Probability - Part 2

Jan. 30, 2025

Probability - Part 2

Jan. 30, 2025

By the end of this lecture, you will be able to:

1. Use Bayes' rule to compute the probability that you are infected with COVID, after getting a positive test result
2. Compute the probability of getting k heads when flipping n coins, using Pascal's triangle
3. Define the normal distribution
4. Give examples of the central limit theorem

Recap question:

Jan. 30, 2025

In which of these cases are the two events A,B independent?

1. You draw two cards and **throw away** the first drawn card before drawing the second. A = the first is a queen, B = the second is a queen
2. You draw two cards and **put back** the first one and shuffle the deck before drawing the second. A = the first is a queen, B = the second is a queen
3. You flip 3 coins. A = the first is heads, B = the second is tails

Recap question:

Jan. 30, 2025

In which of these cases are the two events A,B independent?

1. Not independent! If we draw a queen in the first draw, the probability of drawing a queen in the second draw changes compared to if we don't draw the first card
2. Independent! If we put back the card and shuffle it (without cheating) then the first draw does not influence the second one
3. Independent! The coin has no memory so the first flip does not influence the future

Example 1: Consider one COVID test with 84% sensitivity and 99% specificity. (I = infected, nI = not infected)

P(test positive | I) = 0.84	True positive
P(test negative | I) = 0.16	False negative
P(test negative | nI) = 0.99	True negative
P(test positive | nI) = 0.01	False positive

Assume that 10% of the population is infected, and that you receive a positive result on your test. What is the probability that you are infected?

Example 1: Consider one COVID test with 84% sensitivity and 99% specificity. (I = infected, nI = not infected)

P(test positive | I) = 0.84	True positive
P(test negative | I) = 0.16	False negative
P(test negative | nI) = 0.99	True negative
P(test positive | nI) = 0.01	False positive

Assume that 10% of the population is infected, and that you receive a positive result on your test. What is the probability that you are infected?

P(I | test positive)

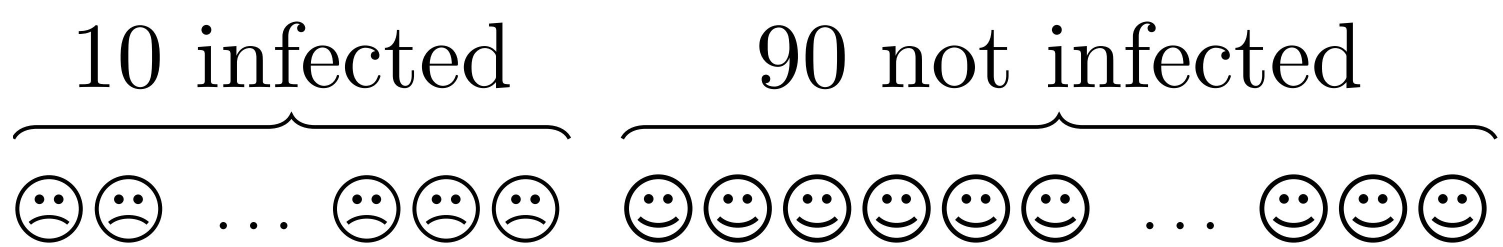
Example 1:

