The Box Model

Sampling Data | Chance Variability

STAT5002

The University of Sydney

Mar 2025

CRICOS 00026A TEQSA PRV12057

Sampling Data

Topic 5: Understanding chance and chance simulation

Topic 6: Chance variability

Topic 7: Central limit theorem

Outline

Box model

Random draws

Sum of random draws

Averages of random draws

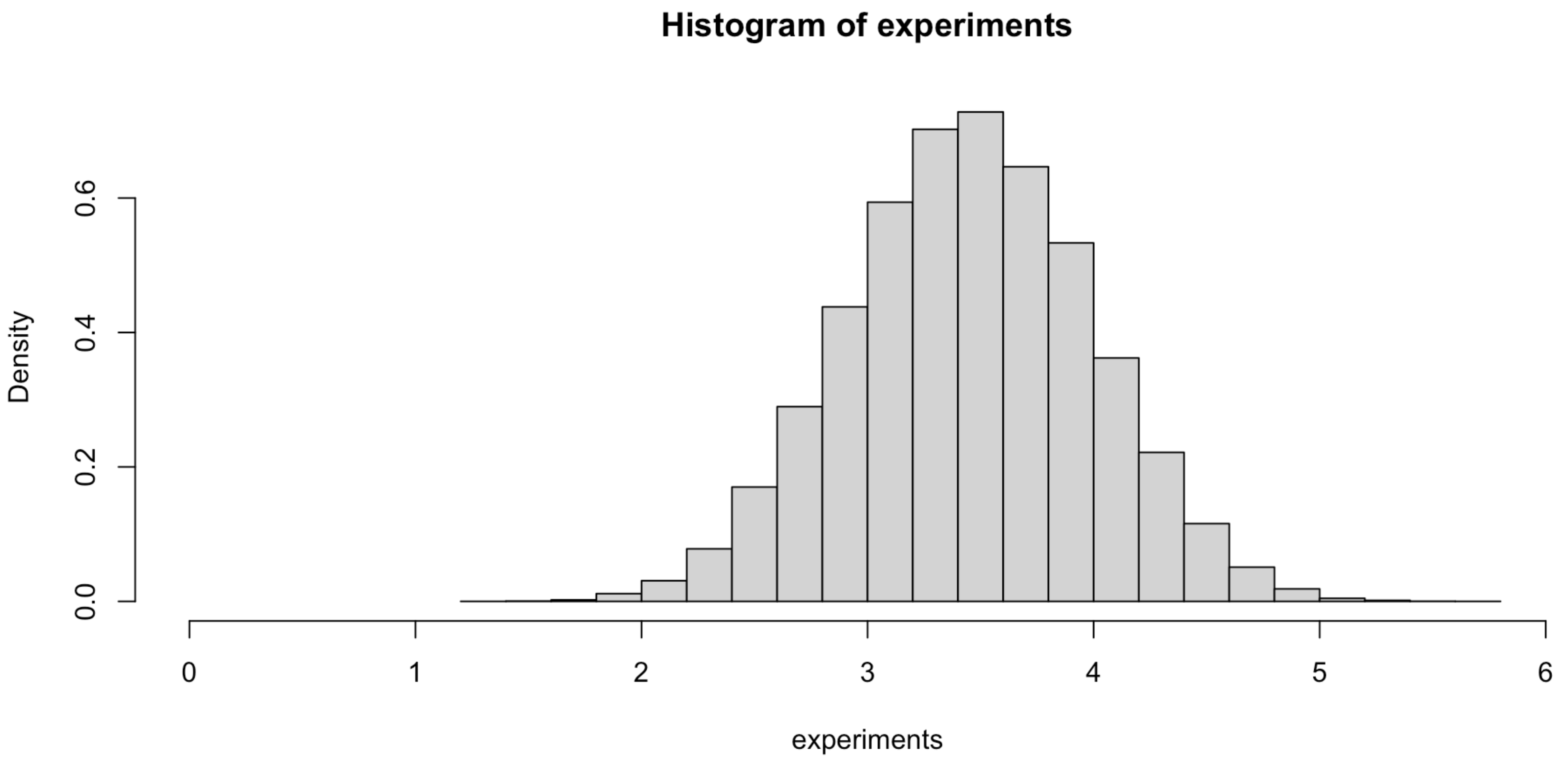
Motivation: average of rolling dice

- Consider a sample consists of rolling a fair 6-sided die n times .
- Take the sample mean - which is the average over the n rolls of the fair die.
- What is the behaviour (e.g., mean, SD) of possible sample means for increasing sample size $n = 10, 100, 1000$?
  - ➟ Simulate in R using 100,000 experiments.

```
rolling = function(n) {
    # rolling n times, sample with replacement
    rolls = sample(1:6, size = n, rep = T)
    # taking the average (mean)
    a = mean(rolls)
    return(a)
}
```

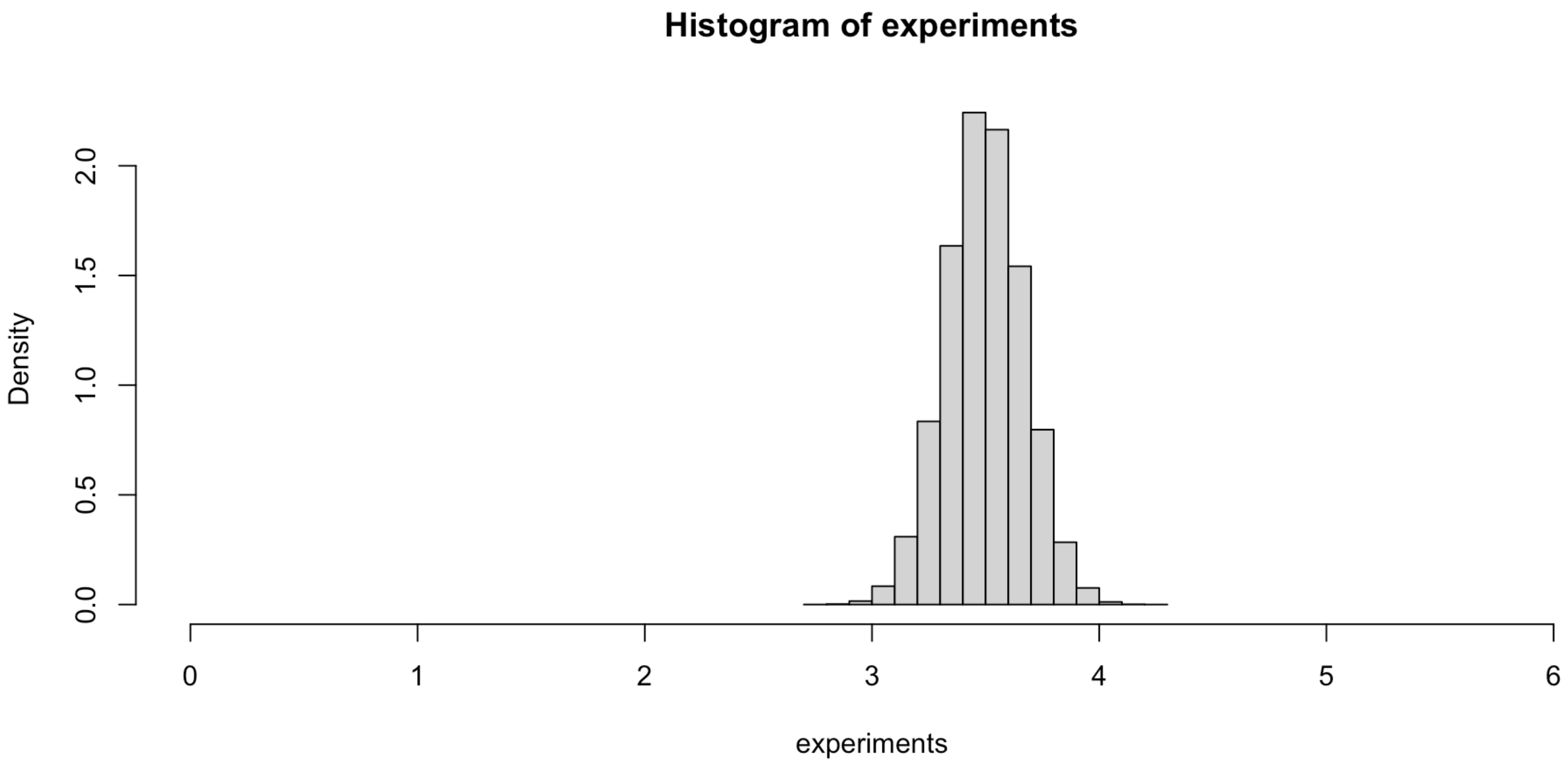
average of $n = 10$ rolls

```
experiments = replicate(1e+05, rolling(10))
hist(experiments, freq = F, xlim = c(0, 6))
```



