Central Limit Theorem

Sampling Data | Chance Variability

STAT5002

The University of Sydney

Mar 2025

THE UNIVERSITY OF SYDNEY

Sampling Data

Topic 5: Understanding chance and chance simulation

Topic 6: Chance variability

Topic 7: Central limit theorem

Outline

A review of box models

Increasing the sample size

The Central Limit Theorem (CLT)

A review of box models

Single draws from box models

- Suppose we have a "box" containing tickets each bearing a number: $\{x_1, \ldots, x_N\}$.
- The probability a random draw X from the box takes a value is just the *proportion of* x_i *values*
- Recall the example:

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

➟ if each "ticket" is equally likely, we have

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

➟ the random draw X then has **distribution**

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

Expectation and standard error

- $\mu = \frac{1}{N}(x_1 + \cdots + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i$
  - the mean of the box,
  - also called $E(X)$, the **expected value** of the random draw X;
- $\sigma = \sqrt{\frac{1}{N}[(x_1 - \mu)^2 + \cdots + (x_N - \mu)^2]}$
  - the (population) SD of the box,
  - also called $SE(X)$, the **standard error** of the random draw X.

Chance error

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

$$X = \underbrace{E(X)}_{\text{mean, not random}} + \underbrace{[X - E(X)]}_{\text{chance error, random}} = E(X) + \varepsilon\,.$$

- The first part $E(X) = \mu$ is *not random*.
- All randomness is included in the **chance error** ε, which is a random draw from an **error box** $\{x_1 - \mu, \ldots, x_N - \mu\}$.
  - The error box has zero mean and its SD is the same as the SD of the original box
  - the chance error has $E(\varepsilon) = 0$ and $SE(\varepsilon) = SE(X)$
- So $SE(X)$ is interpreted as the "likely size" of the chance error ε, i.e. the likely size of the deviation of X from its expected value $E(X)$.

Sum of random draws

Consider the sum of n random draws (which is a sample)

$$S = X_1 + X_2 + \cdots + X_n$$

where each X_j is random draw **with replacement** from a box $\{x_1, \ldots, x_N\}$ with mean μ and SD σ.

Then, the sum S is a single random draw from a much larger box. We have

- $E(S) = n\mu$
- $SE(S) = \sqrt{n}\sigma$

In other words, the box containing all possible sums S has the mean $n\mu$ and the SD $\sqrt{n}\sigma$

Average of random draws

Now consider the sample mean of n random draws

$$\bar{X} = \frac{1}{n}S = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

Then, the sample mean $\bar{X}$ is also a single random draw from a much larger box

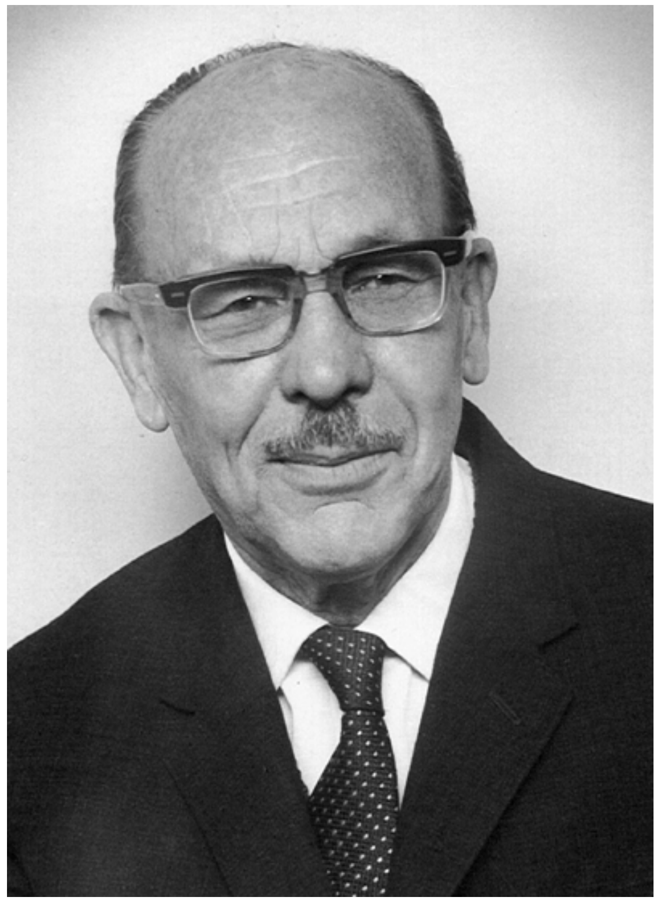
- similar to the box of all possible sums, but each ticket is scaled by $\frac{1}{n}$.
- $E(\bar{X}) = \mu$
- $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

In other words, the box containing all possible sample means $\bar{X}$ has the mean μ and the SD $\frac{\sigma}{\sqrt{n}}$

Examples: the law of average and MORE

Example: coin tossing in WWII

- John Edmund Kerrich (1903–1985) was a mathematician noted for a series of experiments in probability which he conducted while interned in Nazi-occupied Denmark (Viborg, Midtjylland) in the 1940s.
- Kerrich had travelled from South Africa to visit his in-laws in Copenhagen, and arrived just 2 days after Denmark was invaded by Nazi Germany!



Various “random experiments”

With a fellow internee Eric Christensen, Kerrich set up a sequence of experiments demonstrating the empirical validity of a number of fundamental laws of probability.

- **They tossed a (fair) coin 10,000 times and counted the number of heads (5,067).**
- They made 5,000 draws from a container with 4 ping pong balls (2x2 different brands), ‘at the rate of 400 an hour, with - need it be stated - periods of rest between successive hours.’
- They investigated tosses of a “biased coin”, made from a wooden disk partly coated in lead.

In 1946 Kerrich published his finding in a monograph An Experimental Introduction to the Theory of Probability.

Simulating Kerrich's 1st experiment

- Each coin flip (assuming the coin is fair) is like a random draw from the "box"

$$\boxed{\boxed{0}\,\boxed{1}}$$

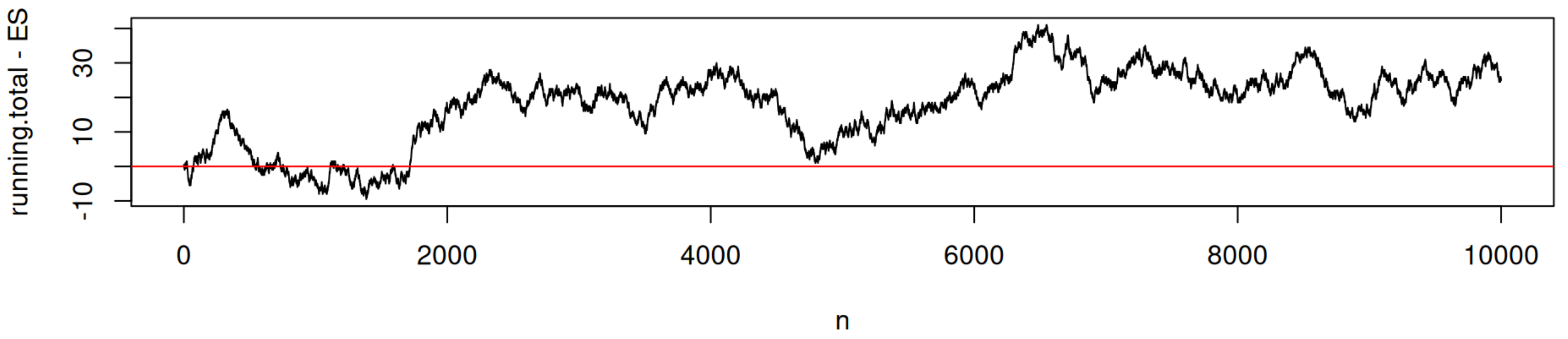
- This box has average $\mu = \frac{1}{2}$ and also SD

$$\sigma = \sqrt{\textbf{mean square} - (\textbf{mean})^2} = \sqrt{\frac{1}{2} - \left(\frac{1}{2}\right)^2} = \sqrt{\frac{1}{4}} = \frac{1}{2}\,.$$

