STAT5002 Lab6 Worksheet

Introduction to Statistics

STAT5002

1 More box model: winnings after of 25 rolls

Scenario: in a game, you roll a fair die once, and if it lands on "1", you win \$4, and otherwise, you lose \$1. Suppose you play this game 25 times. Let $X_1, \dots, X_n$ denote your winnings each time (with $n = 25$) and $S = X_1 + \cdots + X_n$ your total winnings after $n = 25$ repetitions of the game.

1.1 Draw the box model.

```
|                        |
| -4  -1  -1  -1  -1  -1 |
|                        |
+------------------------+
```

1.2 Calculate the mean and SD of the box.

Try to work out the calculations "by hand", and then verify your results in R.

Solution: The box has

- sum $= -1$
- mean $\mu = -\frac{1}{6}$
- sum of squares $= 5 + 16 = 21$
- mean square $= \frac{21}{6} = \frac{7}{2}$
- SD $\sigma = \sqrt{\frac{7}{2} - \left(-\frac{1}{6}\right)^2} = \sqrt{\frac{7\times 18 - 1}{36}} = \frac{\sqrt{125}}{6} = \frac{\sqrt{5\times 25}}{6} = \frac{5\sqrt{5}}{6}$.

```r
## write code here
box = c(rep(-1, 5), rep(4, 1))  # equivalent to c(-1,-1,-1,-1,-1, 4)
mu = mean(box)
mu
```

```
[1] -0.1666667
```

```
sig = sqrt(mean(box^2) - mu^2)
sig
```

```
[1] 1.86339
```

```
5 * sqrt(5)/6
```

```
[1] 1.86339
```

1.3 Calculate the $E(S)$ and $SE(S)$, which are the expected value and SE of your total winnings after 25 repetitions of the game.

Try to work out the calculations "by hand", and then verify your results in R.

Solution: We take $n = 25$, then we get

- $E(S) = n\mu = -\frac{25}{6}$
- $SE(S) = \sigma\sqrt{n} = \frac{5\sqrt{5}}{6} \times 5 = \frac{25\sqrt{5}}{6}$

```
## write code here
n = 25
E.S = mu * n
E.S
```

```
[1] -4.166667
```

```
SE.S = sig * sqrt(n)
SE.S
```

```
[1] 9.31695
```

```
25 * sqrt(5)/6
```

```
[1] 9.31695
```

1.4 Using the appropriate normal approximation (by the Central Limit Theorem)

- Estimate the chance that you will break even, that is, not lose money.
- Estimate the chance that you will win more than $10.
- Estimate the chance that you will lose more than $10

Hint: you can first work out the mean and SD of the normal curve, and then you will need to use R for calculating the chance, and it may help to draw a diagram (by hand).

Solution: The (BIG) box of all possible sums (after 25 repetitions of the game) has

- mean equal to $E(S) = -\frac{25}{6} \approx -4.16667$
- SD equal to $SE(S) = \frac{25\sqrt{5}}{6} \approx 9.317$
- by the Central Limit Theorem an (approximately) normal shape (so long as $n = 25$ is "large enough"!)

Estimate the chance that you will break even

We are asked to approximate $P(S > 0)$. The value 0 is (in standard units)

$$\frac{0 - 25\mu}{5\sigma} = \frac{-\left(\frac{25}{6}\right)}{\frac{25\sqrt{5}}{6}} = \frac{1}{\sqrt{5}} \approx 0.447\,.$$

Using R:

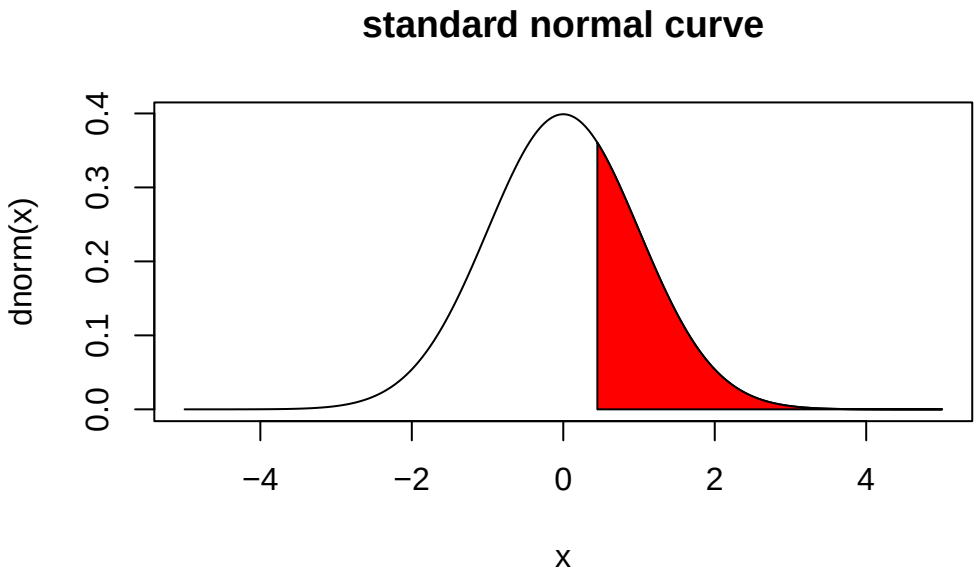
```
## write code here
z = (0 - E.S)/SE.S
z
```

```
[1] 0.4472136
```

```
1/sqrt(5)
```

```
[1] 0.4472136
```

That is, the value 0 is 0.447 SDs above the mean (in the box of all possible sums). The chance is then *approximated* by the area under the *standard* normal curve to the *right* of 0.447:



This is given by

```
1 - pnorm(1/sqrt(5))
```

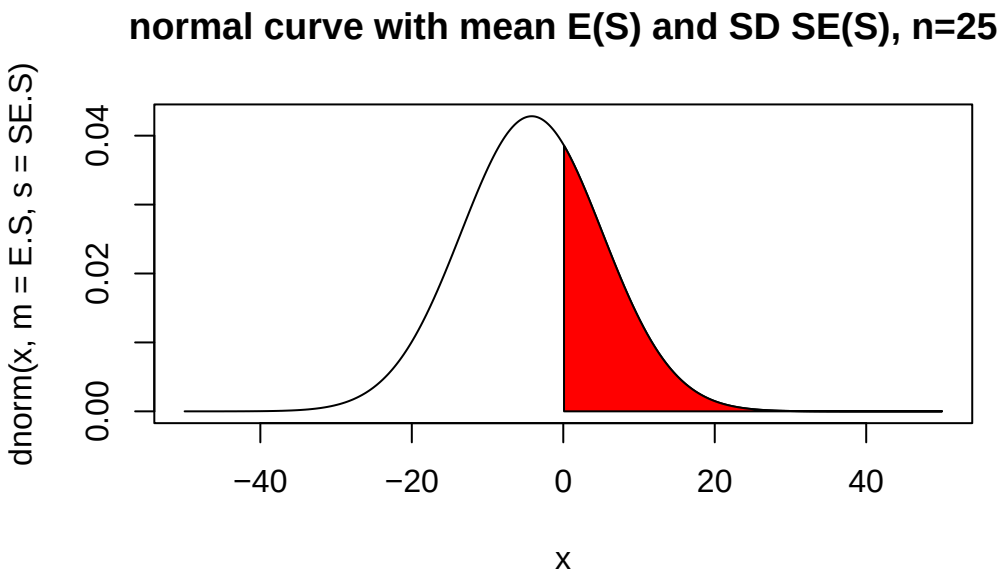
```
[1] 0.3273604
```

or equivalently

```
pnorm(1/sqrt(5), lower.tail = F)
```

```
[1] 0.3273604
```

Note that we can also think on the original scale, and rather than comparing to the *standard* normal curve, use instead the normal curve with the same mean and SD as the box of all possible sums:



The normal approximation can then be obtained using just

```
1 - pnorm(0, m = E.S, s = SE.S)
```

```
[1] 0.3273604
```

