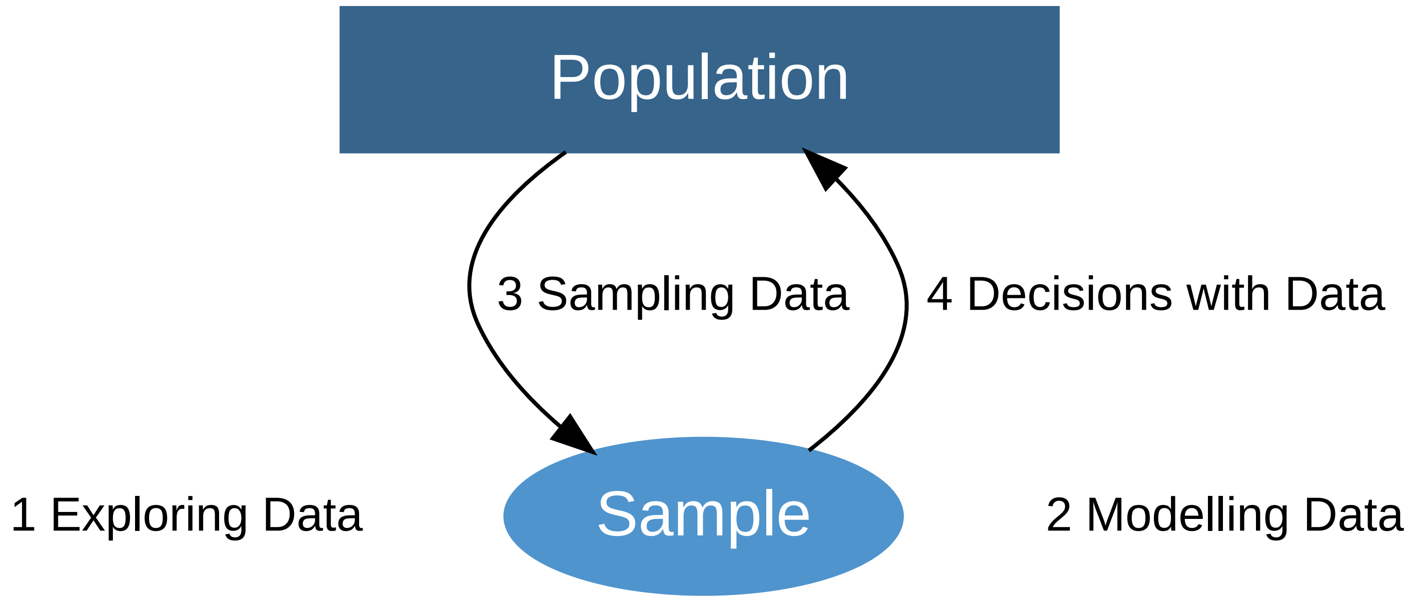
The Box Model

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?

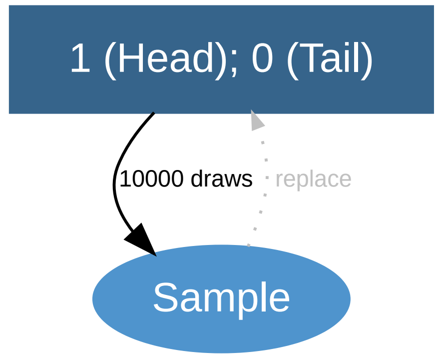
The Box Model

Data Story

Coin tossing in WWII

Coin tossing in WWII

Kerrich and Christensen's experiment can be described a simple Box Model.



- How many heads do we expect? [5000 = EV]
- How many heads did they observe? [5067 = OV]
- How big was the chance error? [5,067-5,000 = CE]

Now we more formally define the box model, so we can model the chance error by the standard error (SE).

The Box Model: Modelling the Sum/Mean of a Sample

Case 1: Sum of draws from a box model

Sum of draws from a box model

For the **Sum** of random draws from a box model with replacement,

$$\text{observed value} = \text{expected value} + \text{chance error}$$

where:

$$\text{expected value (EV)} = \text{number of draws} \times \text{mean of the box}$$

$$\text{standard error (SE)} = \sqrt{\text{number of draws}} \times \text{SD of the box}$$

How to work out the SD of the box

- The result for the standard error (SE) is called the **square root law**.
- As the **box** represents a population, the **SD of the box** is the **population SD**.
- We could call it SD_{pop}, but in this context will will simply use SD.

3 ways to calculate the SD of the box

1. Formula: RMS(gaps) = Root of the Mean of the Squared gaps.
2. R: `popsd()` with package `multicon`
3. Short cut (for simple binary boxes)

Short cut for SD of binary box

If a box only contains 2 different numbers (“big” and “small”), then

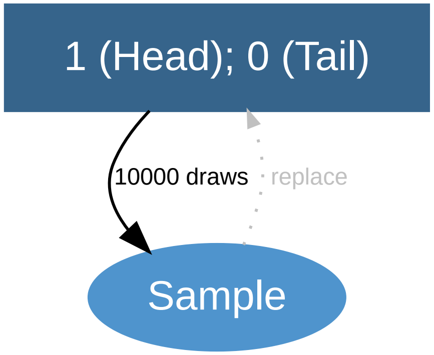
$$\mathrm{SD} = (\text{big -small})\sqrt{\text{proportion of big} \times \text{proportion of small}}$$

How does chance error relate to standard error?