- We may then model n "independent" flips $X_1, \ldots, X_n$ as a random sample with replacement of size n from this box.
- The sum $S = X_1 + \cdots + X_n$ is the **number** of heads.
- The average $\bar{X} = S/n$ is the **proportion** of heads.

Simulating 1st experiment: chance error in sums

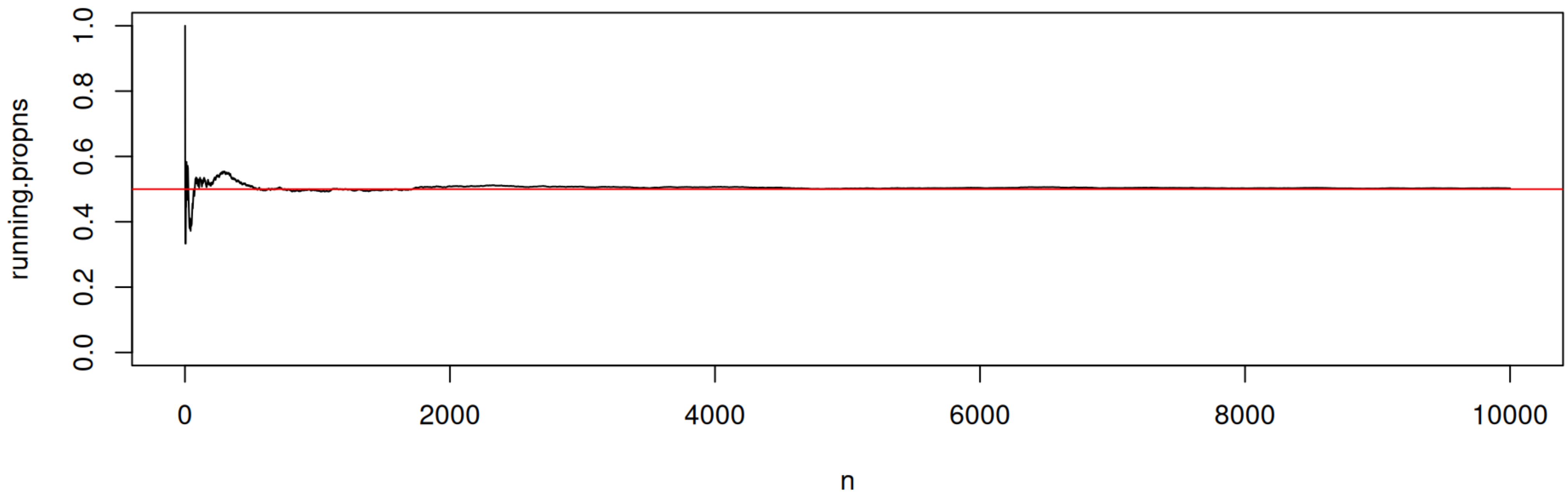
```
flips = sample(c(0, 1), size = 10000, replace = T)  # 'box' is c(0,1)
running.total = cumsum(flips)  # cumulative summation
n = 1:10000  # sample sizes
ES = n/2  # expeced sum for each sample size
plot(n, running.total - ES, type = "l")  # plot the chance error
abline(h = 0, col = "red")
```



- `running.total = cumsum(flips)`: gives the progressive addition of a sequence of numbers in `flips`
  - ➠ `running.total[j]` is the same as `sum(flips[1:j])`

chance error in averages (cumulative proportion)

```r
running.propns = running.total/n  # remember n = 1:10000  is a vector!
plot(n, running.propns, type = "l", ylim = c(0, 1))
abline(h = 0.5, col = "red")
```



We are able to see that as we flip the coin more times, the proportion of heads compared to tails becomes more even. We can also see that there is a lot of fluctuation at the start.

Size of chance errors as n increases

It seems that

- The size of the chance error in the **sums increases**;
- The size of the chance error in the **proportion decreases**;

This makes perfect sense, because

- The "likely size" of the chance error for the **sum**, i.e.

  $$SE(S) = \sigma\sqrt{n} \to \infty$$

  as $n \to \infty$
- The "likely size" of the chance error for the **proportion**, i.e.

  $$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} \to 0$$

  as $n \to \infty$.

Law of Averages

- For the sample mean $\bar{X}$ from *any* box model,

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} \to 0$$

  as $n \to \infty$.
- So the likely size of the chance error between $\bar{X}$ and $E(\bar{X}) = \mu$ gets smaller and smaller as n increases.
- In other words, as the sample size n increases, the distribution of a sample mean $\bar{X}$ gets "more concentrated" about the "population mean" μ.
- Sample mean $\bar{X}$ can be a good estimation of the population mean for large sample size n.
- This "phenomenon" is (loosely) known as the "Law of Averages" or the "Law of Large Numbers".

Demonstration

- We can determine the box of all possible sums for small values of n:

```r
box = c(0, 1)
s2 = outer(box, box, "+")  # forms two-way array of all possible sums for n=2
s2
```

```
     [,1] [,2]
[1,]    0    1
[2,]    1    2
```

```r
as.vector(s2)  # converts matrix to a vector
```

```
[1] 0 1 1 2
```

- We can iterate this procedure to get all sums for $n = 3$:

```r
s3 = as.vector(outer(box, s2, "+"))  # each sum for n=3 adds 0 or 1 to each sum in s2
s3
```

```
[1] 0 1 1 2 1 2 2 3
```

- Again, for $n = 4$:

```
s4 = as.vector(outer(box, s3, "+"))  # each sum for n=4 adds 0 or 1 to each sum in s3
s4
```

```
[1] 0 1 1 2 1 2 2 3 1 2 2 3 2 3 3 4
```

- Again, for $n = 5$:

```
s5 = as.vector(outer(box, s4, "+"))  # each sum for n=5 adds 0 or 1 to each sum in s4
s5
```

```
[1] 0 1 1 2 1 2 2 3 1 2 2 3 2 3 3 4 1 2 2 3 2 3 3 4 2 3 3 4 3 4 4 5
```

- Again, for $n = 6$:

```
s6 = as.vector(outer(box, s5, "+"))  # each sum for n=6 adds 0 or 1 to each sum in s5
s6
```

```
 [1] 0 1 1 2 1 2 2 3 1 2 2 3 2 3 3 4 1 2 2 3 2 3 3 4 2 3 3 4 3 4 4 5 1 2 2 3 2 3
[39] 3 4 2 3 3 4 3 4 4 5 2 3 3 4 3 4 4 5 3 4 4 5 4 5 5 6
```