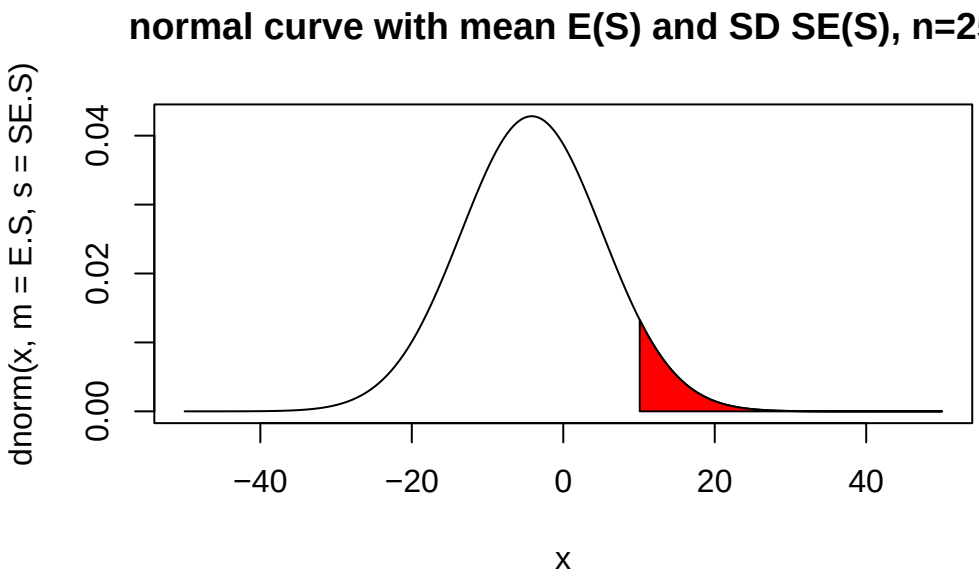
or

```
pnorm(0, m = E.S, s = SE.S, lower.tail = F)
```

```
[1] 0.3273604
```

Estimate the chance that you will win more than $10.

Thinking on the original scale, we have now



The normal approximation is then simply given by

```
1 - pnorm(10, m = E.S, s = SE.S)
```

```
[1] 0.06418939
```

Note: as a "sanity check", we see that the value 10 is

$$\frac{10 - \left(-\frac{25}{6}\right)}{\frac{25\sqrt{5}}{6}} = \frac{85}{25\sqrt{5}} = \frac{17}{5\sqrt{5}} \approx 1.52$$

in standard units (i.e. approx. 1.52 SEs above the expectation), so we should get the same answer using

```
z = (10 - E.S)/SE.S
z
```

```
[1] 1.520526
```

```
17/(5 * sqrt(5))
```

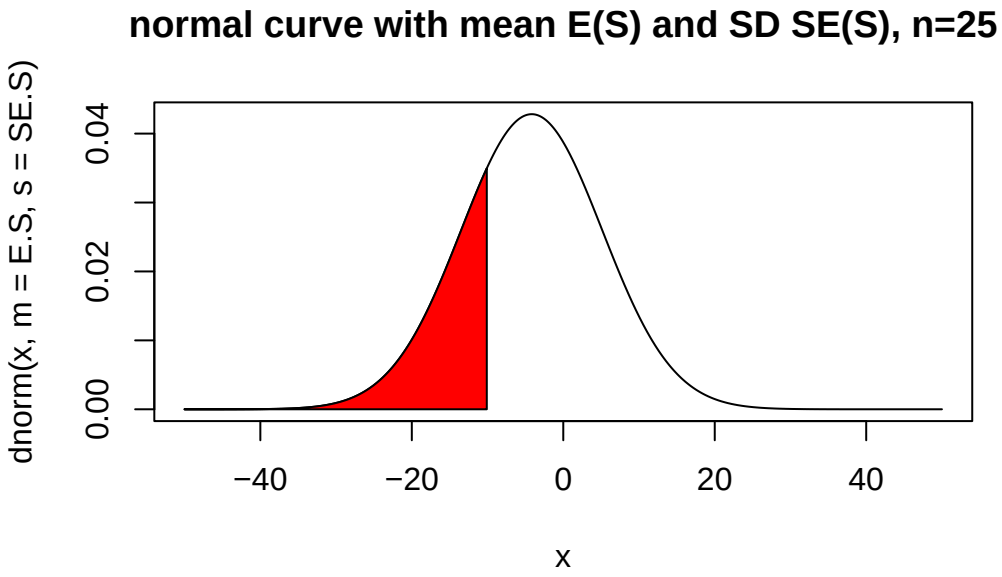
```
[1] 1.520526
```

```r
1 - pnorm(z)
```

```
[1] 0.06418939
```