- An observed value is likely to be around its expected value, with a **chance error** similar to the **SE**.
- Observed values are rarely more than 2 or 3 SEs aways from the expected value.

Example: WWII Coin Tossing

Step1: Draw the box model



Step2: Calculate the mean and SD of the box

- The mean of the box is $\frac{0+1}{2} = 0.5$.
- The SD of the box is $\sqrt{\frac{(1-(0.5))^2+(0-(0.5)^2)}{2}} = 0.5$.
- Or using the short cut, the SD is $(1-0)\sqrt{1/2 \times 1/2} = 0.5$.

Step3: Calculate the EV and SE of the Sum of the Sample

- The EV of the Sum of the draws is $10000 \times 0.5 = 5000$.
- The SE of the Sum of the draws is $\sqrt{10000} \times 0.5 = 50$.

Step4: Conclusion

- We would expect a Sample Sum of 5000 (EV) with SE 50.
- Note: We observed a sample sum of 5067 (OV) with chance error 67.

In R

```
library(multicon)
box=c(1,0)

mean(box)
```

```
## [1] 0.5
```

```
popsd(box)
```

```
## [1] 0.5
```

```
10000*mean(box)
```

```
## [1] 5000
```

```
sqrt(10000)*popsd(box)
```

```
## [1] 50
```

Case 2: Mean of draws from a box model

As the **Mean** of the Sample is just the the **Sum** of the Sample divided by the number of the draws, we get an equivalent result as follows.

Mean of draws from a box model

For the **Mean** of random draws from a box model with replacement,

$$\text{observed value} = \text{expected value} + \text{chance error}$$

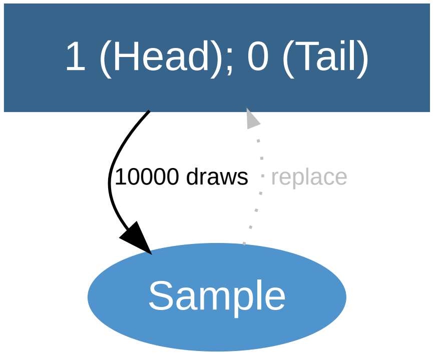
where:

$$\text{expected value (EV)} = \text{mean of the box}$$

$$\text{standard error (SE)} = \frac{\text{SD of the box}}{\sqrt{\text{number of draws}}}$$

Example: WWII Coin Tossing

Step1: Draw the box model



Step2: Calculate the mean and SD of the box

- The mean of the box is $\frac{0+1}{2} = 0.5$.
- The SD of the box is 0.5.

Step3: Calculate the EV and SE of the Mean of the Sample

- The EV of the Mean of the draws is 0.5.
- The SE of the Mean of the draws is $\frac{0.5}{\sqrt{10000}} = 0.005$.

Step4: Conclusion

- We would expect a Sample Mean of 0.5 (EV) with SE 0.005.

The Normal Curve: Modelling the Sum/Mean of a Sample

The Normal Curve in the box model

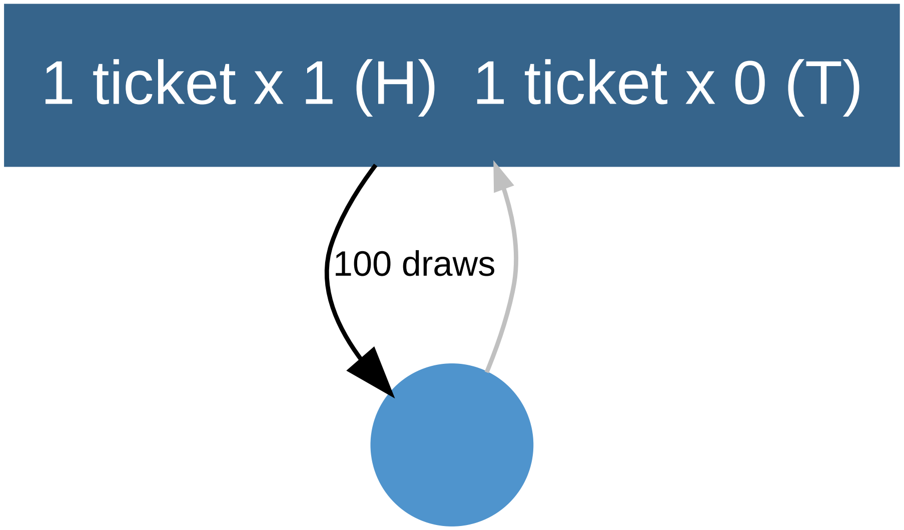
- For large amounts of draws from the box, the observed value of the Sum/Mean often follows the Normal curve.
- We will learn about the Central Limit Theorem (CLT) in the next Topic.
- Here we will just illustrate the main idea, which is that **given a box model we can work out EV and SE**, and then use them to model the Sum/Mean by the **Normal**!

Example (based on WWII)

Example1 (coin tossed 100 times)

A coin is rolled 100 times. What is the chance of getting between 40 and 60 heads?

Step1: Draw the box model



Step2: Calculate the mean and SD of the box

- The mean is $\frac{1+0}{2} = 0.5$.
- The SD is $(1-0)\sqrt{1/2 \times 1/2} = 0.5$.