All possible sums to all possible averages

$n = 2$

```r
m2 = as.vector(s2)/2
m2
```

```
[1] 0.0 0.5 0.5 1.0
```

$n = 3, 4, \ldots$

```r
m3 = s3/3
m3
```

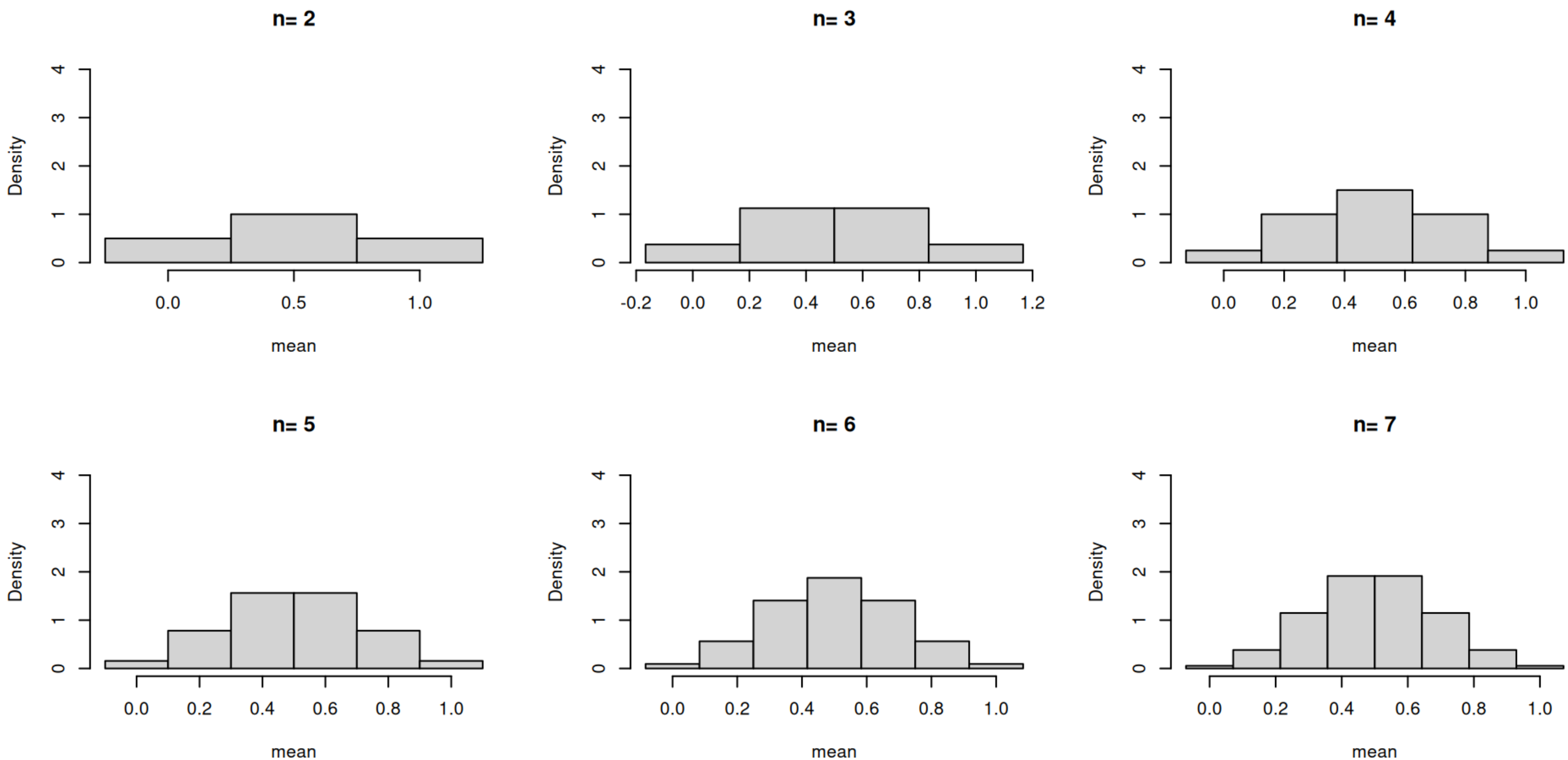
```
[1] 0.0000000 0.3333333 0.3333333 0.6666667 0.3333333 0.6666667 0.6666667
[8] 1.0000000
```

```r
m4 = s4/4
m5 = s5/5
m6 = s6/6
s7 = as.vector(outer(box, s6, "+"))
m7 = s7/7
means = list(`n=2` = m2, `n=3` = m3, `n=4` = m4, `n=5` = m5, `n=6` = m6, `n=7` = m7)
```