Estimate the chance that you will lose more than $10.

Using the same method as above, this looks like



The normal approximation is then

```r
## write code here
pnorm(-10, m = E.S, s = SE.S)
```

```
[1] 0.265625
```

1.5 Compare these approximations to "true" probabilities using a simulation:

- simulate a sum of 25 games 1000 (or more) times
- determine the proportion of simulations where the sum is
  - positive
  - more than $10
  - less than -$10

Code to obtain 1000 of the sums in one line is

- `sums=replicate(1000, sum(sample(box, 25, replace=T)))`.

You can also compare the shape of the histogram to the normal curve with mean $E(S)$ and SE $SE(S)$ with something like

```
hist(sums, pr = T)
curve(dnorm(x, m = E.S, s = SE.S), add = T, lty = 2)
```

What do you notice?

```
sums = replicate(10000, sum(sample(box, 25, replace = T)))
mean(sums > 0)
```

```
[1] 0.2283
```

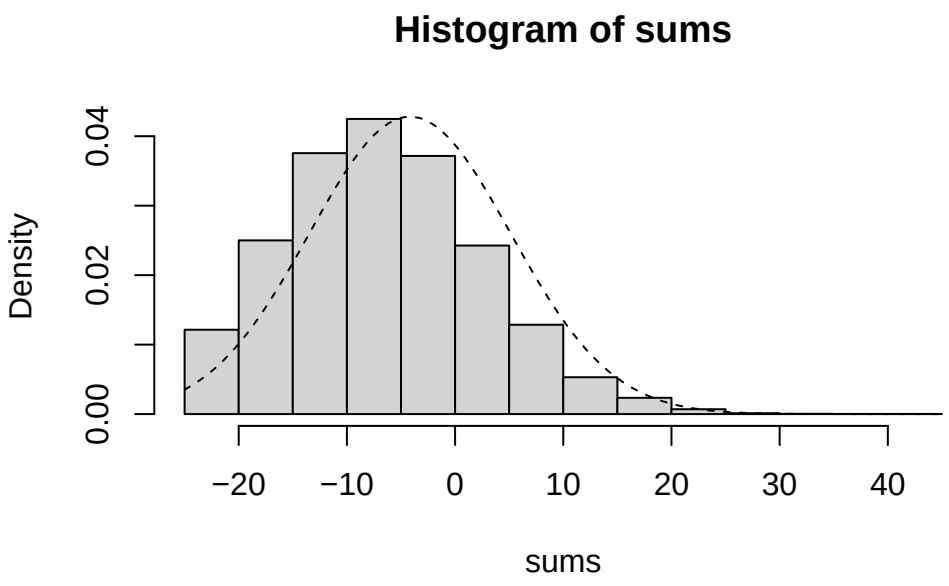
```
mean(sums > 10)
```

```
[1] 0.0427
```

```
mean(sums < (-10))
```

```
[1] 0.1857
```

```
hist(sums, n = 20, pr = T)
curve(dnorm(x, m = E.S, s = SE.S), add = T, lty = 2)
```



It is not a very good match! It is quite skewed. Also, the normal approximations to the 3 probabilities are not very close. The upper tail approximations are too big, while the lower tail approximation is too small.

