T6: Understanding Chance
T7: Chance Variability (The Box Model)
T8: Sample Surveys

LO6: Use the box model to describe chance & chance variability, including sample surveys & the CLT.

Tossing a coin in the long run

$$\#\text{Heads} = \frac{1}{2}\#\text{Tosses} + \text{Chance Error}$$

(observed value) (expected value)

$$OV = EV + CE$$

① Common mistake = Gambler's Fallacy

As # of tosses ↑, the size of CE ↓

② Correct thinking = Law of Large Numbers (LLN)

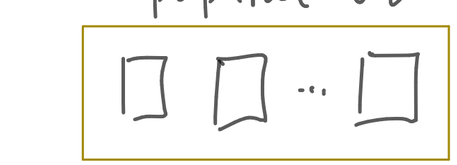
As # of tosses ↑, % Heads $\xrightarrow{\text{stabilises}}$ P(Head)

The Central Limit Theorem (CLT)

When drawing at random from <u>any</u> box, if the sample size is sufficiently large, then

the shape of the Sum or Mean of the Sample is approximately <u>Normal</u>.

The Box Model



← represented by "tickets"
need to know
- what values
- how many of each

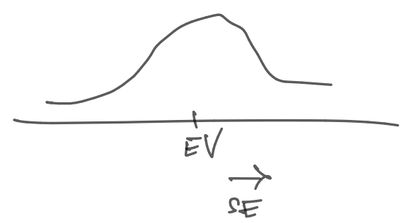
① We want to model the Sum or Mean of an (unknown) Sample.

② $OV = EV + CE$ ← use the Box Model with CE ≒ SE

calculate from a know sample OR simulate samples from the population

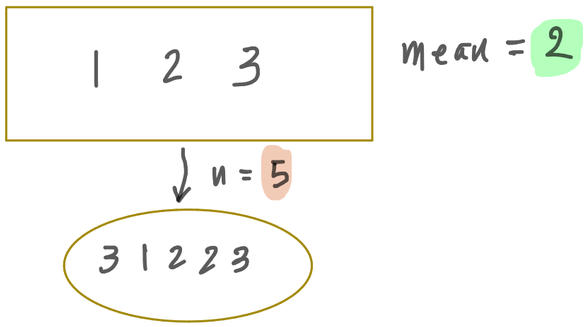
③ Use the CLT.

Eg/ Sample Sum ~ $N(EV, SE^2)$



Examples of the Box Model

① DV22 p23



Consider the Sum of the Sample:

- EV = 5 * 2 = 10
- OV = 3 + 1 + 2 + 2 + 3 = 11

So CE = OV − EV = 11 − 10 = 1

② DV22 p 29

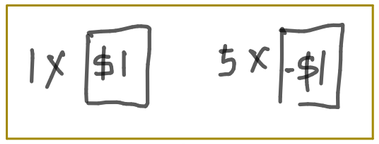
It costs $1 to play a game

If you roll a 6
- you get $1 back
- plus an extra $1

If you don't roll a 6
- you lose your $1

You play 25 times. What is your expected win/loss?

1 game



Consider the Sum of the Sample

- EV = 25 * $-\frac{4}{6}$ = −$16.67
- OV = −13

So CE = OV − EV = $3.67

simulation

```
set.seed(1)
die tosses = sample(c(1,-1,-1,-1,-1,-1), 25, rep=T)
sum(die tosses)
```