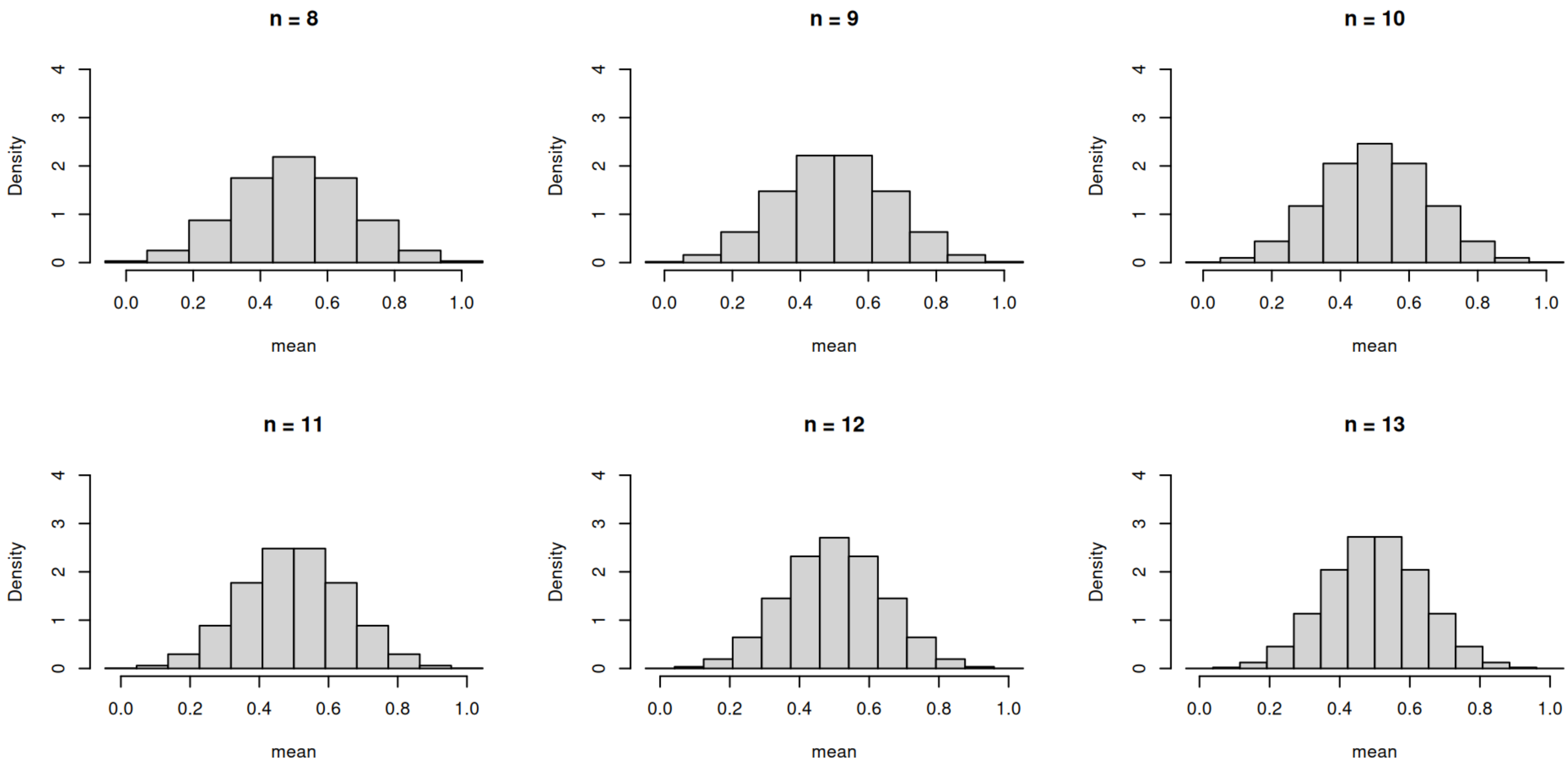
```r
means
```

```
$`n=2`
[1] 0.0 0.5 0.5 1.0

$`n=3`
[1] 0.0000000 0.3333333 0.3333333 0.6666667 0.3333333 0.6666667 0.6666667
[8] 1.0000000

$`n=4`
 [1] 0.00 0.25 0.25 0.50 0.25 0.50 0.50 0.75 0.25 0.50 0.50 0.75 0.50 0.75 0.75
[16] 1.00

$`n=5`
 [1] 0.0 0.2 0.2 0.4 0.2 0.4 0.4 0.6 0.2 0.4 0.4 0.6 0.4 0.6 0.6 0.8 0.2 0.4 0.4
[20] 0.6 0.4 0.6 0.6 0.8 0.4 0.6 0.6 0.8 0.6 0.8 0.8 1.0

$`n=6`
 [1] 0.0000000 0.1666667 0.1666667 0.3333333 0.1666667 0.3333333 0.3333333
 [8] 0.5000000 0.1666667 0.3333333 0.3333333 0.5000000 0.3333333 0.5000000
[15] 0.5000000 0.6666667 0.1666667 0.3333333 0.3333333 0.5000000 0.3333333
[22] 0.5000000 0.5000000 0.6666667 0.3333333 0.5000000 0.5000000 0.6666667
```

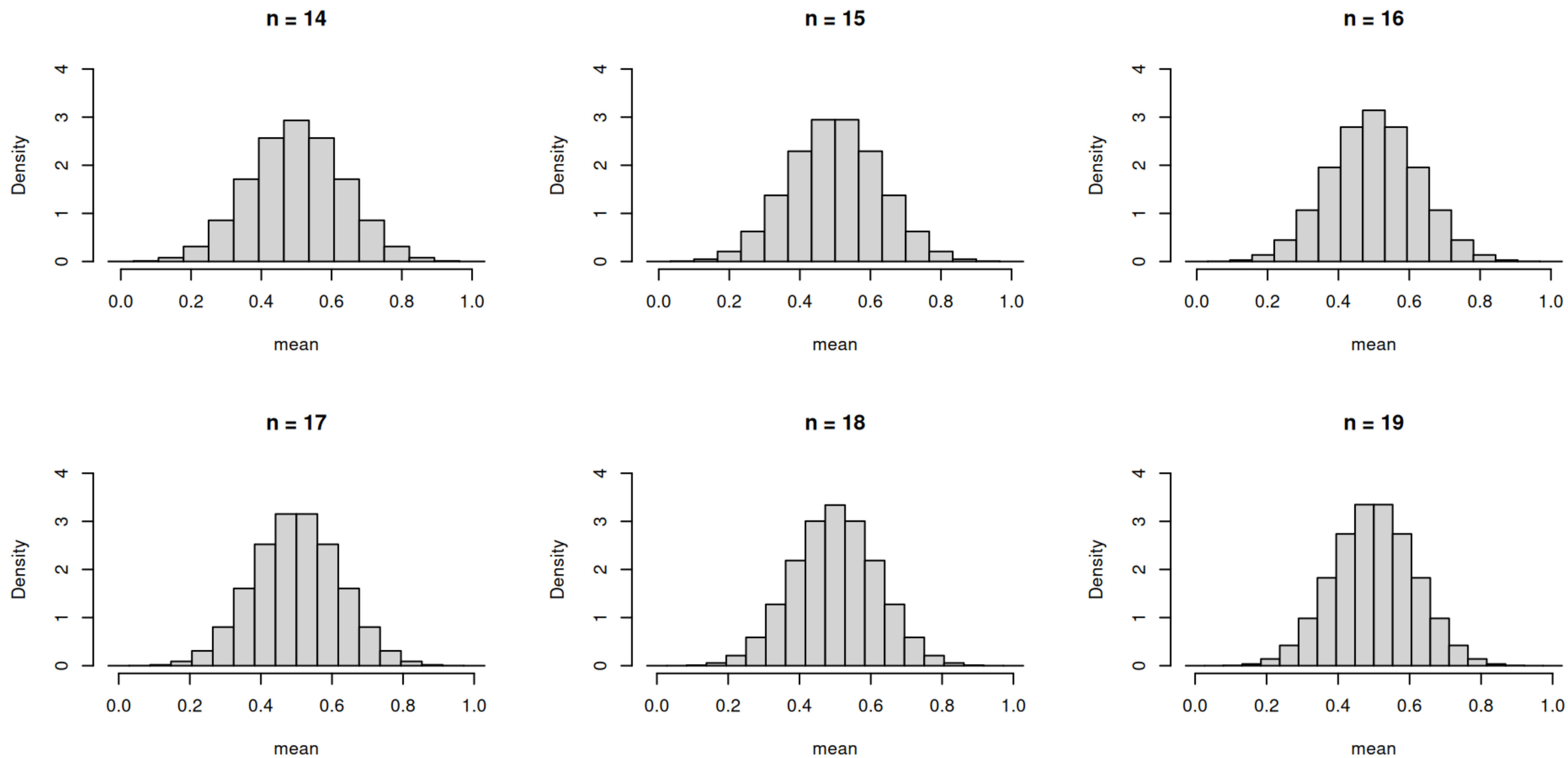
All possible averages for $n = 2, \ldots, 7$