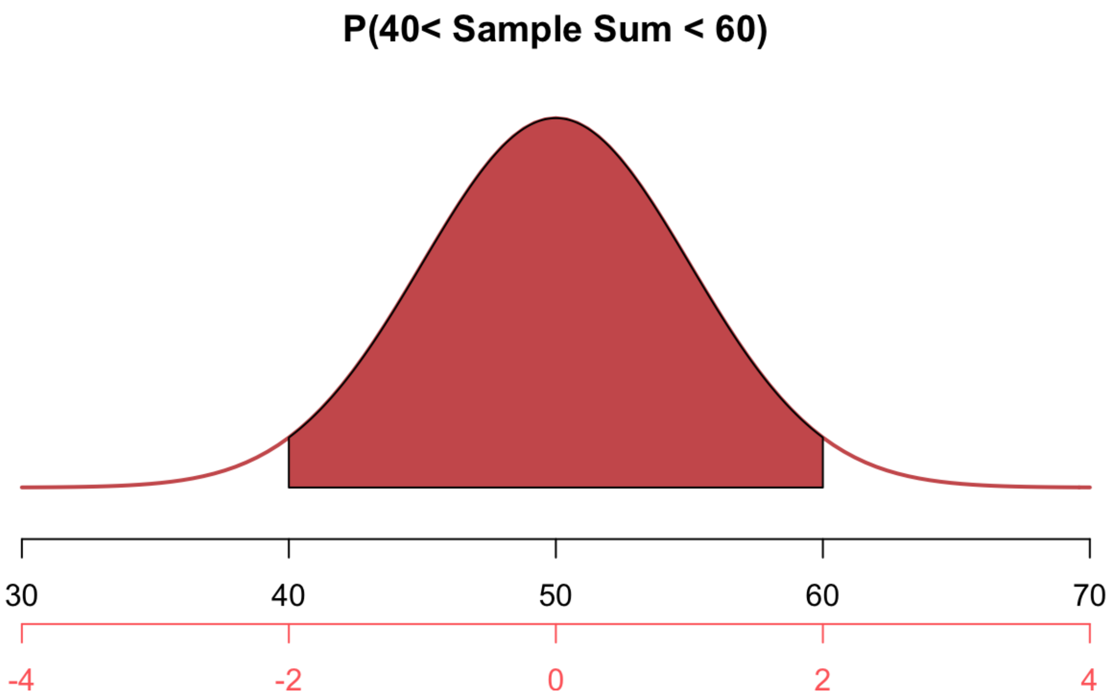
Step3: Calculate the EV and SE of the Sum of the Sample

- The EV of the Sum of the draws is $100 \times 0.5 = 50$.
- The SE of the Sum of the draws is $\sqrt{100} \times 0.5 = 5$.

Step4: Conclusion

- We would expect a Sample Sum of 50 (EV) with SE 5.
- So now model the Sample Sum by a Normal, with mean = 50 and SD = 5. ie Sample Sum ~ N(50, 5^2).

Step5: Draw the Normal curve



Step6: Calculate the chance

- The x values are 40 and 60.
- The z scores are $\frac{40-50}{5} = -\mathbf{2}$ and $\frac{60-50}{5} = \mathbf{2}$.
- So we expect the sum to be between 40 and 60, which is about 95% of the time.

In R

```
box=c(1,0)

100*mean(box)
```

```
## [1] 50
```

```
sqrt(100)*popsd(box)
```

```
## [1] 5
```

```
pnorm(2)-pnorm(-2)
```

```
## [1] 0.9544997
```

```
pbinom(60,100,0.5)-pbinom(40,100,0.5)
```

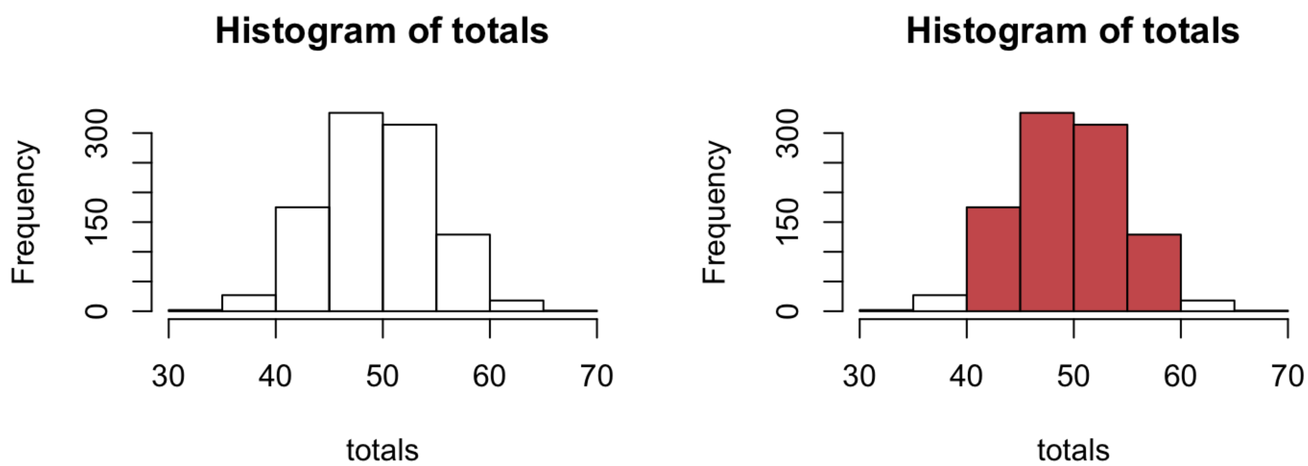
```
## [1] 0.9539559
```

Simulation

```
set.seed(1)
box=c(0,1)
totals = replicate(1000, sum(sample(box, 100, rep = T)))
table(totals)
```

```
## totals
## 32 34 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59
##  1  1  1  3  2  6 15 16 27 29 51 52 58 64 64 80 68 71 59 72 55 57 54 24 29 17
## 60 61 62 63 64 65 68
##  5  7  4  4  2  1  1
```

```
par(mfrow=c(1,2))
h=hist(totals)
cuts<- cut(h$breaks, c(38, 58))
plot(h, col="indianred"[cuts])
```



