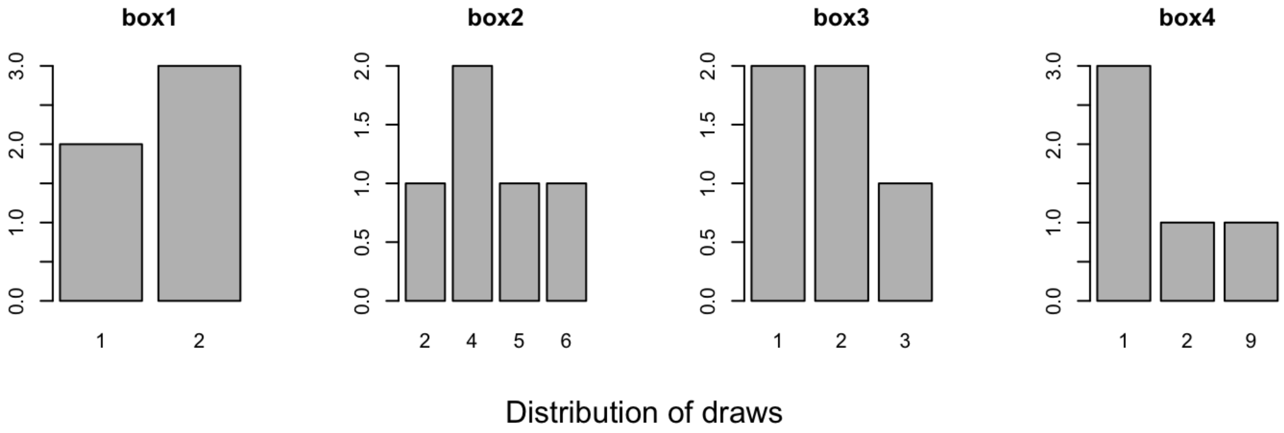
- To approximate a discrete distribution by the Normal distribution (continuous) , we adjust the endpoints by 0.5. This is called the **continuity correction**.
- To work out whether to add or minus 0.5, draw a sketch of the histogram.

Does the CLT work for products?

1. Sample 5 times from each box and plot the results

```
box1=c(1,2)
box2=c(1,2,3,4,5,6)
box3=c(1,1,1,2,2,3)
box4=c(1,2,9,11)
s1=sample(box1,replace=T,5)
s2=sample(box2,replace=T,5)
s3=sample(box3,replace=T,5)
s4=sample(box4,replace=T,5)
```

```r
par(mfrow=c(1,4))
barplot(table(s1),main="box1")
barplot(table(s2),main="box2")
barplot(table(s3),main="box3")
barplot(table(s4),main="box4")
mtext("Distribution of draws", side =1, line = -1, outer = TRUE)
```