1.6 Winnings after 100 rolls

Now consider the same game, but you play 100 times. First use CLT and normal approximations to carry the following estimations

- Estimate the chance that you will break even, that is, not lose money.
- Estimate the chance that you will win more than \$10.
- Estimate the chance that you will lose more than \$10

Then, use a simulation to assess the quality of these normal approximations, and the shape of the histogram. What do you conclude?

Solution: we can replicate the code above after making the change `n=100`. The important difference is that now the sum S has

$$E(S) = 100\mu = -\frac{100}{6} = -\frac{50}{3} \approx -16.6667$$

and

$$SE(S) = 10\sigma = \frac{50\sqrt{5}}{6} = \frac{25\sqrt{5}}{3} \approx 18.634\,.$$

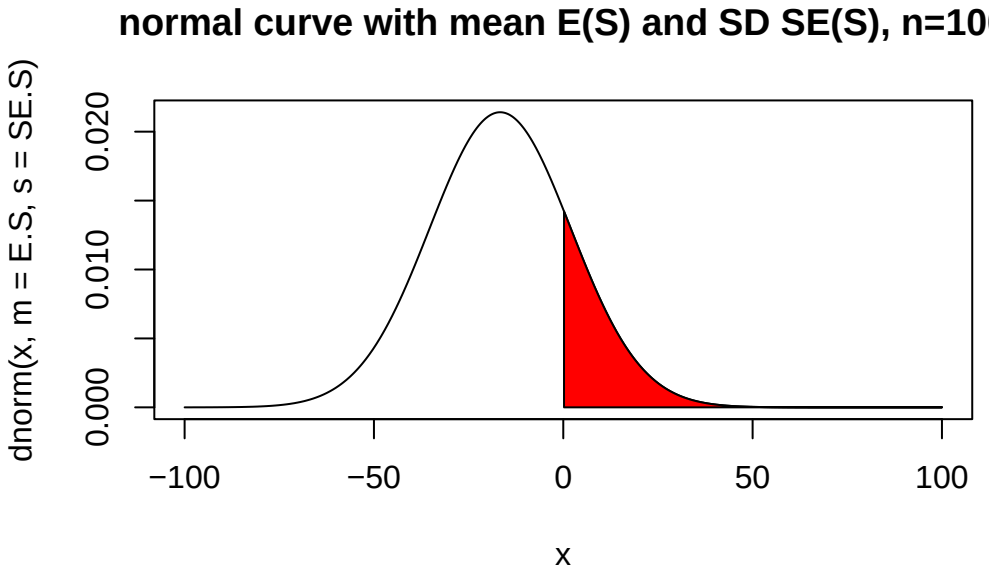
```
n = 100
E.S = n * mu
E.S
```

```
[1] -16.66667
```

```
SE.S = sqrt(n) * sig
SE.S
```

```
[1] 18.6339
```

Estimate the chance that you will break even.

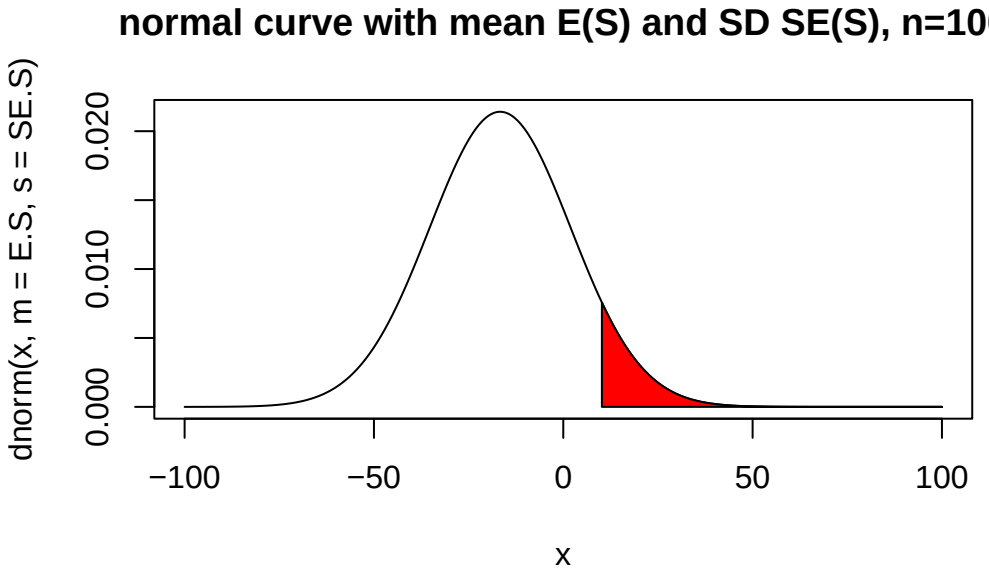



```
## write code here
1 - pnorm(0, m = E.S, s = SE.S)
```

```
[1] 0.1855467
```