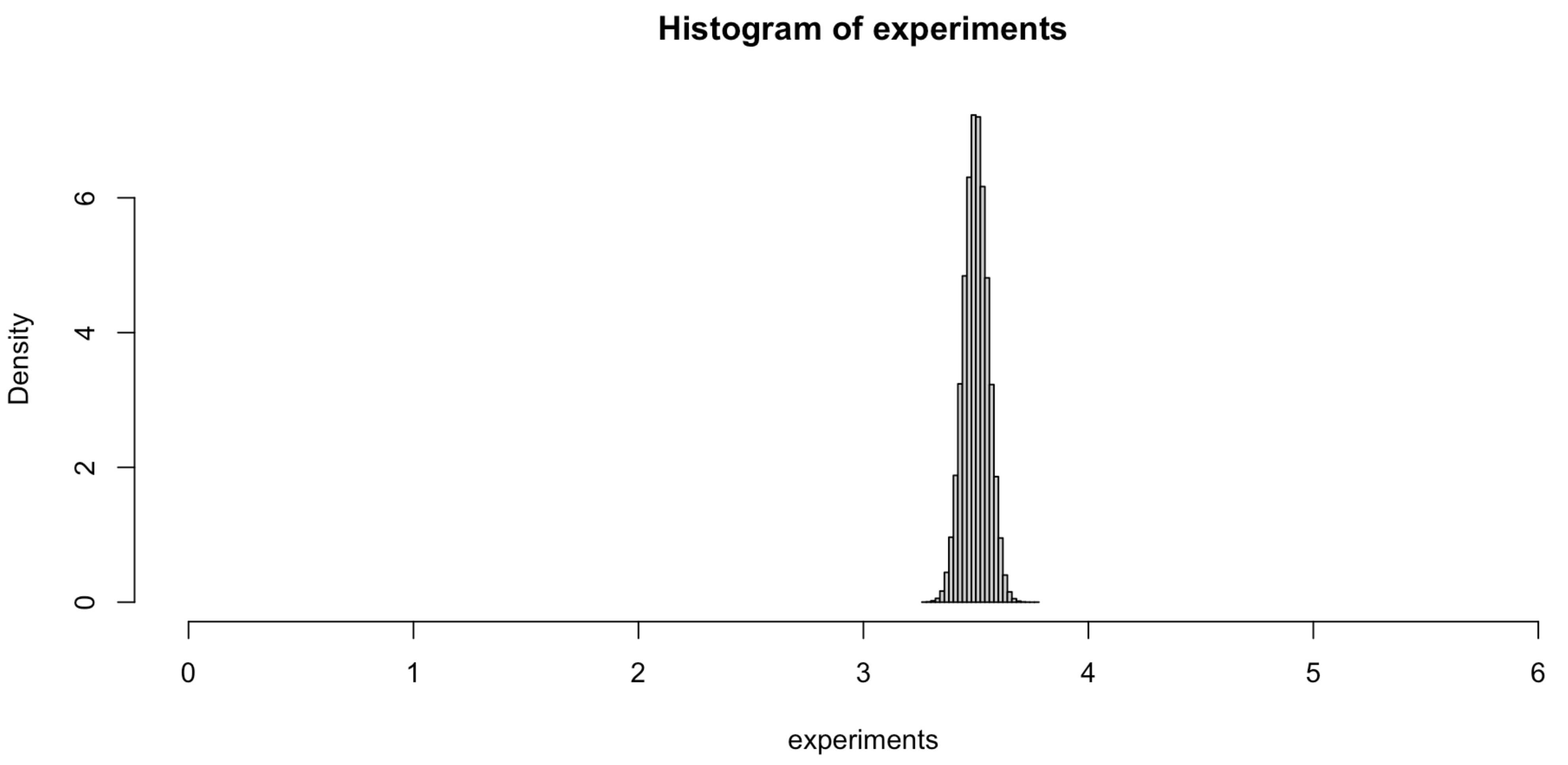
average of $n = 100$ rolls

```r
experiments = replicate(1e+05, rolling(100))
hist(experiments, freq = F, xlim = c(0, 6))
```



average of $n = 1000$ rolls

```r
experiments = replicate(1e+05, rolling(1000))
hist(experiments, freq = F, xlim = c(0, 6))
```



Review: population mean and SD

Given a data $x_1, \dots, x_M$:

- Population mean

$$\bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i$$

- Deviations $D_i = x_i - \bar{x}$.
  - ➟ The mean of deviations is zero, as $\sum_{i=1}^{M} D_i = 0$.
- Population SD (root mean square of deviations)

$$\mathrm{SD}_{pop}(x) = \sqrt{\frac{\sum_{i=1}^{M} D_i^2}{M}} = \sqrt{\frac{\sum_{i=1}^{M}(x_i - \overline{x})^2}{M}}$$

Population mean and SD: dividing by a constant

Given a data $x_1, \ldots, x_M$, we create a new data $y_1, \ldots, y_M$ such that $y_i = \frac{x_i}{b}$ for some $b \neq 0$. What are the population mean and SD of y?

- Population mean

$$\bar{y} = \frac{1}{M}\sum_{i=1}^{M} y_i = \frac{1}{M}\sum_{i=1}^{M} \frac{x_i}{b} = \frac{1}{b}\left(\frac{1}{M}\sum_{i=1}^{M} x_i\right) = \frac{1}{b}\bar{x}$$

- Population SD

$$\mathrm{SD}_{pop}(y) = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - \overline{y})^2} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\frac{x_i - \overline{x}}{b}\right)^2} = \frac{1}{b}\sqrt{\frac{1}{M}\sum_{i=1}^{M}(x_i - \overline{x})^2} = \frac{1}{b}\mathrm{SD}_{pop}(x)$$

Computing formula for population SD

- For a list of numbers $x_1, x_2, \ldots, x_M$, the square of the SD may be written as

$$SD^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i - \bar{x})^2 = \left(\frac{1}{M}\sum_{i=1}^{M} x_i^2\right) - \bar{x}^2$$

  the "mean square minus the square of the mean".

- To see why, recall that $\sum_{i=1}^{M} x_i = M\bar{x}$ and so:

$$\begin{aligned}
\sum_{i=1}^{M}(x_i - \bar{x})^2 &= (x_1^2 - 2\bar{x}x_1 + \bar{x}^2) + \cdots + (x_M^2 - 2\bar{x}x_M + \bar{x}^2) \\
&= (x_1^2 + \cdots + x_M^2) - 2\bar{x}(x_1 + \cdots + x_M) + \underbrace{\bar{x}^2 + \cdots + \bar{x}^2}_{M \text{ terms}} \\
&= \sum_{i=1}^{M} x_i^2 - 2\bar{x}M\bar{x} + M\bar{x}^2 = \sum_{i=1}^{M} x_i^2 - M\bar{x}^2
\end{aligned}$$

Easy way to compute population SD in R

- The computing formula above can be used to write a quick-and-easy R function to compute the (population) SD of a list of numbers.

```r
popsd = function(x) {
    pop = sqrt(mean(x^2) - mean(x)^2)
    return(pop)
}
```

- Let's try it out:

```r
x = 1:10
x  # this list has mean 5.5
```

```
 [1]  1  2  3  4  5  6  7  8  9 10
```

```r
mean(x)
```

```
[1] 5.5
```

```r
sqrt(mean((x - 5.5)^2))
```

```
[1] 2.872281
```

```r
popsd(x)
```

```
[1] 2.872281
```

The box model

Statistical models

A **model** is a representation of something which

- Is **simpler** but at the same time captures the **key features** of the original.

Data obtained "in real life" is generated (in general) by quite complicated processes.

Statistical models are models for data-generating processes:

- They are much simpler than the "real" data-generating process but
- (Hopefully) they capture the key features, at least in terms of the **random variability** of the data.

For example, the normal curve is a model.

The box model

- The **box model** is a very simple statistical model for representing a population.
- A collection of N objects, e.g. tickets, balls is imagined "in a box".
  - ➠ For example, here is the box for a die

$$\boxed{\boxed{1}\,\boxed{2}\,\boxed{3}\,\boxed{4}\,\boxed{5}\,\boxed{6}}$$

  - ➠ Each ticket bears a number – let's deal with only numerical data here.
- We can take a **random sample** of a certain size n from the box.
  - ➠ The sampling may be **with** or **without** replacement.
- What does **a random sample is taken** mean exactly?
  - ➠ Consider all possible ways of selecting n objects from the box. A random sample is when each possible of these selection is equally likely.