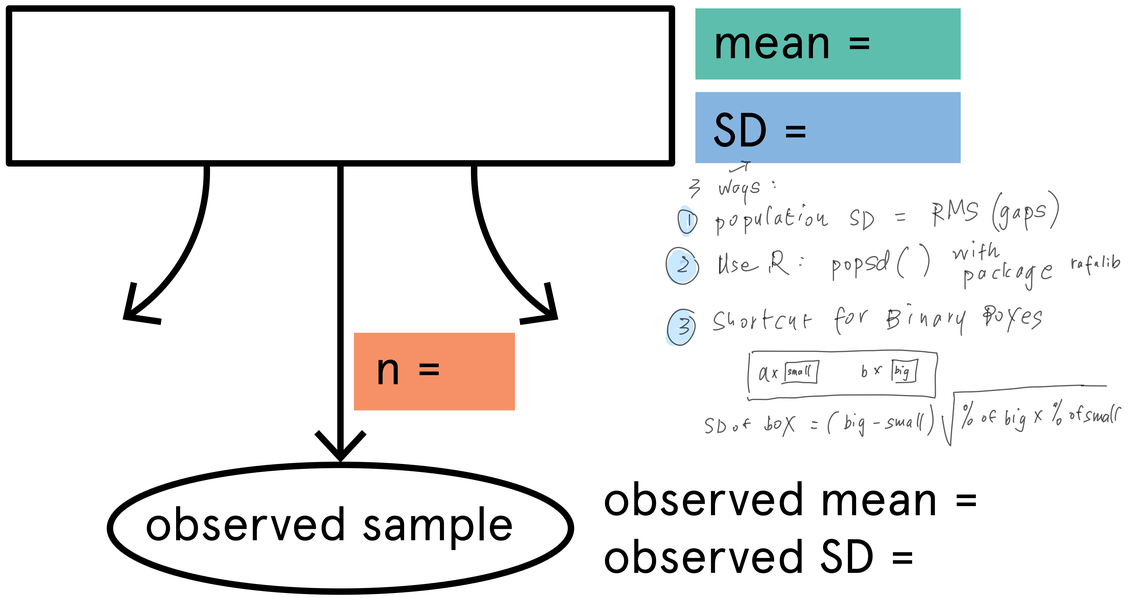
output is -13

Summary of the Box Model

Eg/ 0 1 mean = 0.5, SD = 0.5
① SD = $\sqrt{\frac{(0-0.5)^2 + (1-0.5)^2}{2}}$
② box = c(0,1) library(rafalib) popsd(box)
③ $(1-0)\sqrt{0.5 \times 0.5} = 0.5$

The Box Model

population = "box" = ...



3 ways:
① population SD = RMS (gaps)
② Use R: popsd( ) with package rafalib
③ Shortcut for Binary Boxes

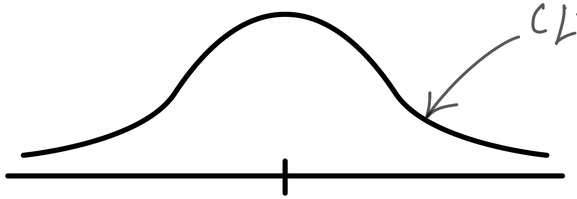
a × small b × big

SD of box = (big − small) $\sqrt{\%\text{ of big} \times \%\text{ of small}}$

observed mean =
observed SD =

Model of Sample

Sample Sum	EV = n mean	
	SE = $\sqrt{n}$ SD	
Sample Mean	EV = mean	
	SE = $\frac{SD}{\sqrt{n}}$	



Sample Sum / Mean

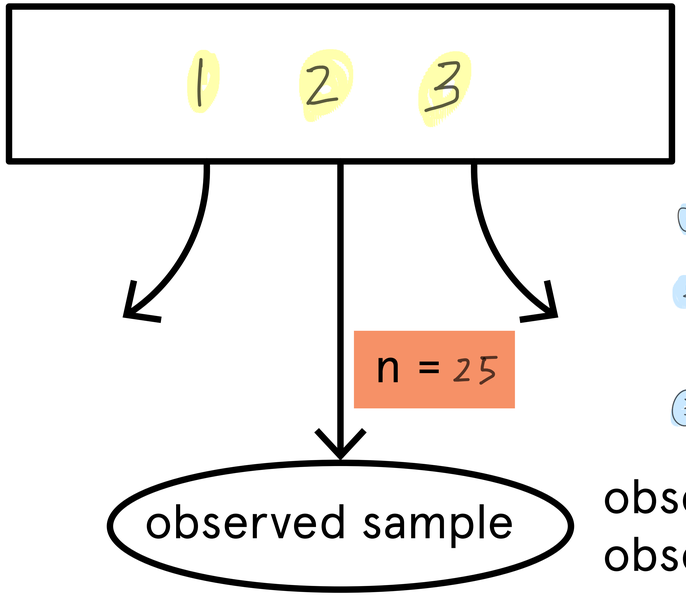
N(EV , SE)