Using the box model for classifying and counting

Using the box model for classifying and counting

Often we are just interested in 1 particular “ticket” in the box (binary box), which may summarise other tickets.

Example: Toss a dice 100 times and count the number of 6s. The box would have a 1 (representing “6”) and 5 x 0 (representing non “6”).

Box model for classifying (binary box)

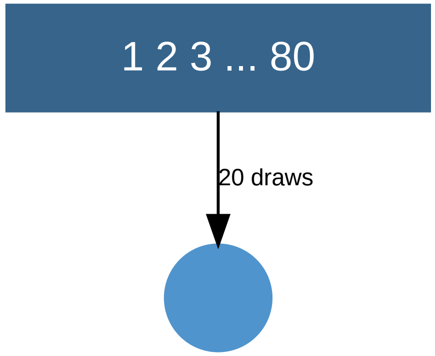
- If you want to classify and count the draws from a box model:
  - mark 1 on the tickets you are counting;
  - mark 0 on the others.

More Examples

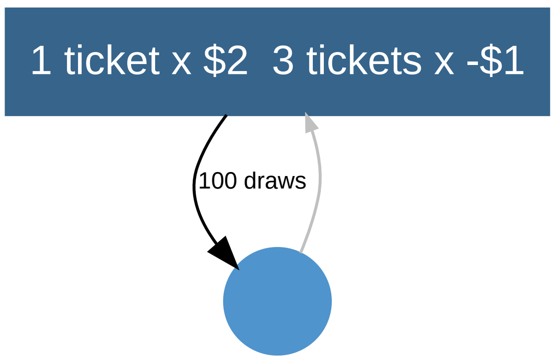
Example2 (Keno Classic)

- In Keno Classic, there are 80 balls numbered 1 to 80. 20 balls are chosen at random without replacement.
- You pick 1 single number from the 80, and *win* if your number is equal to one of the 20 chosen numbers.
  - If you win, you get your dollar back plus $2.
  - If you lose, the house keeps your dollar.
- If you play 100 times, how much would you expect to win/lose?
- How often would you lose more than $20? (Assume a Normal curve).

Preliminary set up (draw 20 numbers from the 80, without replacement)



Step1: Draw the box model (for game)



Step2: Calculate the mean and SD of the box

- The mean of the box is $\frac{2-1-1-1}{4} = -0.25$.
- The SD of the box is $(2-(-1))\sqrt{1/4 \times 3/4} = 1.299038$.

Step3: Calculate the EV and SE of the Sum of the Sample

- The EV of the Sum of the draws is $100 \times -0.25 = -25$.
- The SE of the sum of draws is $\sqrt{100} \times 1.299038 = 12.99$.

Step4: Conclusion

- In 100 plays of Classic Keno we expect to lose $25 (EV) with a SE of $13.
- Hence, it would be very common to lose between $12 and $38.

In R

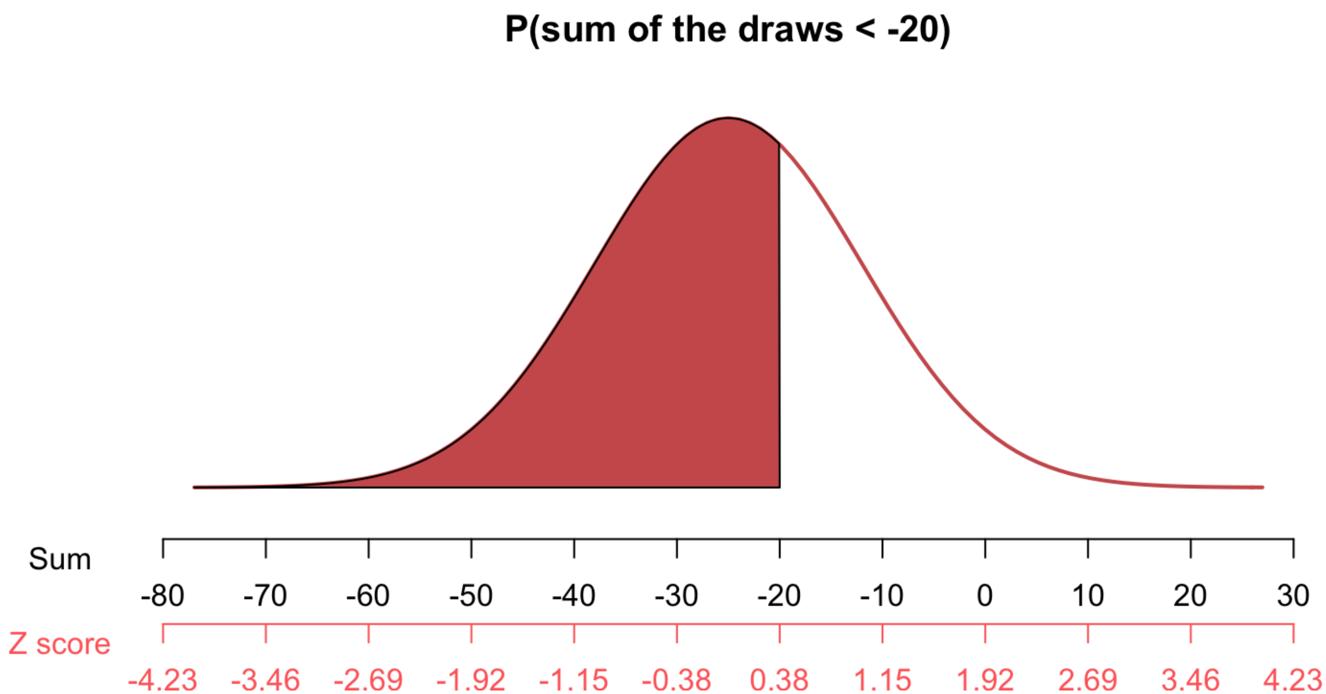
```
box=c(2,-1,-1,-1)
n=100

n*mean(box)
```

```
## [1] -25
```

```
sqrt(n)*popsd(box)
```

```
## [1] 12.99038
```