Random draws

Single random draws (samples of size $n = 1$)

A random draw is a random sample with $n = 1$.

- If a single draw is taken, then each object in the “box” has an equal chance of being picked.
- If we *completely know* the contents of the box, we can write down the chance of each possible value.

We let X denote the **random draw**:

- This represents the “value we might get”
- X can take different values with different probabilities/chances.

The **distribution** of X is a **table** with two “rows”:

- Each possible value x that X can take (note the capitalisation!) *and*
- The corresponding probability/chance of that value.

Simple examples (Box 1)

For example, suppose X is a random draw from the following box (box 1):

$$\boxed{\boxed{1}\ \boxed{2}\ \boxed{3}}$$

There are then three possible tickets: $\boxed{1}$, $\boxed{2}$ and $\boxed{3}$ and each has (equal) chance of $\frac{1}{3}$ of being picked, so:

$$P(X = 1) = P(X = 2) = P(X = 3) = \frac{1}{3}\,.$$

Here we write $P(\cdot)$ to denote the "probability" or "chance" of each event.

The distribution of X is

x	1	2	3
$P(X = x)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Non-equal chances (Box 2)

We can have box models where the different possible *values* are not necessarily equally likely.

For the box (box 2)

$$\boxed{\boxed{1}\ \boxed{2}\ \boxed{2}\ \boxed{3}\ \boxed{3}\ \boxed{3}}$$

if each "ticket" is equally likely, we have

$$P(X=1)=\frac{1}{6}\,,\quad P(X=2)=\frac{2}{6}=\frac{1}{3}\,,\quad P(X=3)=\frac{3}{6}=\frac{1}{2}\,.$$

X then has distribution

x	1	2	3
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{2}$

Larger box example

Consider the box defined by the file `y.dat` in the R code below:

```r
y = scan("y.dat")
y
```

```
  [1]  3  4  5  6  7  8  4  5  6  7  8  9  5  6  7  8  9 10  6  7  8  9 10 11  7  8  9 10 11 12
 [31]  8  9 10 11 12 13  4  5  6  7  8  9  5  6  7  8  9 10  6  7  8  9 10 11  7  8  9 10 11 12
 [61]  8  9 10 11 12 13  9 10 11 12 13 14  5  6  7  8  9 10  6  7  8  9 10 11  7  8  9 10 11 12
 [91]  8  9 10 11 12 13  9 10 11 12 13 14 10 11 12 13 14 15  6  7  8  9 10 11  7  8  9 10 11 12
[121]  8  9 10 11 12 13  9 10 11 12 13 14 10 11 12 13 14 15 11 12 13 14 15 16  7  8  9 10 11 12
[151]  8  9 10 11 12 13  9 10 11 12 13 14 10 11 12 13 14 15 11 12 13 14 15 16 12 13 14 15 16 17
[181]  8  9 10 11 12 13  9 10 11 12 13 14 10 11 12 13 14 15 11 12 13 14 15 16 12 13 14 15 16 17
[211] 13 14 15 16 17 18
```

What is the chance that a single draw from this is less than 8?

Find the *proportion* less than 8

Use the frequency table

```r
table(y)  # note: first two rows below are only labels: the 'real' output is the third line
```

```
y
 3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 
 1  3  6 10 15 21 25 27 27 25 21 15 10  6  3  1 
```

```r
sum(table(y))  # gives total freq, i.e. size of the box
```

```
[1] 216
```

```r
length(y)  # same as above
```

```
[1] 216
```

```r
round(100 * table(y)/length(y), 1)  # chance for getting each ticket (in percentage)
```

```
y
   3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18 
 0.5  1.4  2.8  4.6  6.9  9.7 11.6 12.5 12.5 11.6  9.7  6.9  4.6  2.8  1.4  0.5 
```

```r
sum(y < 8)  # the vector 'y<8' is of length 216, with TRUE=1 and FALSE=0 if each value <8 or >=8
```

```
[1] 35
```

```r
sum(y < 8)/length(y)
```

```
[1] 0.162037
```

```r
mean(y < 8)  # mean of a vector of 0's and 1's is the *proportion* of 1's
```

```
[1] 0.162037
```

- The chance of drawing a value less than 8 is $\frac{35}{216} \approx 16\%$.
- Note: $35 = 1 + 3 + 6 + 10 + 15$ (the frequencies of 3, 4, 5, 6 and 7 respectively).

Histogram, normal curve

- In some situations, we may not know the *exact* contents of the box, but we might have access to some summary statitsics, so we are able to build an approximation to the box.
- For example, what if the histogram of the box has a normal shape?
- In that case, knowing only the mean and SD of the box, we can approximate *proportions*, and hence chances of getting different values.
- Firstly note the mean and SD for our example `y`:

```r
mn.y = mean(y)
mn.y
```

```
[1] 10.5
```

```r
SD.y = sqrt(mean((y - mn.y)^2))
SD.y
```

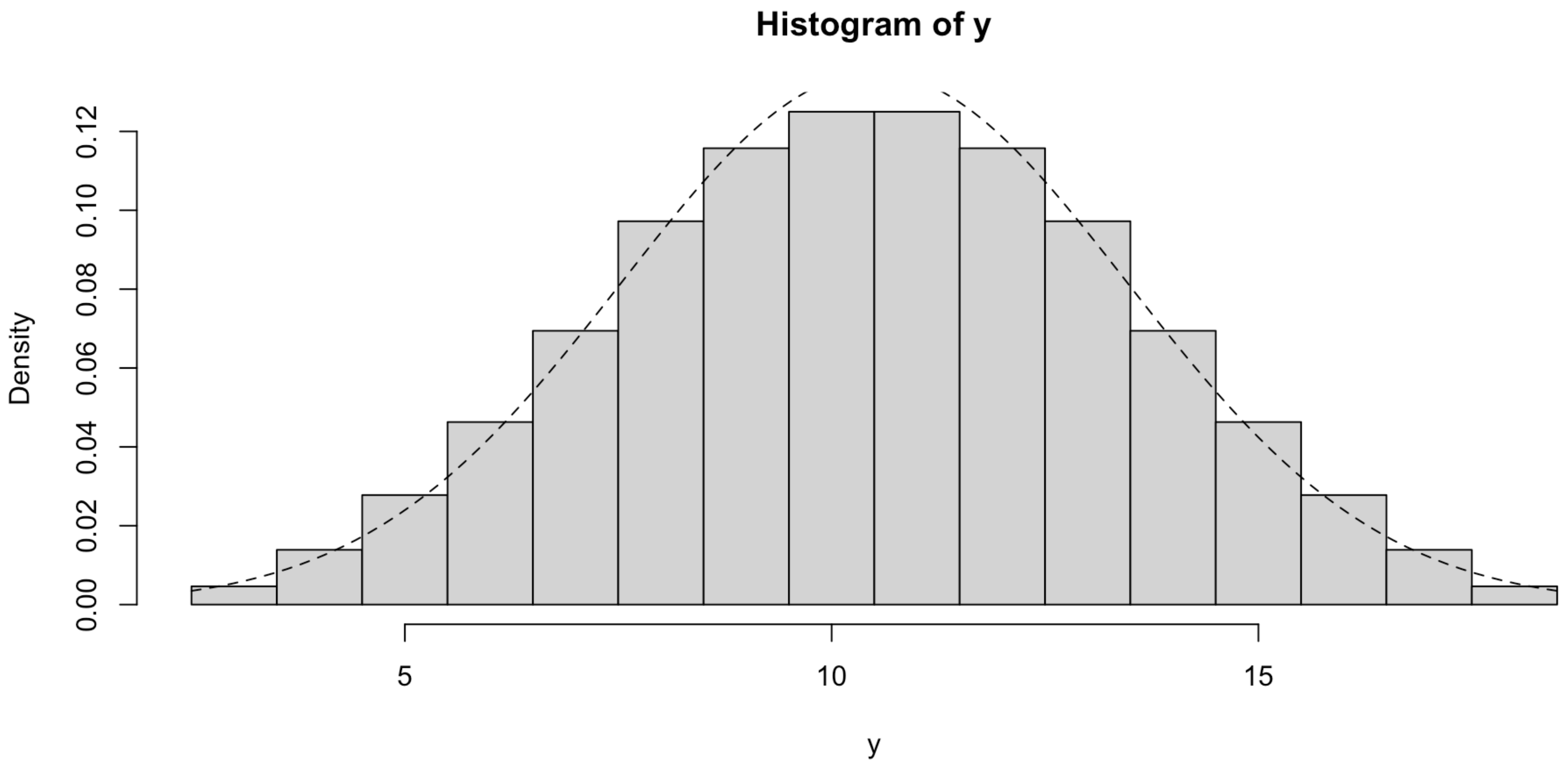
```
[1] 2.95804
```

- Note: box is a population

```r
br = (2:18) + 0.5
br  # this gives rectangles centred on each integer 3,4,...,18
```

```
[1]  2.5  3.5  4.5  5.5  6.5  7.5  8.5  9.5 10.5 11.5 12.5 13.5 14.5 15.5 16.5 17.5 18.5
```

```r
hist(y, breaks = br, pr = T)
curve(dnorm(x, mn.y, SD.y), add = T, lty = 2)  # lty=2 gives a dashed line
```