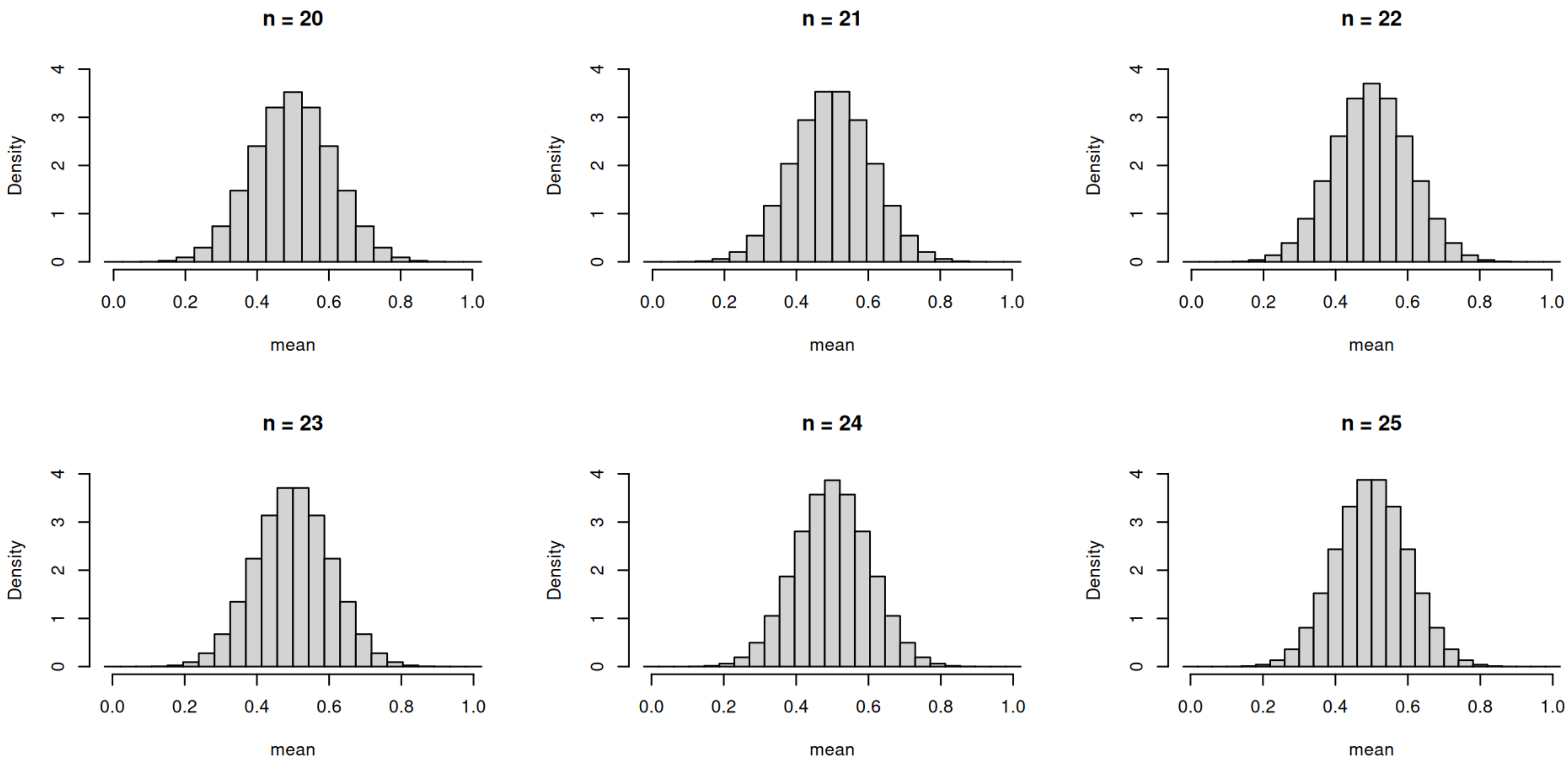
All possible averages for $n = 8, \ldots, 13$



All possible averages for $n = 14, \ldots, 19$



All possible averages for $n = 20, \ldots, 25$



…and so on…

Two important things

- In this example it is very clear that **TWO** important things are happening:
    1. The spread of the distribution of all possible averages/proportions is getting **more concentrated about $\mu = 0.5$ as n increases**;
    2. The shape of the histogram of all possible averages/proportions is becoming "normal-shaped".
- The normal shape means we can approximate probabilities, knowing only $E(\bar{X}) = \mu$ and $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
- Is the "normal shape" due to something special about this particular simple box?
    - ➠ **Not really!**

Example: rolling a 6-sided die

- Suppose we are interested in rolling a 6-sided die n times. How does the sum of the rolls behave?
- This is like taking a random sample of size n from the box

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

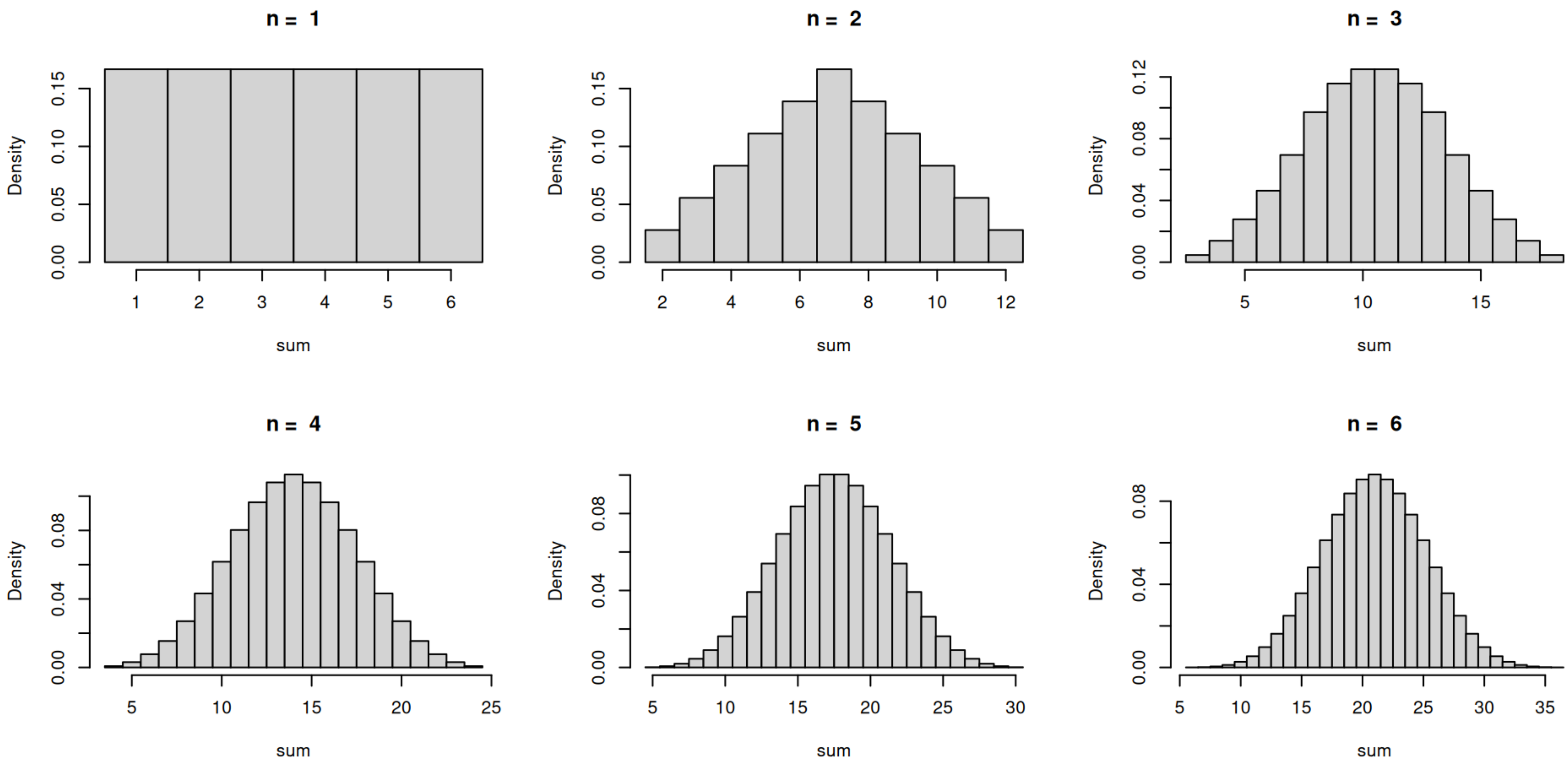
- This box has
  - mean $\mu = 3.5 = \frac{7}{2}$
  - mean square $\frac{1+4+9+16+25+36}{6} = \frac{91}{6}$
  - SD $\sigma = \sqrt{\frac{91}{6} - \left(\frac{7}{2}\right)^2} = \sqrt{\frac{182-(3\times 49)}{12}} = \sqrt{\frac{35}{12}} \approx 1.708$.

```r
box = 1:6
box
```

```
[1] 1 2 3 4 5 6
```