Step5: Draw the Normal curve using EV and SE.



Step6: Work out the chance

- The Normal curve has EV -25 and SE 12.99 (from previous working).
- The x value is -20.
- The z score is $\frac{-20-(-25)}{12.99} = 0.3849$.
- So we expect the loss to be $20 or more around 0.65 of the time.

In R

```
ev=100*mean(box)
se=sqrt(100)*popsd(box)
pnorm(-20,ev,se)
```

```
## [1] 0.6498443
```

```
pnorm(0.3849)
```

```
## [1] 0.6498442
```

In R: Simulation (size 20)

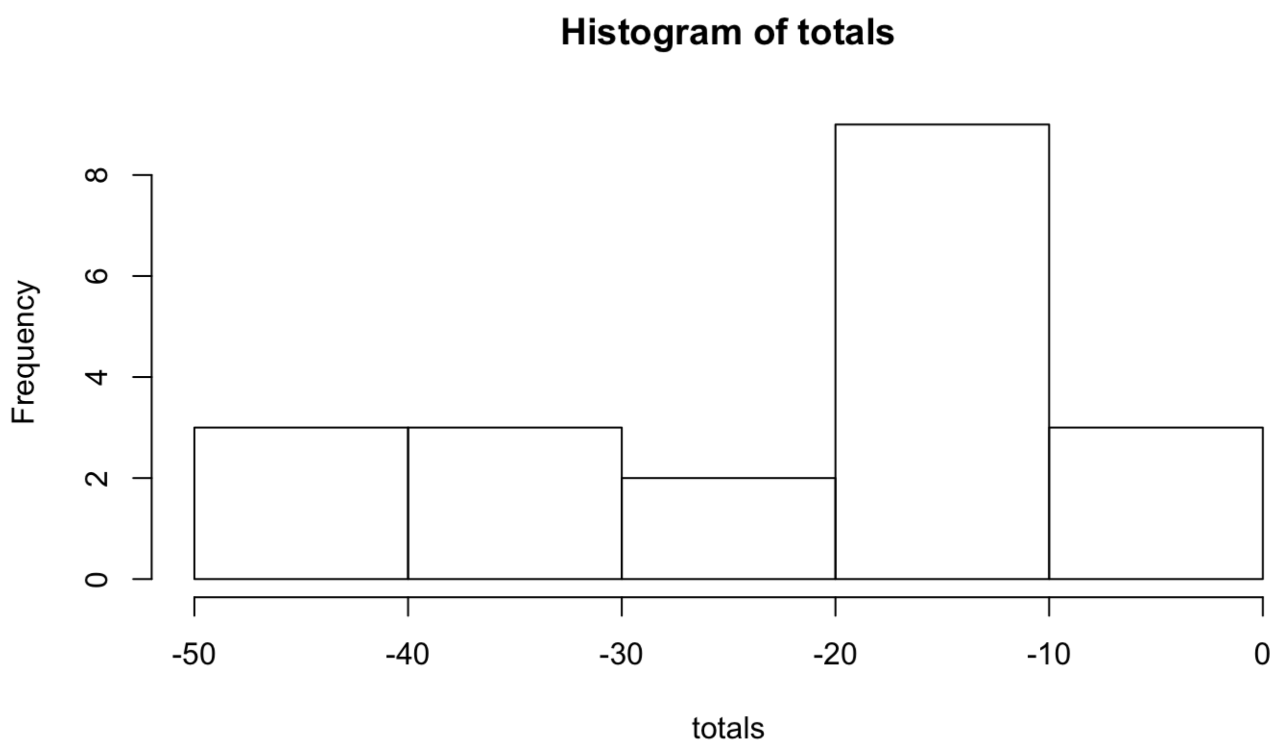
```
set.seed(1)
totals = replicate(20, sum(sample(box, 100, rep = T)))
table(totals)
```

```
## totals
## -49 -46 -40 -37 -34 -28 -25 -19 -16 -13 -10  -7  -4
##   1   1   1   1   2   1   1   3   2   2   2   2   1
```

```
length(totals[totals>=-38 & totals<=-12])/20
```

```
## [1] 0.6
```