Summary of the Box Model

7.3 Challenge
Scenario: Box 1

The Box Model

population = "box" = ... Box 1



mean = 2

SD ≈ 0.82

① SD = RMS(gaps) = $\sqrt{\frac{(1-2)^2+(2-2)^2+(3-2)^2}{3}}$ = $\sqrt{\frac{2}{3}}$ ≈ 0.816

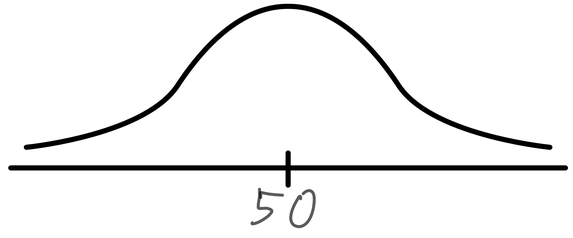
② box1 = c(1,2,3)
library (rafalib)
popsd(box1)

③ not Binary Box

observed mean =
observed SD =

Model of Sample

Sample Sum	EV = n mean	25 × 2 = 50
	SE = √n SD	$\sqrt{25} \times \sqrt{\frac{2}{3}} \approx 4.1$
Sample Mean	EV = mean	
	SE = $\frac{SD}{\sqrt{n}}$	



Sample Sum / Mean

N(50 , 4.1^2)

CLT

Summary of the Box Model

7.3 Challenge
Scenario: Box 2

The Box Model

population = "box" = ... Box 2

mean = 3.5

SD ≈ 1.71

① SD = RMS(gaps) = $\sqrt{\frac{(1-3.5)^2+\dots+(6-3.5)^2}{6}}$ ≈ 1.71

② box 2 = 1:6
library (rafalib)
popsd(box2)

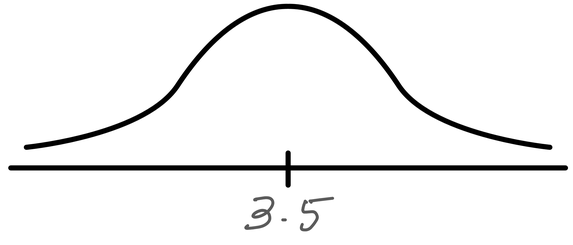
n = 16

observed sample

observed mean =
observed SD =

Model of Sample

Sample Sum	EV = n mean	
	SE = √n SD	
Sample Mean	EV = mean	3.5
	SE = $\frac{SD}{\sqrt{n}}$	$\frac{1.71}{\sqrt{16}} \approx 0.43$



Sample Sum / Mean

$N(3.5, 0.43^2)$

CLT

Continuity Correction

When we apply the Box Model:
As the box is discrete,
then we can (often) improve
the use of the Normal approximation
by slightly adjusting the thresholds.

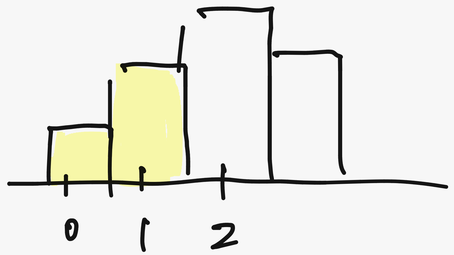
Example:

Suppose the Sample Sum $\dot{\sim} N(2, 3^2)$
(discrete) ← continuous

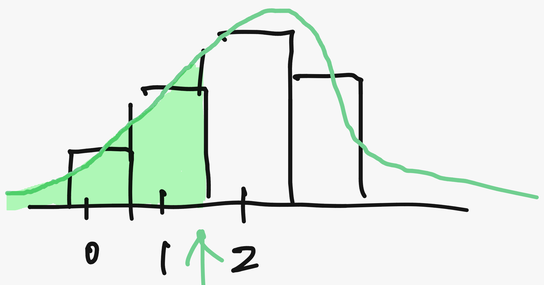
Find $P(X \leq 1)$

① Exact

$P(X \leq 1)$ →



② Approximation



Hence we work out

$P(X \leq 1.5)$ for the Normal curve.

> pnorm(1.5, 2, 3)