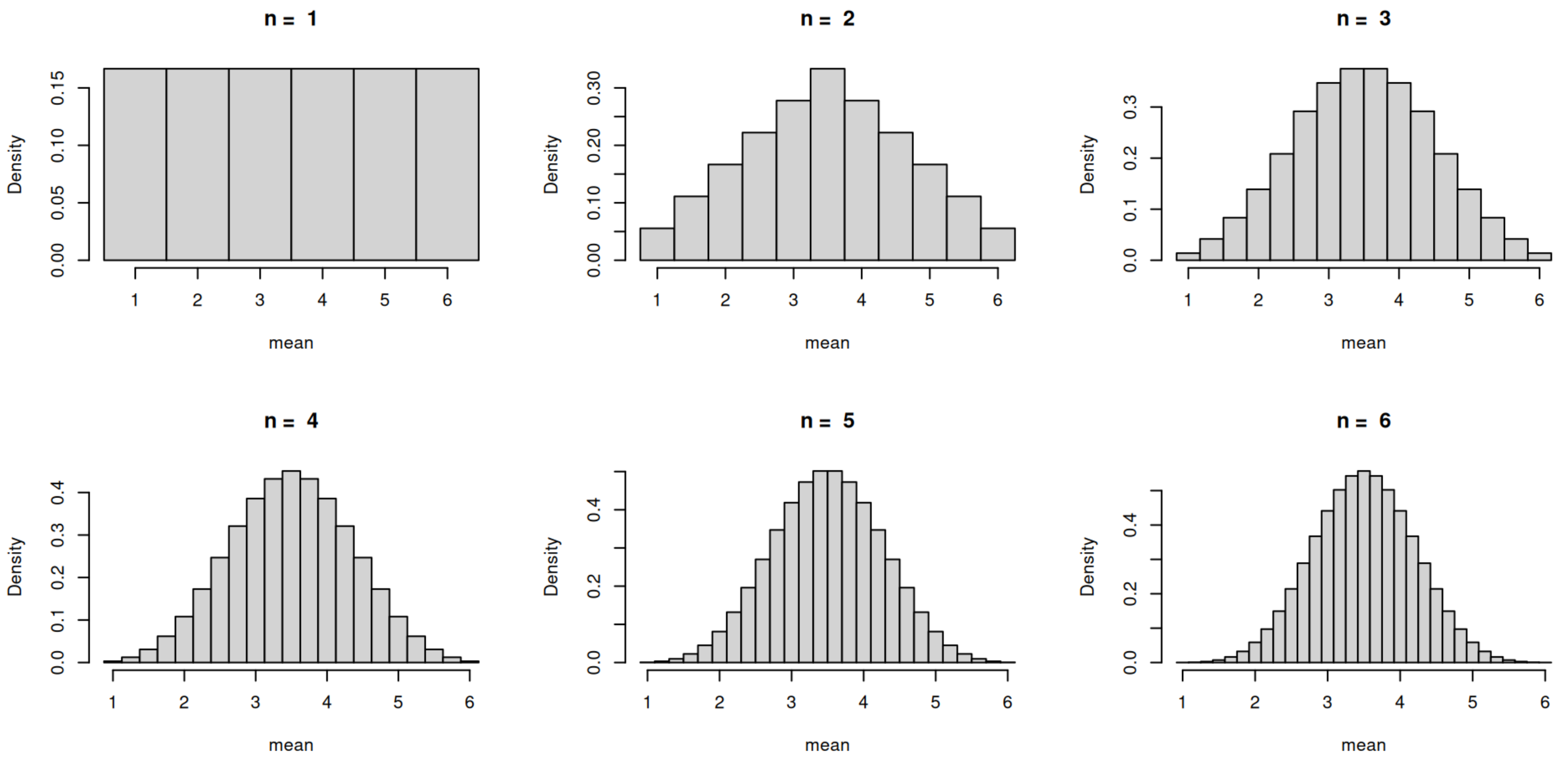
```r
s2 = as.vector(outer(box, box, "+"))
s3 = as.vector(outer(s2, box, "+"))
s4 = as.vector(outer(s3, box, "+"))
s5 = as.vector(outer(s4, box, "+"))
s6 = as.vector(outer(s5, box, "+"))
sums.rolls = list(box, s2, s3, s4, s5, s6)
```

Histograms of all possible sums-of-n-rolls



For $n = 6$ this is normal-shaped too!

Histograms of all possible average-of-n-rolls



Same shape, but different scaling.

Asymmetric example

- Instead of the sum of the rolls, how about the number of $\boxed{6}$s we get out of n rolls?
- the original box for the die