Estimate the chance that you will win more than $10.

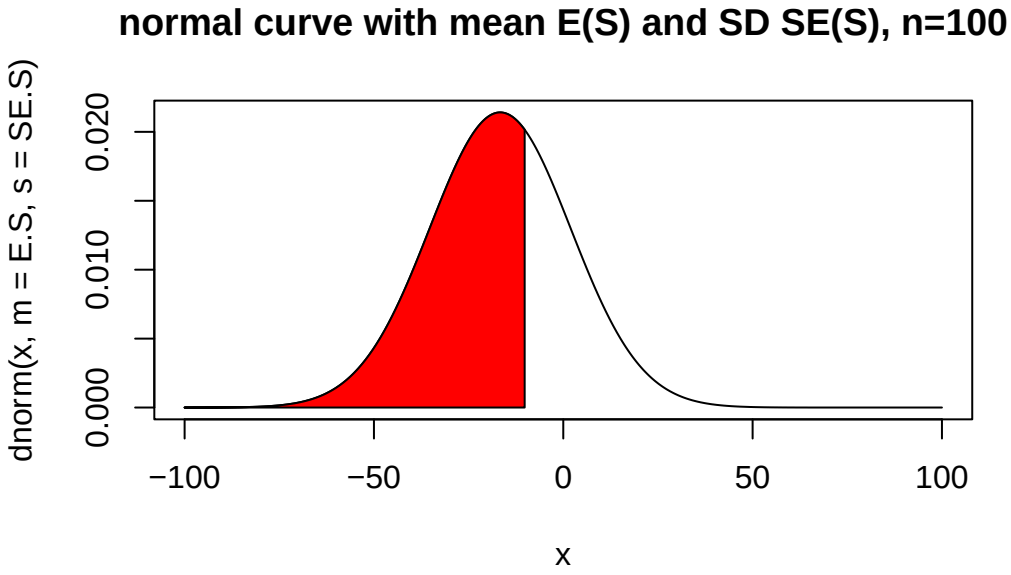
```
## write code here
```



```
## write code here
1 - pnorm(10, m = E.S, s = SE.S)
```

```
[1] 0.07620314
```

Estimate the chance that you will lose more than $10.




```
## write code here
pnorm(-10, m = E.S, s = SE.S)
```

```
[1] 0.6397426
```

Use a simulation to assess the quality of these normal approximations, and the shape of the histogram. What do you conclude?

```
sums = replicate(10000, sum(sample(box, 100, replace = T)))
mean(sums > 0)
```

```
[1] 0.1481
```