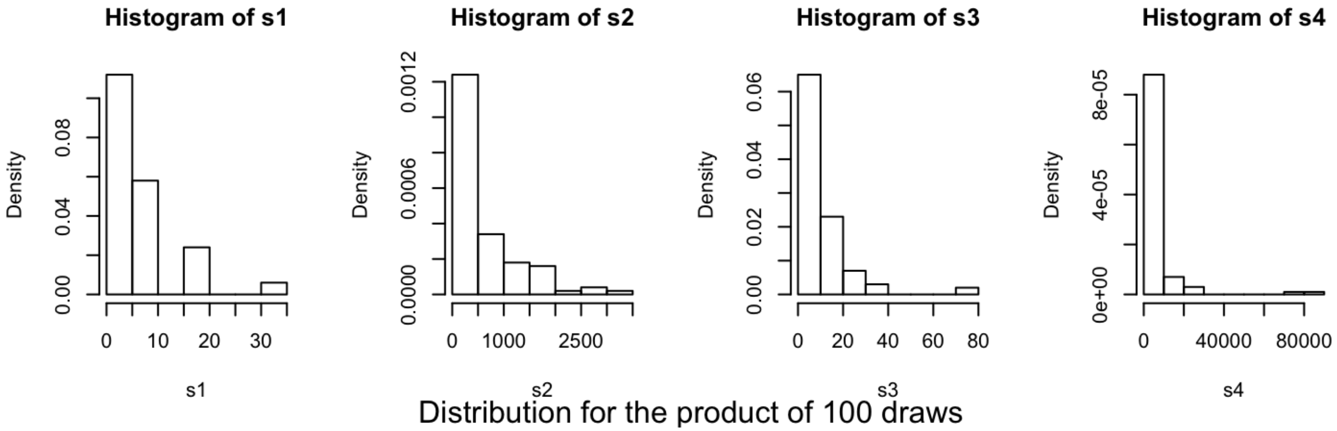


Notice the difference between the composition of the boxes.

2. Take the product of the sample from each box, repeat 100 times, and plot the results

```
s1=replicate(100,prod(sample(box1,replace=T,5)))
s2=replicate(100,prod(sample(box2,replace=T,5)))
s3=replicate(100,prod(sample(box3,replace=T,5)))
s4=replicate(100,prod(sample(box4,replace=T,5)))
```

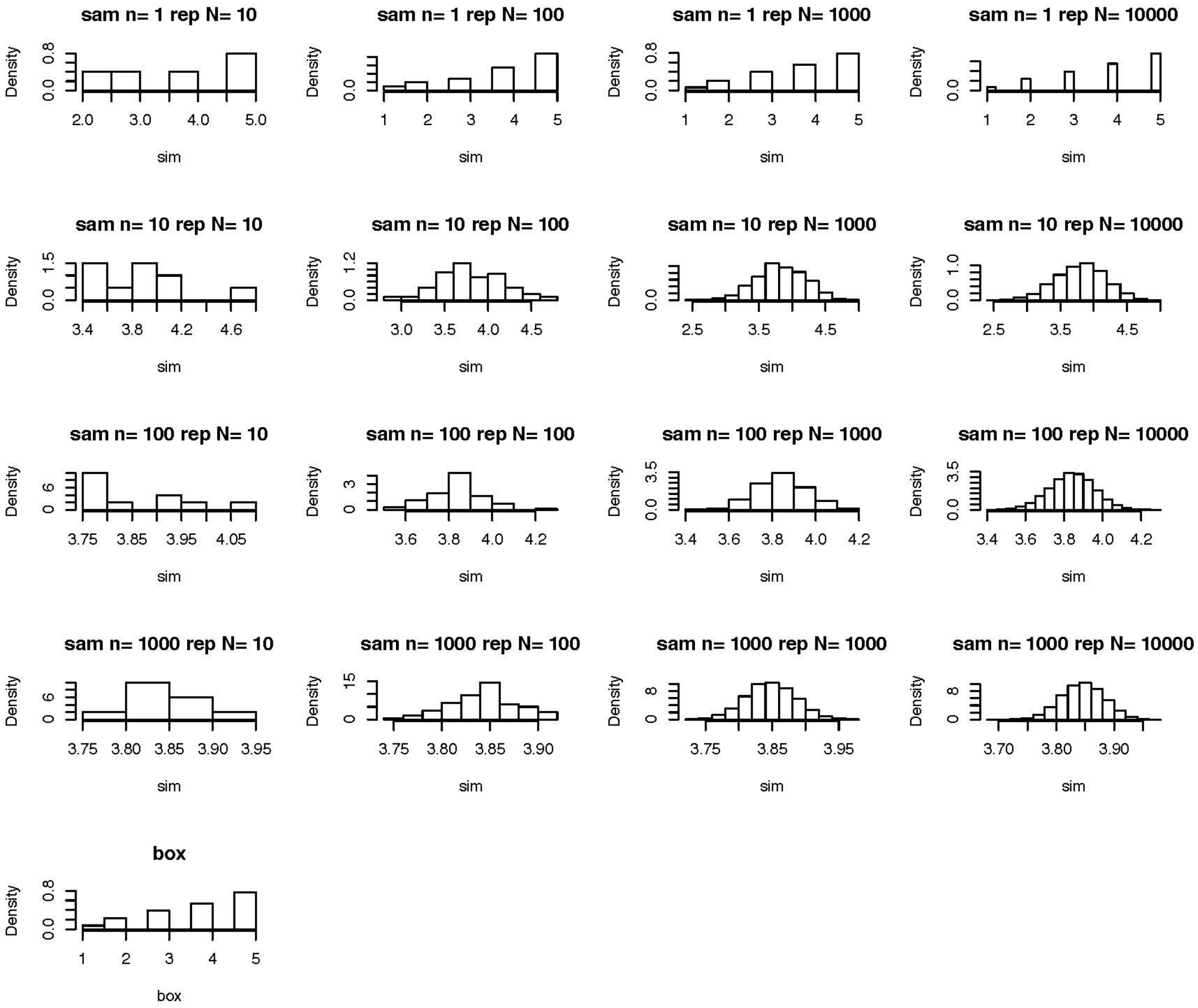
```
par(mfrow=c(1,4))
hist(s1, prob=T)
hist(s2, prob=T)
hist(s3, prob=T)
hist(s4, prob=T)
mtext("Distribution for the product of 100 draws", side =1, line = -1, outer = TRUE)
```