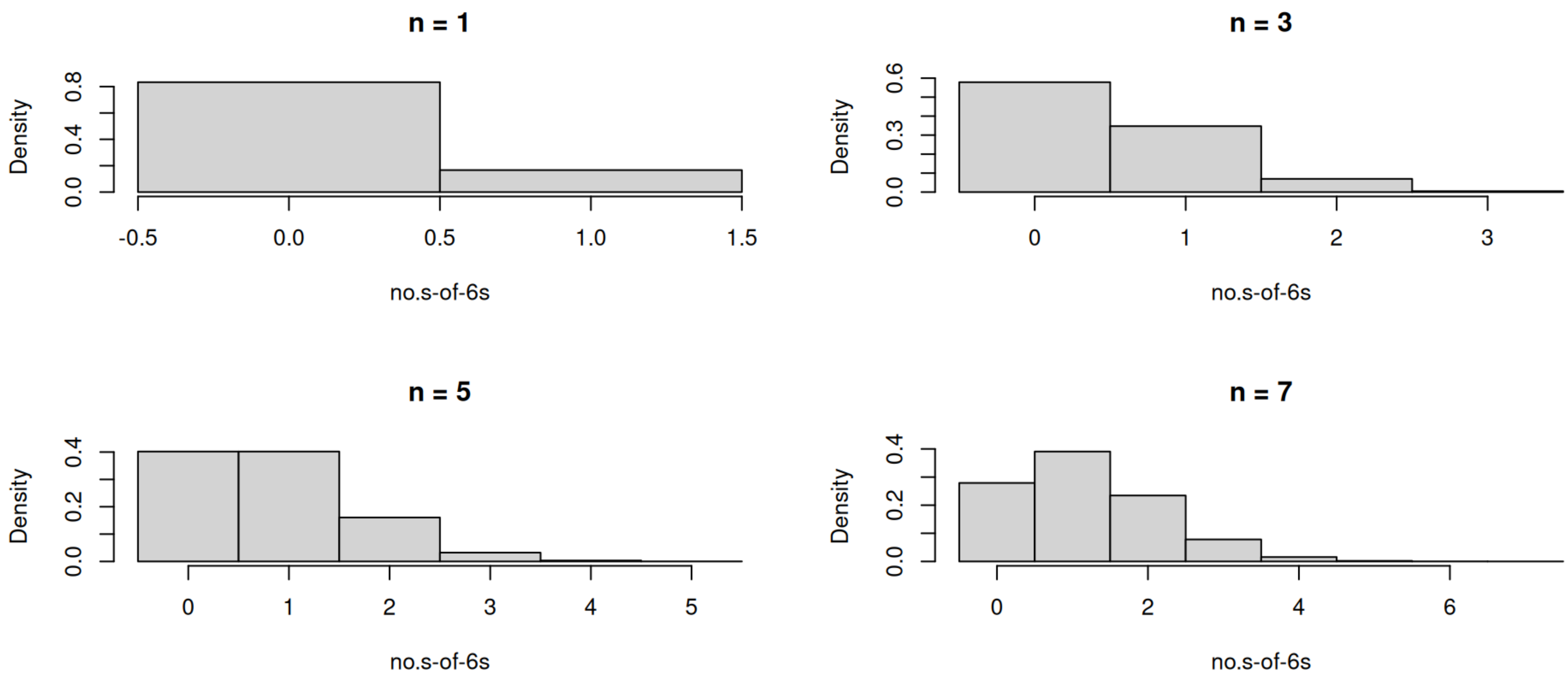
$$\boxed{\boxed{1}\ \boxed{2}\ \boxed{3}\ \boxed{4}\ \boxed{5}\ \boxed{6}}$$

can be converted to a new box representing if we get a $\boxed{6}$ (1) or not (0)

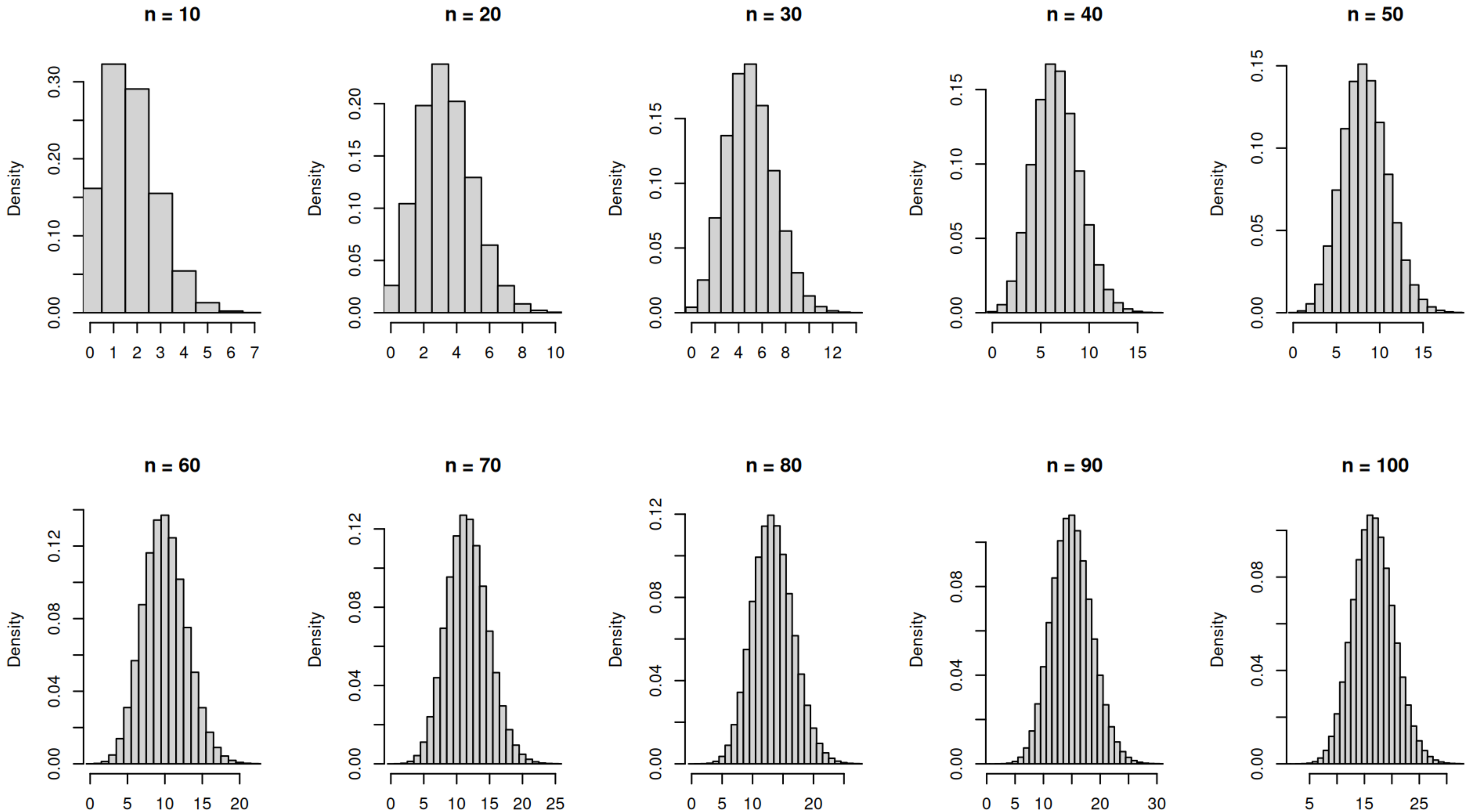
$$\boxed{\boxed{0}\ \boxed{0}\ \boxed{0}\ \boxed{0}\ \boxed{0}\ \boxed{1}}$$

- The number of 6s we get in n rolls is just like the sum S when we take a random sample of size n from this new box.
- This new box has
  - mean $\mu = \frac{1}{6}$
  - mean square $\frac{1}{6}$
  - SD $\sigma = \sqrt{\frac{1}{6} - \left(\frac{1}{6}\right)^2} = \sqrt{\frac{6-1}{36}} = \frac{\sqrt{5}}{6} \approx 0.373$.