Normal approximation

We can find the “area” to the left of 8, for a normal curve with the same mean and SD:

```
pnorm(8, mn.y, SD.y)  # not a bad approximation, but a bit big
```

```
[1] 0.1990124
```

Compare this to the “true” value of 16%

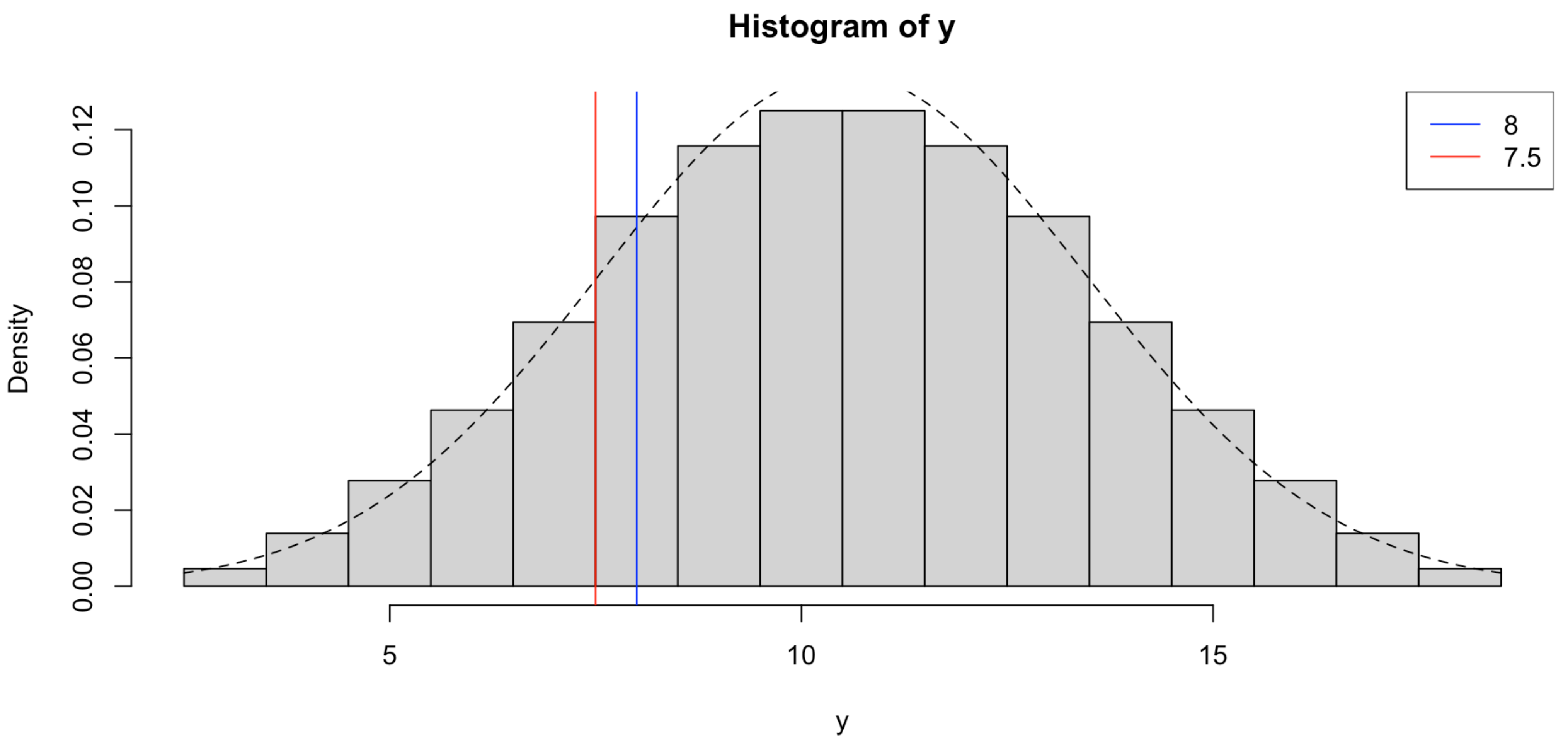
Non-examinable: Note that we can have a *better* approximation:

- all tickets taking integer values (whole numbers), 3, 4, 5, …
- so < 8 is the same as < 7.5, so the area under the rectangles we want is actually to the left of 7.5 (see the histogram repeated on the next slide):

```
pnorm(7.5, mn.y, SD.y)  # much closer to the true value!
```

```
[1] 0.1552472
```

```r
hist(y, breaks = br, pr = T)
curve(dnorm(x, mn.y, SD.y), add = T, lty = 2)  # lty=2 gives a dashed line
abline(v = 8, col = "blue")
abline(v = 7.5, col = "red")
legend("topright", leg = c("8", "7.5"), lty = c(1, 1), col = c("blue", "red"))
```



New interpretation of mean and SD of box

When we are taking a random draw X from a box, we see that the mean and SD of the box have a new, special interpretation.

We call the mean of the box the **expected value** of the random draw:

- We write this as $E(X)$.

We call the SD of the box the **standard error** of the random draw:

- We write this as $SE(X)$.

Random draw = Expected value + Chance error

- The random draw may be "decomposed" into two pieces:

$$X = E(X) + [X - E(X)] = E(X) + \varepsilon\,.$$

- The first part $E(X)$ is *not random*.
- All randomness is included in the chance error ε, which is itself a random draw from an **error box** (a box with mean zero).
- **Example**: a random draw X from the box (box 1)

$$\boxed{\boxed{1}\,\boxed{2}\,\boxed{3}}$$

(which has mean 2) may instead be thought of as $X = 2 + \varepsilon$ where the chance error ε is a random draw from the error box

$$\boxed{\boxed{-1}\,\boxed{\ 0\ }\,\boxed{+1}}\,.$$

- Note that the error box just contains all the deviations (and hence zero mean).

Standard error

- The **standard error** is the "root-mean-square" of the error box.
  - ➟ It is also the (population) SD of the errors (deviations) – the error box has zero mean

$$SE(X) = SD(\epsilon) = \sqrt{\frac{1}{3}[(-1-0)^2 + (0-0)^2 + (1-0)^2]} = \sqrt{\frac{2}{3}}$$

.

$$SD(X) = \sqrt{\frac{1}{3}[(1-2)^2 + (2-2)^2 + (3-2)^2]} = \sqrt{\frac{2}{3}}$$

.

- It measures the spread of the errors, and thus the size of the variation of errors.
- For two different random draws, the one with the larger SE is likely to differ from its expected value by a larger amount.

Sums of random draws

New interpretation of mean and SD

We have introduced the concepts of

- A random draw X from a box;
- Its expected value $E(X)$ (fixed value for a given box);

$$X = E(X) + [X - E(X)] = E(X) + \varepsilon\,.$$

- Its standard error $SE(X)$, measuring the size of variation of the error ϵ.

The expected value and standard error are not “new” things;

- Rather, they are new interpretations of old things.

Is it really “worth the effort” to introduce these new names for these things are already know about?

- They are the standard ways to describe random behavious in text books.
- The expected value and standard error become very useful when we have **more than one draw**.

Sum of two random draws

- Consider the two boxes (box 1 with equal chance to get each ticket)

$$\boxed{\boxed{1}\,\boxed{2}\,\boxed{3}} \text{ and } \boxed{\boxed{2}\,\boxed{4}\,\boxed{6}\,\boxed{8}}.$$

- The first box has mean 2 and SD $\sqrt{\frac{1}{3}[(-1)^2+0^2+1^2]} = \sqrt{\frac{2}{3}} \approx 0.816$.
- The second box has mean 5 and SD

$$\sqrt{\frac{1}{4}[(-3)^2+(-1)^2+1^2+3^2]} = \sqrt{5} \approx 2.236\,.$$

- Suppose we take a random draw from each, X from the first box, Y from the second box, in such a way that **each possible pair of values is equally likely**.
- What is the behaviour of the (random) **sum** $S = X + Y$?