Notice the products do not follow a normal shape.

Sample Size vs Replicates

- It is important to distinguish between the effect of sample size n and replicates N.
- Consider a skewed box. Note that increasing the replicates across row 1 will approach the box distribution, NOT the Normal curve. However, increasing the sample size (going down the columns) approaches the Normal curve.


sam n= 1 rep N= 10
Density
0.0 0.8
2.0 3.0 4.0 5.0
sim
sam n= 1 rep N= 100
Density
0.0
1 2 3 4 5
sim
sam n= 1 rep N= 1000
Density
0.0 0.8
1 2 3 4 5
sim
sam n= 1 rep N= 10000
Density
0.0
1 2 3 4 5
sim
sam n= 10 rep N= 10
Density
0.0 1.5
3.4 3.8 4.2 4.6
sim
sam n= 10 rep N= 100
Density
0.0 1.2
3.0 3.5 4.0 4.5
sim
sam n= 10 rep N= 1000
Density
0.0
2.5 3.5 4.5
sim
sam n= 10 rep N= 10000
Density
0.0 1.0
2.5 3.5 4.5
sim
sam n= 100 rep N= 10
Density
0 6
3.75 3.85 3.95 4.05
sim
sam n= 100 rep N= 100
Density
0 3
3.6 3.8 4.0 4.2
sim
sam n= 100 rep N= 1000
Density
0.0 3.5
3.4 3.6 3.8 4.0 4.2
sim
sam n= 100 rep N= 10000
Density
0.0 3.5
3.4 3.6 3.8 4.0 4.2
sim
sam n= 1000 rep N= 10
Density
0 6
3.75 3.80 3.85 3.90 3.95
sim
sam n= 1000 rep N= 100
Density
0 15
3.75 3.80 3.85 3.90
sim
sam n= 1000 rep N= 1000
Density
0 8
3.75 3.85 3.95
sim
sam n= 1000 rep N= 10000
Density
0 8
3.70 3.80 3.90
sim
box
Density
0.0 0.8
1 2 3 4 5
box

Summary

For repeated simulations of a chance process resulting in a sum, the simulation histogram of the observed values converges to the Normal probability histogram.

Key Words

data histogram, simulation histogram, probability histogram, The Central Limit Theorem

Extension

The Central Limit Theorem

Let $X_1, X_2, X_3, \ldots X_n$ be iid (independent and identically distributed) random variables with population mean μ and variance $\sigma^2 > 0$ Then as $n \to \infty$,

$$\bar{X} \to N(\mu, \frac{\sigma^2}{n})$$

🔗 Proof of the CLT