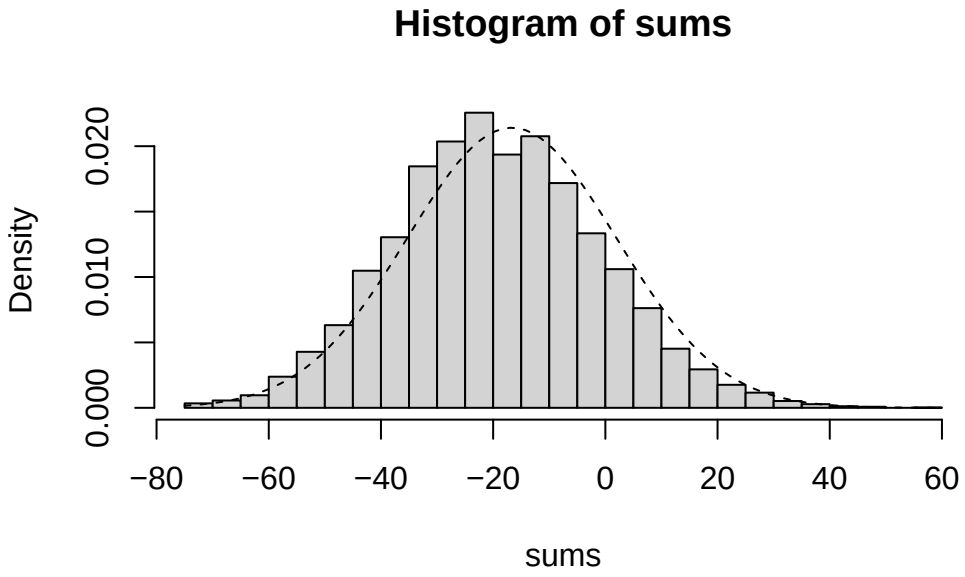
```
mean(sums > 10)
```

```
[1] 0.057
```

```r
mean(sums < (-10))
```

```
[1] 0.5955
```

```r
hist(sums, n = 20, pr = T)
curve(dnorm(x, m = E.S, s = SE.S), add = T, lty = 2)
```



The shape is closer to a normal, and while the approximations are still too small in the upper tail, and too big in the lower tail, they are closer.

2 Simulate the CLT

2.1 Example: Simulate simple box (symmetric)

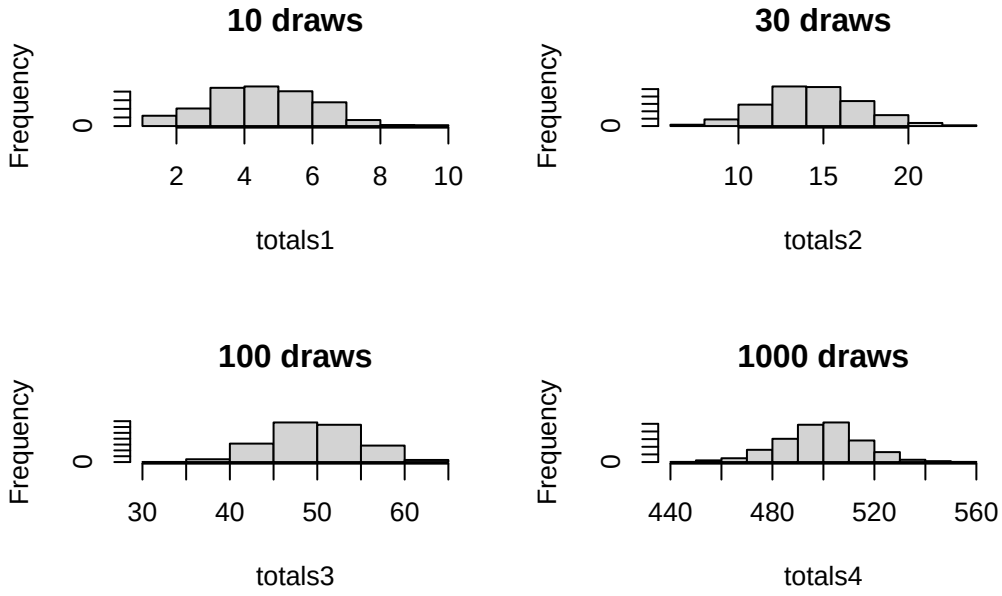
Experiment to find out what minimum size number of draws it takes, for the distribution of the sample sum to start looking like a Normal curve.

Method: Use `replicate` to simulate 1000 samples of 10, 30, 100 and 1000 draws from a box, and compare your results.