All possible pairs/sums

- There are 12 possible pairs:

$$\left(\boxed{1}, \boxed{2}\right), \left(\boxed{1}, \boxed{4}\right), \left(\boxed{1}, \boxed{6}\right), \left(\boxed{1}, \boxed{8}\right),$$

$$\left(\boxed{2}, \boxed{2}\right), \left(\boxed{2}, \boxed{4}\right), \left(\boxed{2}, \boxed{6}\right), \left(\boxed{2}, \boxed{8}\right),$$

$$\left(\boxed{3}, \boxed{2}\right), \left(\boxed{3}, \boxed{4}\right), \left(\boxed{3}, \boxed{6}\right), \left(\boxed{3}, \boxed{8}\right).$$

Table of all possible pairs and their sums

Sample	Sum
(1,2)	3
(1,4)	5
(1,6)	7
(1,8)	9
(2,2)	4
(2,4)	6
(2,6)	8
(2,8)	10
(3,2)	5
(3,4)	7
(3,6)	9
(3,8)	11

Single random draw from a "bigger" box

- Thus getting a random pair (X, Y) and forming the sum $S = X + Y$ is **equivalent** to a *single random draw* from the bigger box

$$\boxed{\boxed{3}\ \boxed{4}\ \boxed{5}\ \boxed{5}\ \boxed{6}\ \boxed{7}\ \boxed{7}\ \boxed{8}\ \boxed{9}\ \boxed{9}\ \boxed{10}\ \boxed{11}}$$

- What are the mean and SD of this "bigger" box?

Using `outer()`

- The R function `outer()` forms a two-way array by applying an operation to each pair of elements from two vectors:

```r
bx = c(1, 2, 3)
by = c(2, 4, 6, 8)
bs = outer(bx, by, "+")
bs
```

```
     [,1] [,2] [,3] [,4]
[1,]    3    5    7    9
[2,]    4    6    8   10
[3,]    5    7    9   11
```

```r
mean(bs)   # mean
```

```
[1] 7
```

```r
mean((bs - mean(bs))^2)   # population variance
```

```
[1] 5.666667
```

```r
sqrt(mean((bs - mean(bs))^2))   # population SD
```

```
[1] 2.380476
```

Expected value and standard error of the sum

- So we have that $E(S) = 7$ and $SE(S) = \sqrt{5 + \frac{2}{3}} \approx 2.38$.
- Note that we have

$$7 = E(S) = E(X+Y) = E(X) + E(Y) = 2 + 5\,.$$

- We also have

$$5 + \frac{2}{3} = SE(S)^2 = SE(X+Y)^2 = SE(X)^2 + SE(Y)^2 = \frac{2}{3} + 5\,.$$

- So in this case we have
    - expected value of sum is sum of expected values;
    - *squared* SE of the sum is the sum of the *squared* SEs
- These results hold quite generally.

Sum of two random draws.

- Consider two boxes (box 1)

$$\boxed{\boxed{x_1}\ \boxed{x_2}\ \cdots\ \boxed{x_M}} \text{ and } \boxed{\boxed{y_1}\ \boxed{y_2}\ \cdots\ \boxed{y_N}}$$

- Suppose we are going to take a random draw from each: X from the first box, Y from the second box, in such a way that **each possible pair of values is equally likely**.
- The expected value of the sum is the sum of the expected values

$$E(S) = E(X+Y) = E(X) + E(Y)\,.$$

- The squared SE of the sum is the sum of the squared SEs

$$SE(S)^2 = SE(X+Y)^2 = SE(X)^2 + SE(Y)^2\,.$$

All possible sums

- There are MN possible sums, we may arrange them in a two-way array with M (horizontal) rows and N (vertical) columns.
- Noting that $\sum_{i=1}^{M} x_i = M\bar{x}$, we may write the column sums below the line:

$$
\begin{array}{cccc}
x_1 + y_1 & x_1 + y_2 & \cdots & x_1 + y_N \\
x_2 + y_1 & x_2 + y_2 & \cdots & x_2 + y_N \\
\vdots & \vdots & \ddots & \vdots \\
x_M + y_1 & x_M + y_2 & \cdots & x_M + y_N \\
\hline
M\bar{x} + My_1 & M\bar{x} + My_2 & \cdots & M\bar{x} + My_N
\end{array}
$$

- The sum of column sums is

$$\underbrace{M\bar{x} + \cdots + M\bar{x}}_{N \text{ terms}} + M(y_1 + \cdots + y_N) = NM\bar{x} + MN\bar{y}\,.$$

- Thus the *average* of all possible sums is

$$\frac{\text{sum of all possible sums}}{\text{no. of all possible sums}} = \frac{NM\bar{x} + MN\bar{y}}{MN} = \bar{x} + \bar{y} = E(X) + E(Y)\,.$$

- That is,

$$E(S) = E(X + Y)\,.$$

Not Examinable: SE of a sum

- It is possible to deduce the SE of our general sum $S = X + Y$.
- We do so by first working out the mean-square of the bigger box of all possible sums.
- Write each squared sum $(x_i + y_j)^2 = x_i^2 + 2x_iy_j + y_j^2$ in an array and add over columns:

$$\begin{array}{ccc}
x_1^2 + 2x_1y_1 + y_1^2 & \cdots & x_1^2 + 2x_1y_N + y_N^2 \\
x_2^2 + 2x_2y_1 + y_1^2 & \cdots & x_2^2 + 2x_2y_N + y_N^2 \\
\vdots & \ddots & \vdots \\
x_M^2 + 2x_My_1 + y_1^2 & \cdots & x_M^2 + 2x_My_N + y_N^2 \\
\hline
\sum_i x_i^2 + 2M\bar{x}y_1 + My_1^2 & \cdots & \sum_i x_i^2 + 2M\bar{x}y_N + My_N^2
\end{array}$$

Not Examinable

- The sum of squares (of all possible sums) is then

$$
\begin{aligned}
&N \sum_i x_i^2 + 2M\bar{x}(y_1 + \cdots + y_N) + M(y_1^2 + \cdots + y_N^2) \\
&= N \sum_i x_i^2 + 2MN\bar{x}\bar{y} + M \sum_j y_j^2 \,.
\end{aligned}
$$

- Since there are MN possible sums, the mean square is

$$
\frac{1}{M} \sum_i x_i^2 + 2\bar{x}\bar{y} + \frac{1}{N} \sum_j y_j^2 \,.
$$

Not Examinable

- Since mean of all possible sums is $\bar{x} + \bar{y}$, the squared SD of all possible sums is

$$\underbrace{\frac{1}{M}\sum_i x_i^2 + 2\bar{x}\bar{y} + \frac{1}{N}\sum_j y_j^2}_{mean\ sq.} - \underbrace{\left(\bar{x}^2 + 2\bar{x}\bar{y} + \bar{y}^2\right)}_{sq.\ of\ mean}$$
$$= \frac{1}{M}\sum_i x_i^2 - \bar{x}^2 + \frac{1}{N}\sum_j y_j^2 - \bar{y}^2$$
$$= \frac{1}{M}\sum_i (x_i - \bar{x})^2 + \frac{1}{N}\sum_j (y_j - \bar{y})^2$$
$$= SE(X)^2 + SE(Y)^2\,.$$

- That is,

$$SE(S)^2 = SE(X)^2 + SE(Y)^2\,.$$

Sums and averages of random samples of size n

Random samples with replacement of size $n = 2$

- A special case of our general sum is where we have a **single** box (box 1)

$$\boxed{\boxed{x_1}\ \boxed{x_2}\ \cdots\ \boxed{x_N}}$$

  but take two random draws with replacement.
    - This means each of the N^2 possible pairs $(x_1, x_1), \ldots, (x_1, x_n), \ldots, (x_n, x_1), \ldots, (x_n, x_n)$ is **equally likely**.
- This is where both boxes are (effectively) the same, so $E(X) = E(Y)$ and $SE(X) = SE(Y)$.
- If we write the mean of the box as μ and the SD of the box as σ, then the sum S of the two random draws has
    - $E(S) = 2\mu$
    - $SE(S) = \sqrt{2}\sigma$ – because $SE(S)^2 = 2\sigma^2$

Random samples of size n and sample average

- We may easily extend the results to any $n \geq 2$.
- Suppose
  - We have a box with mean μ and SD σ;
  - We are going to take a random sample of size n from the box **with replacement**;
  - So each possible sample of size n is equally likely.
- Let us write
  - The random draws as $X_1, X_2, \ldots, X_n$;
  - The sum as $S = X_1 + \cdots + X_n$;
  - The **sample average** as $\bar{X} = \frac{S}{n} = \frac{1}{n}(X_1 + \cdots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$.
- What are the expected value and standard error of both S and $\bar{X}$?