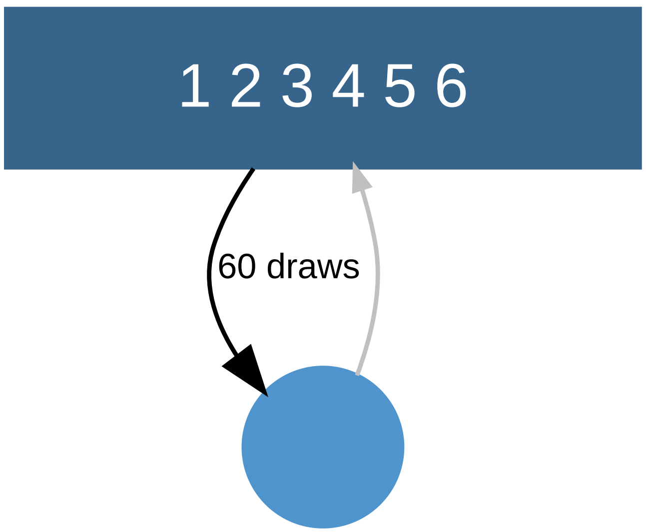
```
hist(totals)
```



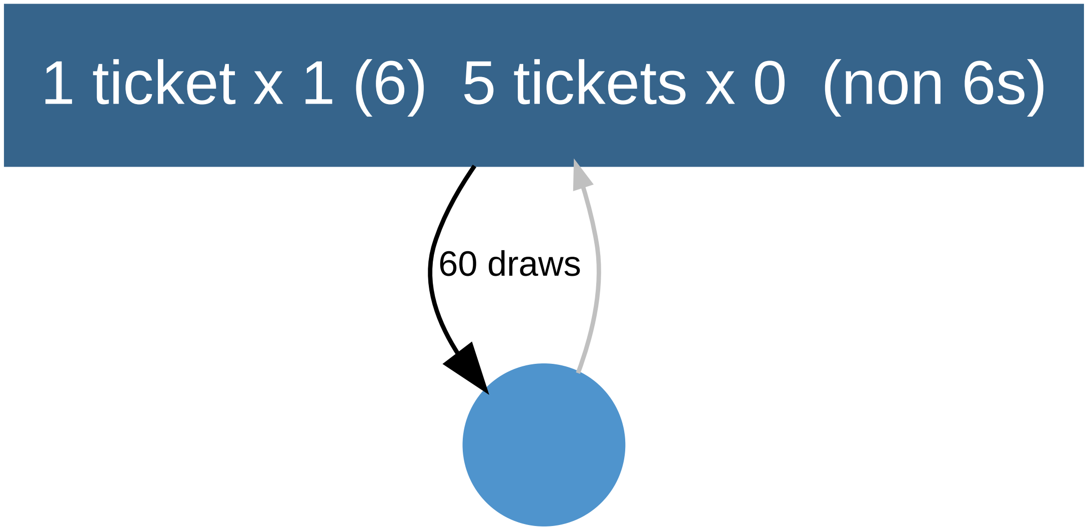
Example3 (die rolled 60 times)

A die is rolled 60 times. How many 6's do we expect?

Step0: Draw the full box model



Step1: Draw the (binary) box model (with 1="6", 0="non 6")



Step2: Calculate the mean and SD of the box

- The mean of the box is $\frac{1+0+0+0+0+0}{6} = 1/6$.
- The SD of the box is $(1-0)\sqrt{5/6 \times 1/6} = 0.372678$.

Step3: Calculate the EV and SD of the Sum of the Sample

- The EV is $60 \times 1/6 = 10$.
- The SE is $\sqrt{60} \times 0.372678 \approx 2.89$.

Step4: Conclusion

Hence, in 60 plays we would expect to see 10 6's with a SE of around 3.

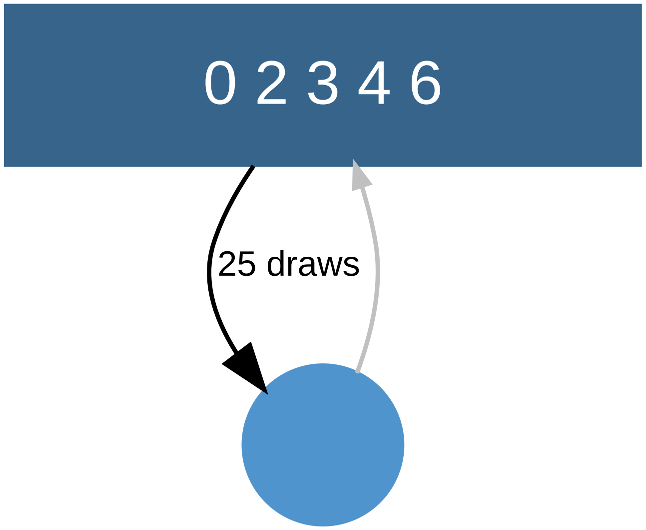
```
box=c(1,0,0,0,0,0)
c(60*mean(box),sqrt(60)*popsd(box))
```

```
## [1] 10.000000  2.886751
```

Example4 (Box 0,2,3,4,6)

- A box contains the numbers 0,2,3,4,6.
- You pick 1 single number, note what it is, and then replace it in the box.
- If you play 25 times, adding each number you draw, what would your expected sum be?
- If you play 25 times, how often is the sum between 50 and 100? (Assume a Normal curve).