- Take draws from 0,1 box

```r
set.seed(1)
box = c(0, 1)
totals1 = replicate(1000, sum(sample(box, 10, rep = T)))
totals2 = replicate(1000, sum(sample(box, 30, rep = T)))
totals3 = replicate(1000, sum(sample(box, 100, rep = T)))
totals4 = replicate(1000, sum(sample(box, 1000, rep = T)))

par(mfrow = c(2, 2))
hist(totals1, main = "10 draws")
hist(totals2, main = "30 draws")
hist(totals3, main = "100 draws")
hist(totals4, main = "1000 draws")
```



How many draws do we need to take before the histogram starts looking like a normal distribution?

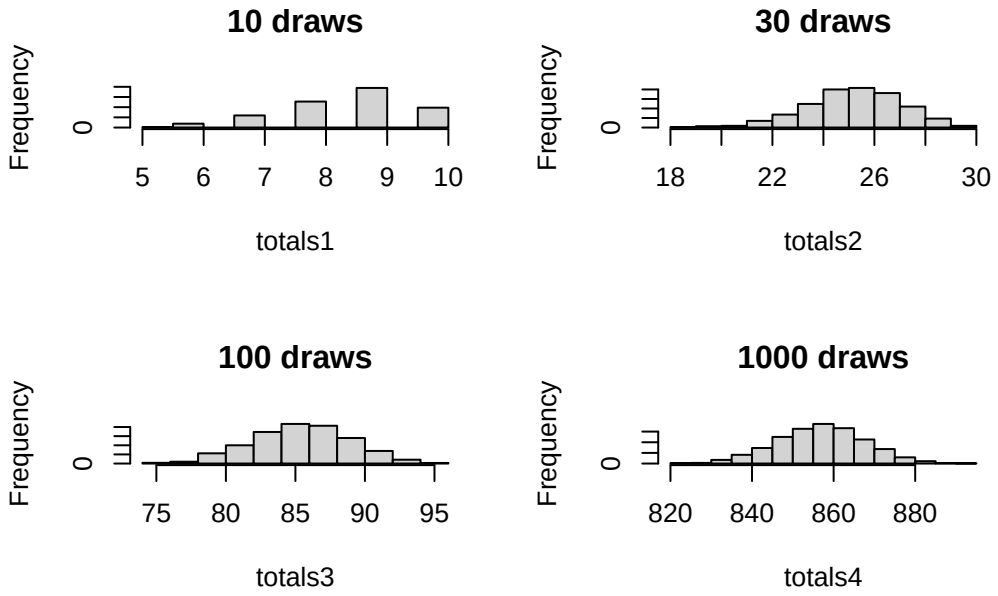
2.2 Simulate simple box (asymmetric)

- Take draws from 0,1,1,1,1,1,1 box

```r
## write code here
box = rep(c(0, 1), c(1, 6))
totals1 = replicate(1000, sum(sample(box, 10, rep = T)))
totals2 = replicate(1000, sum(sample(box, 30, rep = T)))
```

```r
totals3 = replicate(1000, sum(sample(box, 100, rep = T)))
totals4 = replicate(1000, sum(sample(box, 1000, rep = T)))

par(mfrow = c(2, 2))
hist(totals1, main = "10 draws")
hist(totals2, main = "30 draws")
hist(totals3, main = "100 draws")
hist(totals4, main = "1000 draws")
```



How many draws do we need to take before the histogram starts looking like a normal distribution?

2.3 Simulate the box from Q1

Using the box from Q1, experiment to find out what minimum size number of draws it takes, for the distribution of the sample sum to start looking like a Normal curve for your own custom box.

Solution:

```r
# write code here

box = c(-1, -1, -1, -1, -1, 4)
totals1 = replicate(1000, sum(sample(box, 10, rep = T)))
totals2 = replicate(1000, sum(sample(box, 30, rep = T)))
```

```r
totals3 = replicate(1000, sum(sample(box, 100, rep = T)))
totals4 = replicate(1000, sum(sample(box, 1000, rep = T)))

par(mfrow = c(2, 2))
hist(totals1, main = "10 draws")
hist(totals2, main = "30 draws")
hist(totals3, main = "100 draws")
hist(totals4, main = "1000 draws")
```