The sum S

- We may extend our results from $n = 2$ easily.
- Each single draw has the same behaviour.
- X_1 (the first draw) is a single random draw and so has
  - $E(X_1) = \mu$
  - $SE(X_1) = \sigma$.
- The same is true for each other draw.
- Expected value of sum is sum of expected values:

$$E(S) = E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n) = \underbrace{\mu + \cdots + \mu}_{n \text{ terms}} = n\mu\,.$$

- Also, $SE(S)^2 = SE(X_1)^2 + \cdots + SE(X_n)^2 = n\sigma^2$, so

$$SE(S) = \sigma\sqrt{n}\,.$$

Going from the sum to the average

- So the "box of all possible sums" has mean $n\mu$ and SD $\sigma\sqrt{n}$.
- How about the box of all possible sample averages?
- The box of all possible sample averages is obtained by taking each possible sum and dividing it by n.
- This has the effect of
    - dividing the mean (of the sample sum) by n;
    - also dividing the SD (of the sample sum) by n.

The sample average $\bar{X}$

- We thus obtain immediately that for the average $\bar{X} = \frac{S}{n} = \frac{X_1 + \cdots + X_n}{n}$,

$$E(\bar{X}) = \frac{E(S)}{n} = \frac{n\mu}{n} = \mu\,;$$

- So the “bigger box” of all possible sample means has average equal to the “population mean” μ;
  - ➟ this is not surprising.
- As for the standard error we have

$$SE(\bar{X}) = \frac{SE(S)}{n} = \frac{\sigma\sqrt{n}}{n} = \frac{\sigma}{\sqrt{n}}\,.$$

Example

6-sided die

- Consider rolling a fair 6-sided die.
- In this case each of the numbers 1,2,3,4,5,6 are equally likely.
- This is equivalent to a random draw three times from the box (with replacement)

$$\boxed{\boxed{1}\ \boxed{2}\ \boxed{3}\ \boxed{4}\ \boxed{5}\ \boxed{6}}$$

which has expectation $\mu = 3.5 = \frac{7}{2}$, mean-square $\frac{1+4+9+16+25+36}{6} = \frac{91}{6}$ and thus SE

$$\sigma = \sqrt{\frac{91}{6} - \frac{49}{4}} = \sqrt{\frac{182 - 147}{12}} = \sqrt{\frac{35}{12}} \approx 1.71\,.$$

Rolling the die 3 times: Sum of rolls

- Suppose we roll the die ("independently") 3 times.
- What is the random behaviour of the **sum** of the values of the three rolls?
- Let X_1, X_2, X_3 denote 3 random draws with replacement from the box

$$\boxed{\boxed{1}\ \boxed{2}\ \boxed{3}\ \boxed{4}\ \boxed{5}\ \boxed{6}}$$

- Then the sum of the 3 rolls $S = X_1 + X_2 + X_3$ has $E(S) = 3\mu = \frac{21}{2} = 10.5$ and

$$SE(S) = \sigma\sqrt{3} = \sqrt{\frac{35}{12} \times 3} = \sqrt{\frac{35}{4}} = \frac{\sqrt{35}}{2} \approx 2.958\,.$$

- The box of all possible sums here is exactly the dataset `y.dat` from earlier in the lecture!

Rolling the die 3 times: Average of rolls

- What is the random behaviour of the **average** of the values of the three rolls?
- Writing $\bar{X} = \frac{X_1+X_2+X_3}{3} = \frac{S}{3}$, we have

$$E(\bar{X}) = \frac{E(S)}{3} = \frac{3\mu}{3} = \mu = 3.5$$

and

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{3}} = \sqrt{\frac{35}{12} \times \frac{1}{3}} = \sqrt{\frac{35}{36}} = \frac{\sqrt{35}}{6} \approx 0.956\,.$$

Demonstration

- Let us simulate 3 rolls of a 6-sided die 100,000 times, and look at the corresponding 100,000 sums and averages of each triplet.

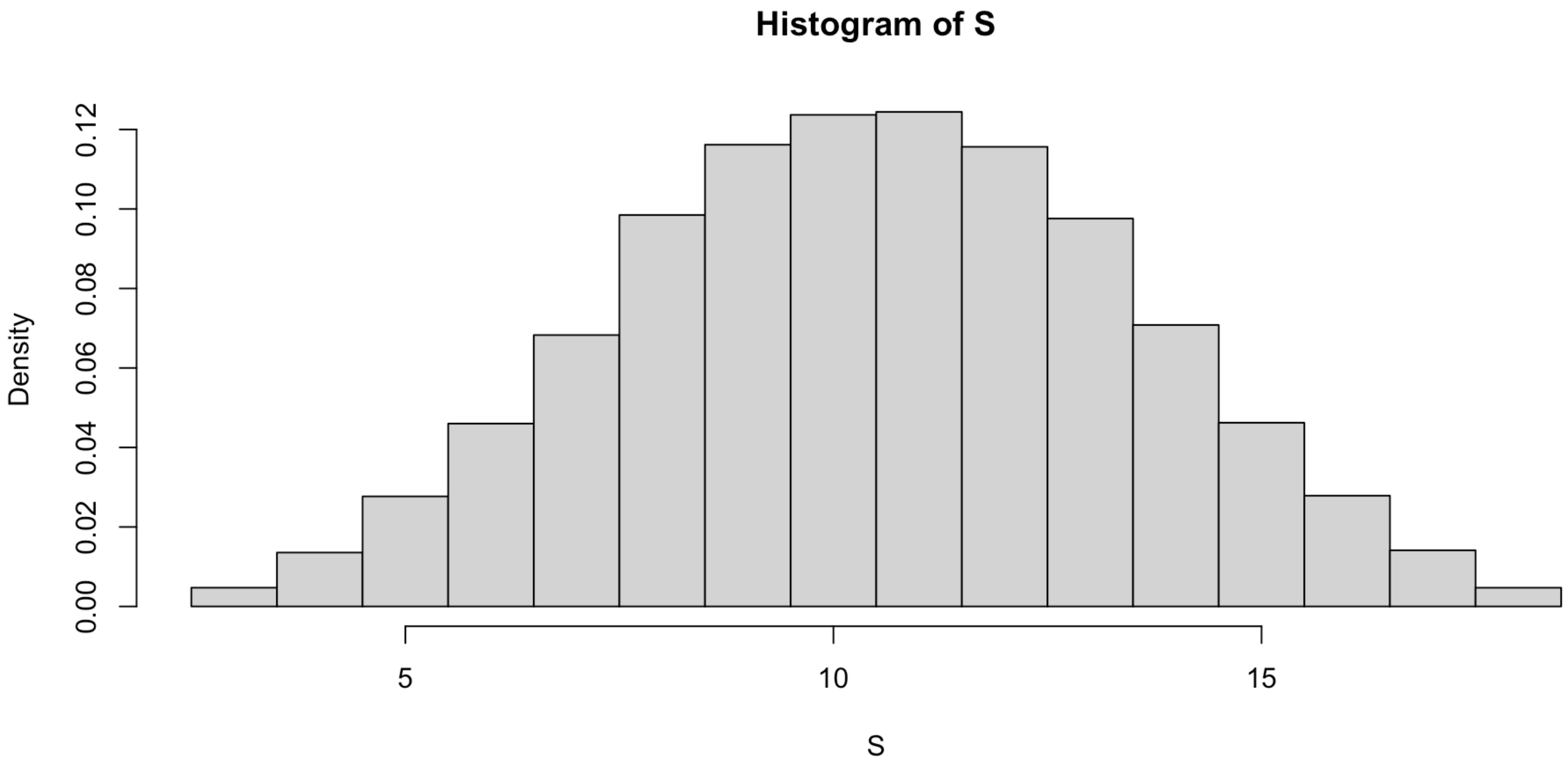
```r
rolling_sum = function(n) {
    # rolling n times, sample with replacement
    rolls = sample(1:6, size = n, rep = T)
    # taking the sum
    return(sum(rolls))
}
S = replicate(1e+05, rolling_sum(3))
mean(S)
```

```
[1] 10.51179
```

```r
sd(S)
```

```
[1] 2.960045
```

```r
hist(S, pr = T, breaks = br)
```



Note these proportions are *close* to (but not *exactly* equal to) the corresponding proportions in `y.dat`.