Step1: Draw the box model



Step2: Calculate the mean and SD of the box

- The mean of the box is $\frac{0+2+3+4+6}{5} = 3$.
- The SD of the box is $\sqrt{\frac{(0-3)^2+(2-3)^2+\ldots(6-3)^2}{5}} = 2$.

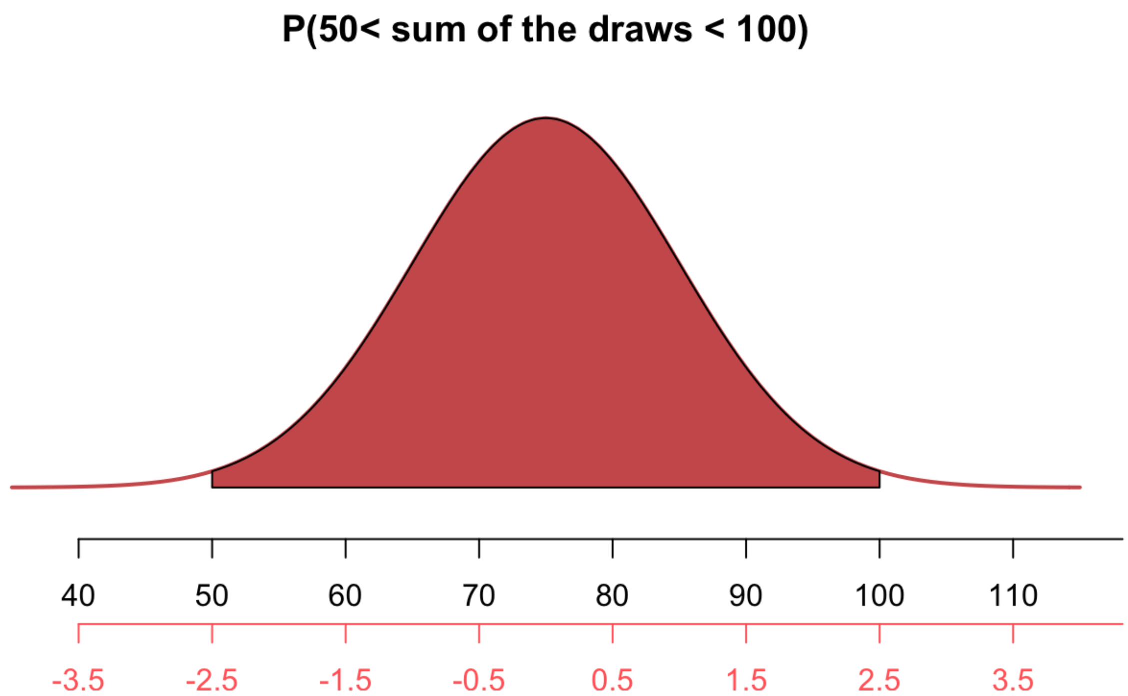
Step3: Calculate the EV and SE of the Sum of the Sample

- The EV is $25 \times 3 = 75$.
- The SE is $\sqrt{25} \times 2 = 10$.

Step4: Conclusion

- In 25 plays we expect to get a sum of 75 (EV) with a SE of 10.
- Hence, it would be very common to get a sum between 65 and 85.

Step5: Draw the Normal curve



Step6: Work out the chance

- The Normal curve has EV 75 and SE 10.
- The x values are 50 and 100.
- The z scores are $\frac{50-75}{10} = -2.5$ and $\frac{100-75}{10} = 2.5$.
- So we expect the sum to be between 50 and 100 almost 99% of the time.

```r
pnorm(2.5)-pnorm(-2.5)
```

```
## [1] 0.9875807
```

In R

```
box=c(0,2,3,4,6)
n=25
n*mean(box)
```

```
## [1] 75
```

```
sqrt(n)*popsd(box)
```

```
## [1] 10
```

Simulation (size 30)