Histograms of all possible no.s-of-6s



Not looking very normal-shaped…what about if we let n get larger?


n = 10
n = 20
n = 30
n = 40
n = 50
n = 60
n = 70
n = 80
n = 90
n = 100
Density
0.00 0.10 0.20 0.30
0 1 2 3 4 5 6 7
0.00 0.05 0.10 0.15 0.20
0 2 4 6 8 10
0.00 0.05 0.10 0.15
0 2 4 6 8 12
0.00 0.05 0.10 0.15
0 5 10 15
0.00 0.05 0.10 0.15
0 5 10 15
0.00 0.04 0.08 0.12
0 5 10 15 20
0.00 0.04 0.08 0.12
0 5 10 15 20 25
0.00 0.04 0.08 0.12
0 5 10 20
0.00 0.04 0.08
0 5 10 20 30
0.00 0.04 0.08
5 15 25

We get a normal shape, but only for larger n

- So although the histograms of all possible sums ("no.s-of-times-we-roll-$\boxed{6}$") are not normal-shaped for smaller n, as n increases the shape gets closer to a normal.
- By the time $n > 100$, the shape is quite normal.
- It turns out that for essentially any box, we get the same phenomenon occuring:
  - ➡ as n gets larger and larger, the box of all possible sums gets a "more normal" shape.

The Central Limit Theorem

Most important result in Statistics

- This phenomenon can be *mathematically proven* to hold for any finite box.
- This result is a special case of the **Central Limit Theorem**.
  - ➠ It is a "limit theorem" because it describes what happens "in the limit" as $n \to \infty$.
  - ➠ "Central" here means "most important".
- For the standard normal curve, we have $P(Z < z)$ given in R by `pnorm(z)`.
- A remark: $P(Z < z)$ is often called the CDF of "standard normal" denoted by $\Phi(z)$.

> If $S = X_1 + \cdots + X_n$ is the sum of random sample (with replacement) of size n from a box with mean μ and SD σ, then for **large** n,
>
> $$P(S \leq s) = \underbrace{P\left(\frac{S - n\mu}{\sigma\sqrt{n}} \leq \frac{s - n\mu}{\sigma\sqrt{n}}\right)}_{\text{standard normal}} \approx \Phi\left(\frac{s - n\mu}{\sigma\sqrt{n}}\right)$$

Deconstructing the Central Limit Theorem

- Note that the desired sum value s being considered here, when converted into standard units is

$$z_s = \frac{s - E(S)}{SE(S)} = \frac{s - n\mu}{\sigma\sqrt{n}},$$

which is the ratio inside the $\Phi(\cdot)$.

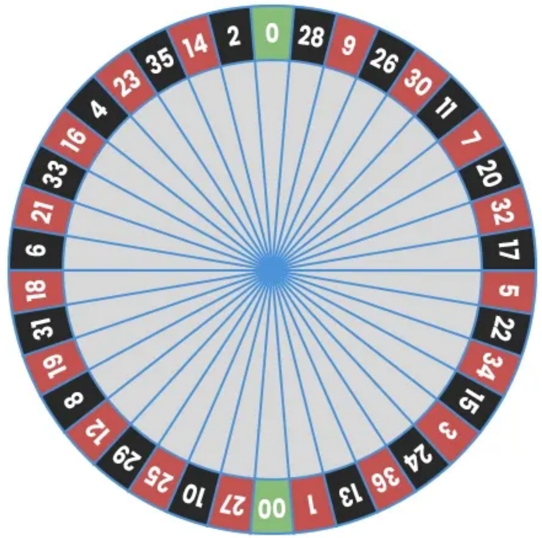
- Therefore, converting to R code, we have

> $P(S \leq s) \approx$ `pnorm(`$(s - n\mu)$`/(`$\sigma\sqrt{n}$`))` $=$ `pnorm(`s`, m =`$n\mu$`, s =`$\sigma\sqrt{n}$`)` .

- The theorem equally applies to the sample mean $\bar{X}$. Let $s = nx$

$$P(\bar{X} \leq x) = \underbrace{P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq \underbrace{\frac{x - \mu}{\sigma/\sqrt{n}}}_{z-\text{score of } x}\right)}_{\text{standard normal} \approx \Phi\left(\frac{x-\mu}{\sigma/\sqrt{n}}\right)} = \Phi\left(\frac{s - n\mu}{\sqrt{n}\sigma}\right)$$

Example: Roulette



- A roulette wheel has slots numbered 1 to 36, plus 2 (or some other number of) slots marked 0.
  - half the positive numbers are coloured black;
  - the remaining positive numbers are coloured red;
  - the zero slots are coloured green (two of them, “0” and “00”).
- If you bet on either “red” or “black”,
  - you double your money if the ball lands in a slot of your colour
  - you lose your money otherwise.
- Suppose each slot is equally likely and a player bets \$1 on “red” for n consecutive spins.

The Roulette Box

- Let S denote the total winnings after n spins. We want to approximate $P(S > 0)$ for $n = 5, 25, 125, 625$.
- There are 38 slots in total, 18 of which are red. If the ball
    - lands in a red slot the player wins \$1;
    - does **not** land in a red slot, the player loses \$1, i.e. they win −\$1.
- Use the following box:

$$\boxed{\underbrace{\boxed{-1}\ \cdots\ \boxed{-1}}_{\text{20 of these}}\ \underbrace{\boxed{+1}\ \cdots\ \boxed{+1}}_{\text{18 of these}}}$$

- mean $\mu = \frac{-2}{38} = -\frac{1}{19}$;
- mean square 1
- SD $\sigma = \sqrt{1 - \left(\frac{1}{19}\right)^2} = \sqrt{\frac{360}{361}} \approx 0.9986$.

Exact answers

- It is possible to work out the exact probabilities (using the “binomial distribution”, more on this later if we have time).
- These are

```r
n = c(5, 25, 125, 625)
prob.win = 1 - pbinom(n/2, n, 18/38)
rbind(n, prob.win)
```

```
              [,1]       [,2]        [,3]         [,4]
n        5.0000000 25.0000000 125.0000000 625.00000000
prob.win 0.4507489  0.3951246   0.2775865   0.09388094
```

Normal approximation

- According to the Central Limit Theorem, for "large n",

$$P(S>0)=1-P(S\leq 0)=1-P(S\leq 0)\approx 1-\Phi\left(z_0\right)=1-\texttt{pnorm}\left(\frac{\sqrt{361n}}{19\sqrt{360}}\right)$$

where z_0 is the z-score of 0

$$z_0=\frac{0-n\mu}{\sqrt{n}\sigma}=\frac{0-\left(-\frac{n}{19}\right)}{\sqrt{\frac{360n}{361}}}=\frac{\sqrt{361n}}{19\sqrt{360}}$$

- This gives

```r
1 - pnorm(sqrt(361 * n)/(19 * sqrt(360)))
```

```
[1] 0.45309281 0.39607370 0.27784490 0.09381616
```

- These are quite good approximations (even for $n=5$)!
- Makes sense, because the box is reasonably symmetric (not that different in shape to Kerrich's box).

Final comments

When we take a random sample of size n (with replacement) from a box with mean μ and SD σ, the box of all possible sums

- Has mean equal to $E(S) = n\mu$;
- Has SD equal to $SE(S) = \sigma\sqrt{n}$;
- Is (approx.) normal-shaped for "large enough n".

For such n we can approximate probabilities for the random sum S or average $\bar{X} = S/n$, using `pnorm()`.

- We don't need to know the exact contents of the box, as long as we have $E(X) = \mu$ and $SE(X) = \sigma$

How large is "large enough n"? **It depends** on the original box. If the original box is

- Reasonably symmetric (without too many outliers), $n = 5$ or 10 may do;
- Very skewed, we may need $n > 100$ before the box of all possible sums has a nice, symmetric normal shape.