Averages

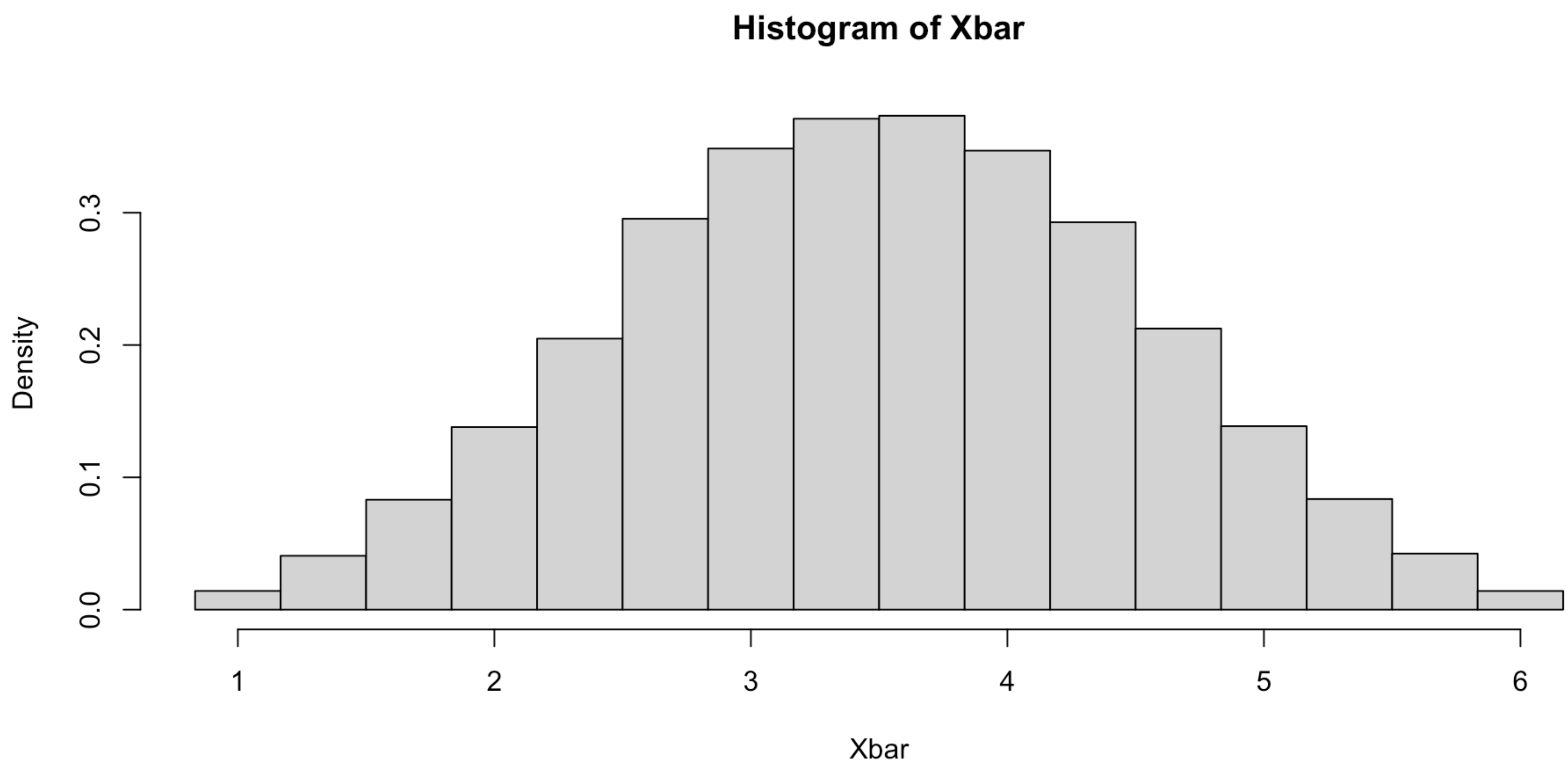
```r
Xbar = S/3
mean(Xbar)
```

```
[1] 3.50393
```

```r
sd(Xbar)
```

```
[1] 0.9866818
```

```r
hist(Xbar, pr = T, breaks = br/3)
```



Same shape as for the sums, but centred on 3.5 and less spread-out.

Increase the number of rolls, n

```r
rolling_average = function(n) {
    # rolling n times, sample with replacement
    rolls = sample(1:6, size = n, rep = T)
    # taking the average (mean)
    a = mean(rolls)
    return(a)
}
```

$n = 10$, $E(\bar{X}) = \mu = 3.5$ and $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\frac{35}{12}}}{\sqrt{10}} \approx 0.54$

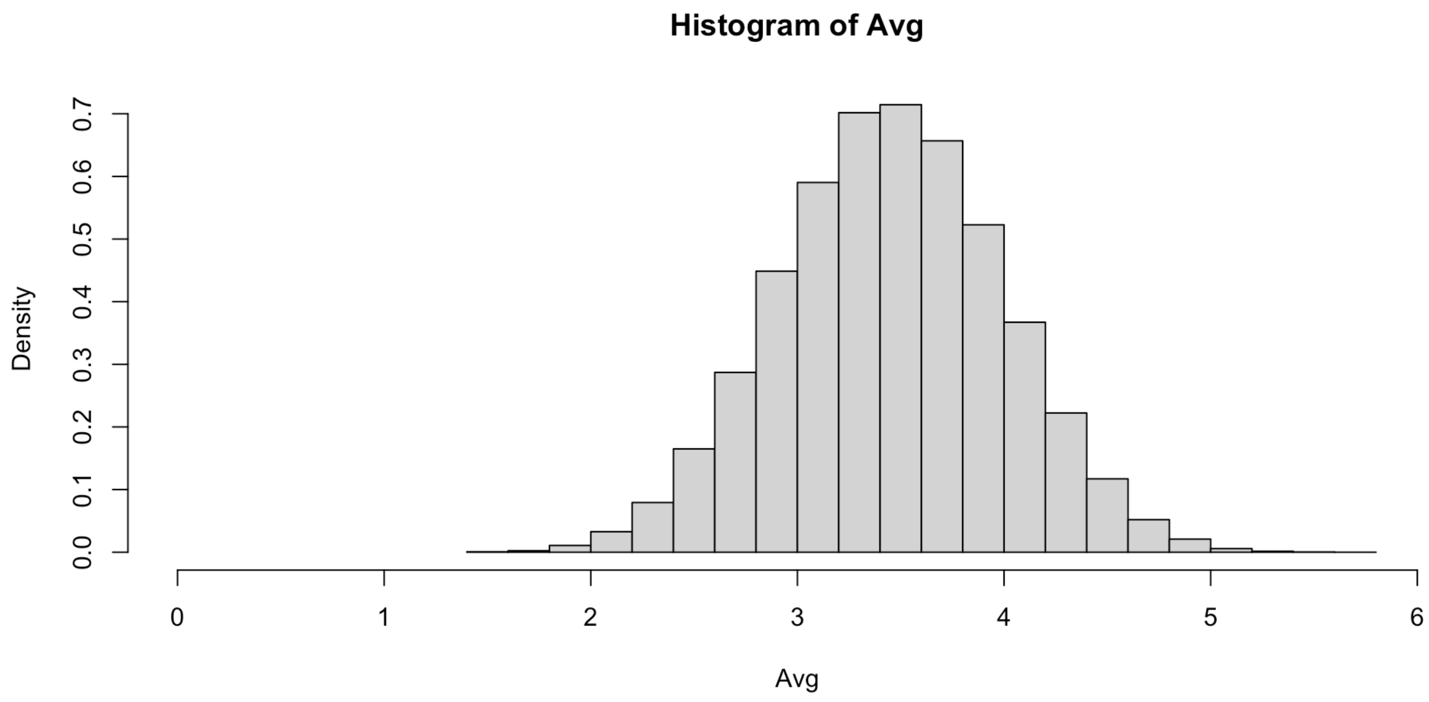
```r
Avg = replicate(1e+05, rolling_average(10))
mean(Avg)
```

```
[1] 3.501198
```

```r
sd(Avg)
```

```
[1] 0.541385
```

```r
hist(Avg, freq = F, xlim = c(0, 6))
```