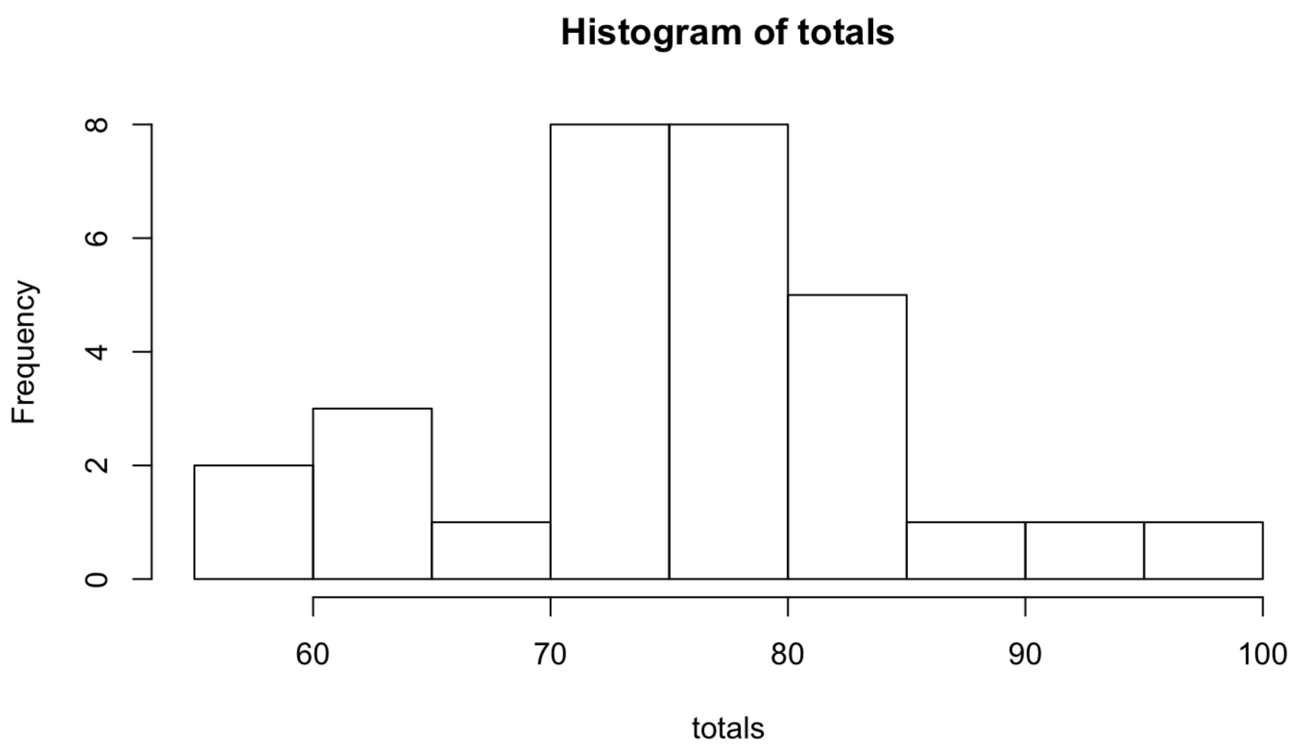
```
set.seed(1)
totals = replicate(30, sum(sample(box, 25, rep = T)))
table(totals)
```

```
## totals
##  59  62  63  68  71  72  73  74  76  78  79  80  81  82  83  84  89  94 100
##   2   2   1   1   2   1   4   1   3   2   2   1   1   1   2   1   1   1   1
```

```
length(totals[totals>=65 & totals<=85])/30
```

```
## [1] 0.7333333
```

```
hist(totals)
```



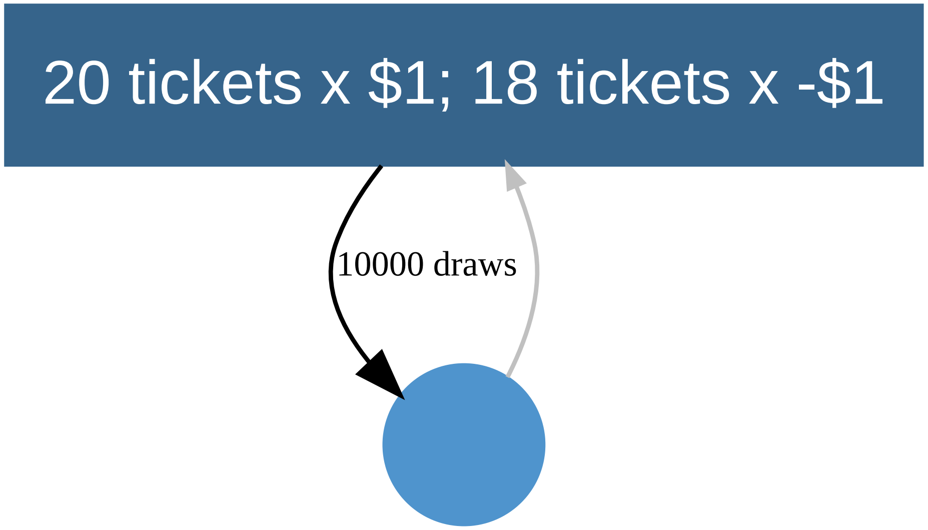
Example5 (Star Casino 00 Roulette on red)

- The Star Casino 00 roulette wheel has 38 pockets,numbered 0 (green), 00 (green) and 1-36 (alternate red and black).
- Suppose players only stake $1 on red at each play.

```
- If they win, they get $1, plus the $1 back.
- If they lose, they lose the $1.
```

- If 10,000 different players have a go over a month, what is the house's expected gain?

Note: Keno Classic is a form of Roulette. We model the box based on the "house" - so "green" and "black" is a win for the house (which is 20 pockets).




```
box=c(rep(1,20),rep(-1,18))
n=10000
n*mean(box)
```

```
## [1] 526.3158
```

```
sqrt(n)*popsd(box)
```

```
## [1] 99.8614
```

Summary

- The sum of n draws from a box model results in a Sample size n.
- We can describe the behaviour of the Sum and Mean of the Sample in terms of the expected value (EV) and standard error (SE), and compare to the observed value (OV).
- We can find SD_{box} by using `popsd()`.
- Given the mean and SD of the population:

	EV	SE
Sum of the Sample	n mean	$\sqrt{n}$ SD
Mean of the Sample	mean	SD / $\sqrt{n}$

- To focus on counting 1 type ticket in the Sample, make that ticket “1” and all the other tickets “0”.

Key Words

chance error, expected value, standard error