$n = 100$, $E(\bar{X}) = \mu = 3.5$ and $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\frac{35}{12}}}{\sqrt{100}} \approx 0.171$

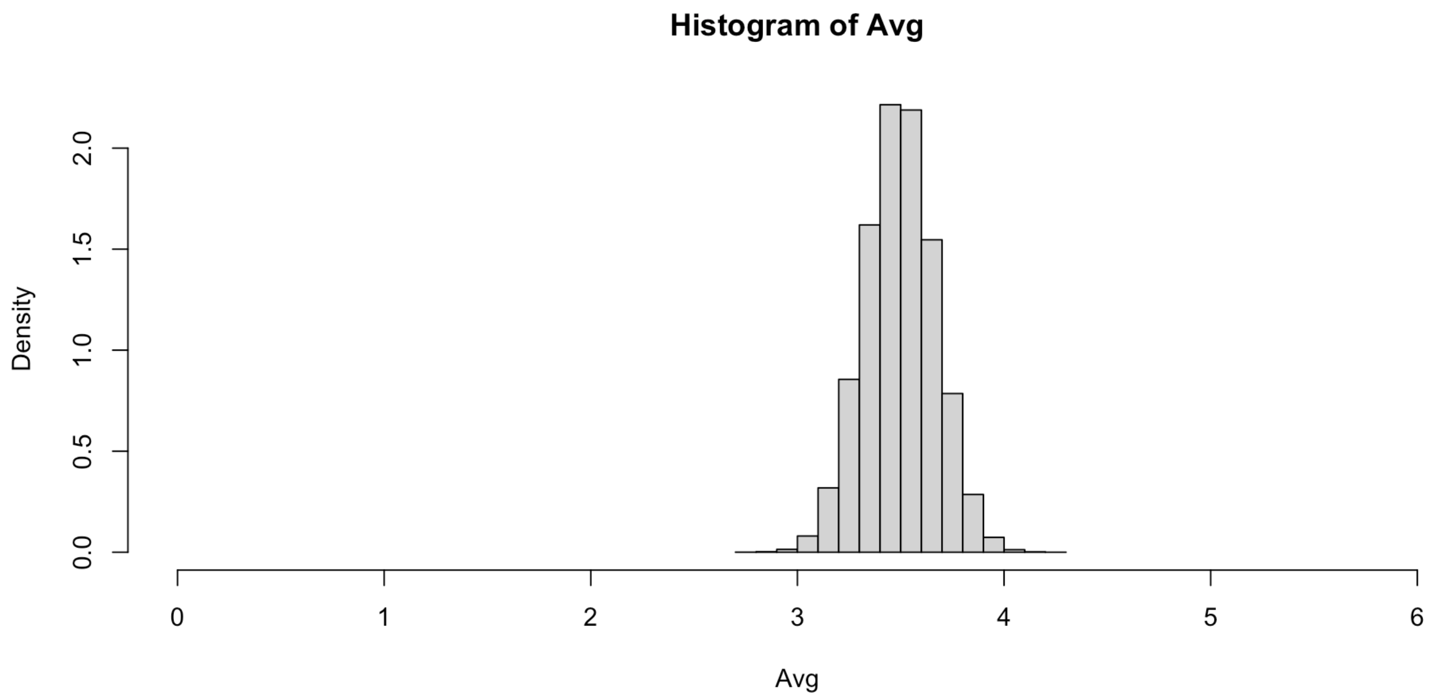
```r
Avg = replicate(1e+05, rolling_average(100))
mean(Avg)
```

```
[1] 3.500501
```

```r
sd(Avg)
```

```
[1] 0.1706137
```

```r
hist(Avg, freq = F, xlim = c(0, 6))
```



$n = 1000$, $E(\bar{X}) = \mu = 3.5$ and $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{\frac{35}{12}}}{\sqrt{1000}} \approx 0.054$

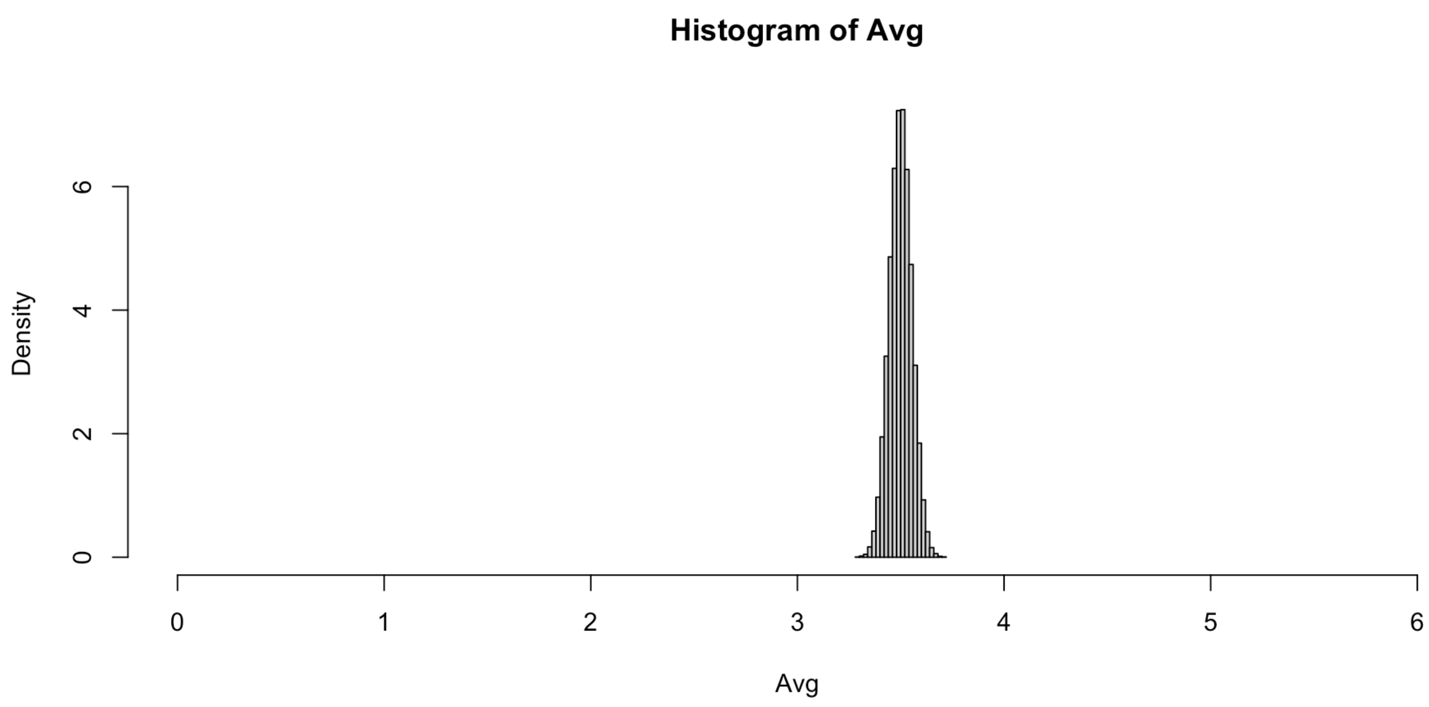
```r
Avg = replicate(1e+05, rolling_average(1000))
mean(Avg)
```

```
[1] 3.499856
```

```r
sd(Avg)
```

```
[1] 0.05398677
```

```r
hist(Avg, freq = F, xlim = c(0, 6))
```



Closing remarks: n getting larger

- Consider a box with mean μ and population SD σ
  - ➟ It has expectation μ and SE σ
- We have seen that for n random draws (with replacement) from this box
  - ➟ the *sum* of draws S has $E(S) = n\mu$ and $SE(S) = \sigma\sqrt{n}$;
  - ➟ the *average* of the draws $\bar{X}$ has $E(\bar{X}) = \mu$ and $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.
- What happens to the SE of each as n gets bigger?
  - ➟ for the sum, $\sigma\sqrt{n}$ gets larger **but**
  - ➟ for the average, $\frac{\sigma}{\sqrt{n}}$ gets **smaller**.
- In particular, for the average $\bar{X}$, the random variability about $E(\bar{X}) = \mu$ gets less as the sample size n increases.

Summary of box model formulas

Box Model	Expected Value E(X)	Standard Error SE(X)
Sum of draws	$n \times$ mean of the box	$\sqrt{n} \times$ SD of the box
Mean of draws	mean of the box	$\frac{\text{SD of the box}}{\sqrt{n}}$

n: number of draws