P(test positive | I) = 0.84 — True positive

P(test negative | I) = 0.16 — False negative

P(test negative | nI) = 0.99 — True negative

P(test positive | nI) = 0.01 — False positive



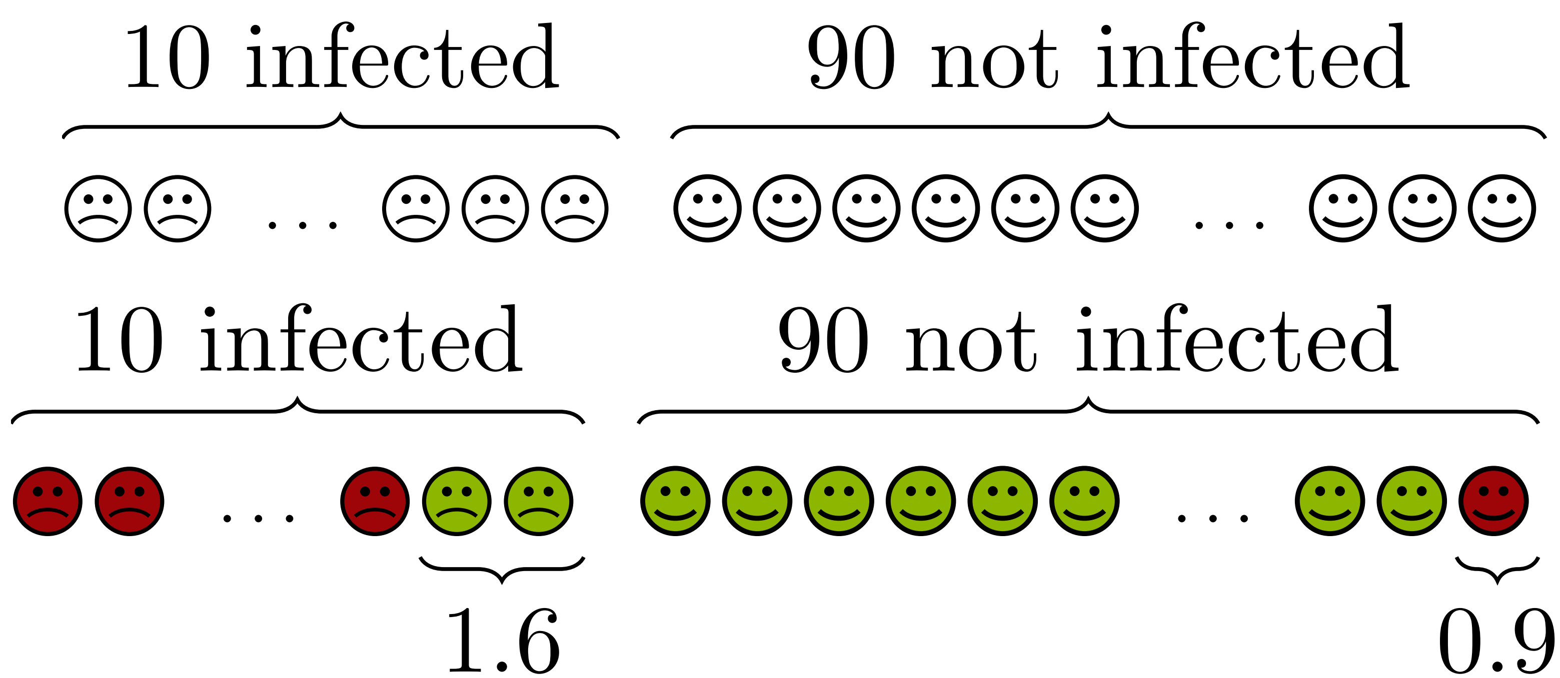
Example 1:

P(test positive | I) = 0.84 True positive

P(test negative | I) = 0.16 False negative

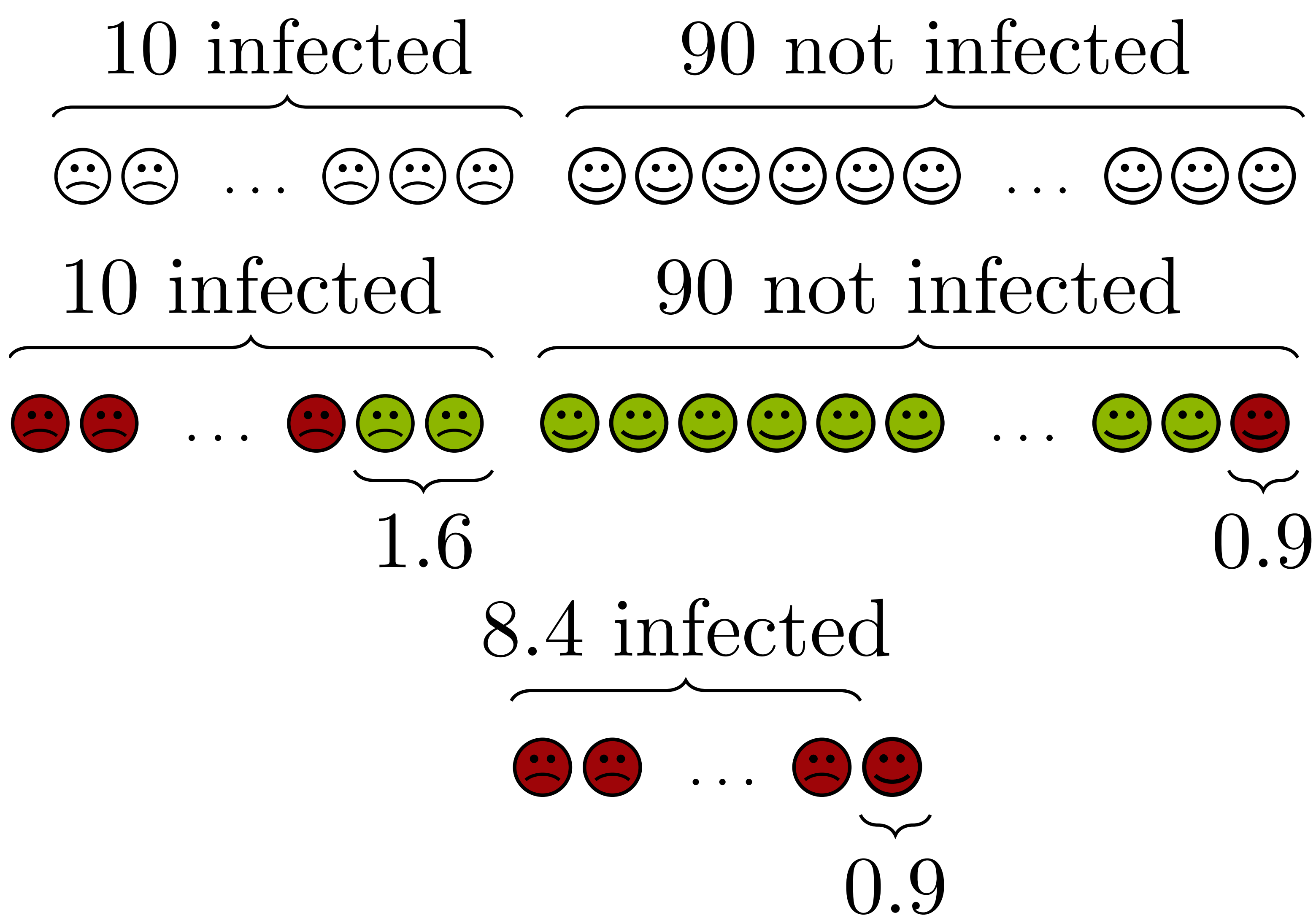
P(test negative | nI) = 0.99 True negative

P(test positive | nI) = 0.01 False positive



Example 1:

P(test positive | I) = 0.84	True positive
P(test negative | I) = 0.16	False negative
P(test negative | nI) = 0.99	True negative
P(test positive | nI) = 0.01	False positive



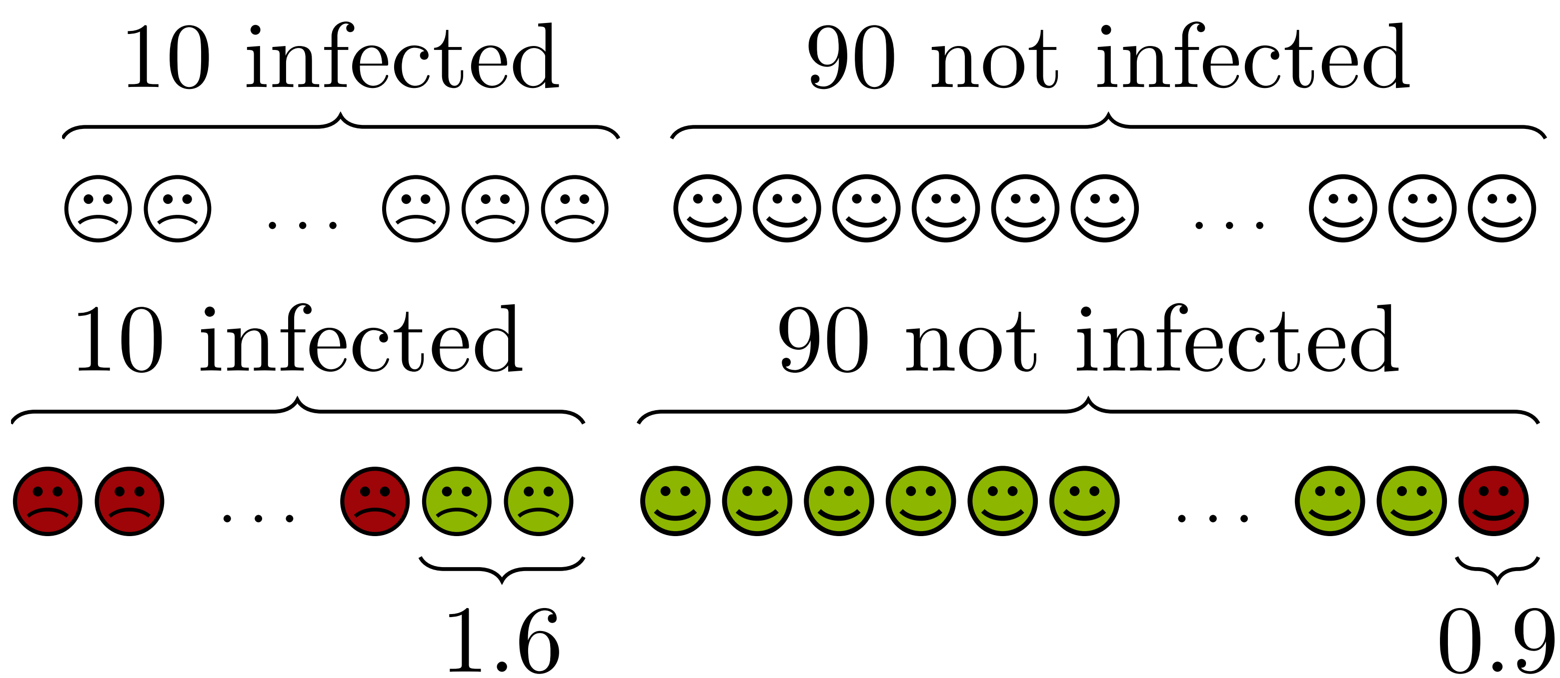
Example 1:

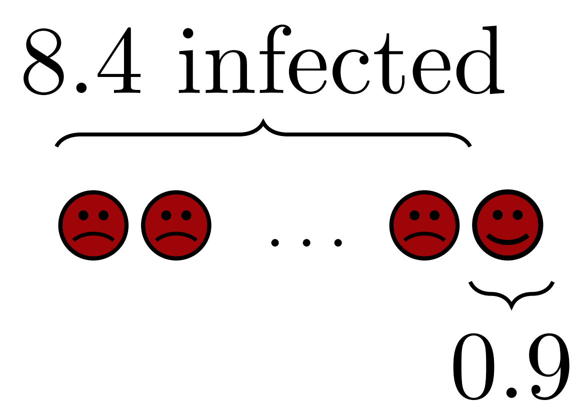
P(test positive | I) = 0.84 True positive

P(test negative | I) = 0.16 False negative

P(test negative | nI) = 0.99 True negative

P(test positive | nI) = 0.01 False positive





$$P(I|\text{test positive}) = \frac{8.4}{8.4 + 0.9} \approx 0.9$$

Bayes' rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(I|\text{positive}) = \frac{P(\text{positive}|I)P(I)}{P(\text{positive})}$$

Bayes' rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(I|\text{positive}) = \frac{P(\text{positive}|I)P(I)}{P(\text{positive})}$$

$$= \frac{P(\text{positive}|I)P(I)}{P(\text{positive and } I) + P(\text{positive and not } I)}$$

Bayes′ rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(I|\text{positive}) = \frac{P(\text{positive}|I)P(I)}{P(\text{positive})}$$

$$= \frac{P(\text{positive}|I)P(I)}{P(\text{positive and } I) + P(\text{positive and not } I)}$$

$$=$$

$$P(A \text{ and } B) = P(B|A)P(A)$$

Bayes' rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(I|\text{positive}) = \frac{P(\text{positive}|I)P(I)}{P(\text{positive})}$$

$$= \frac{P(\text{positive}|I)P(I)}{P(\text{positive and } I) + P(\text{positive and not } I)}$$

$$= \frac{P(\text{positive}|I)P(I)}{P(\text{positive}|I)P(I) + P(\text{positive}|nI)P(nI)}$$

$$P(A \text{ and } B) = P(B|A)P(A)$$

Bayes' rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(I|\text{positive}) = \frac{P(\text{positive}|I)P(I)}{P(\text{positive})}$$

$$= \frac{P(\text{positive}|I)P(I)}{P(\text{positive and } I) + P(\text{positive and not } I)}$$

$$= \frac{P(\text{positive}|I)P(I)}{P(\text{positive}|I)P(I) + P(\text{positive}|nI)P(nI)}$$

$$= \frac{0.84 \times 0.1}{0.84 \times 0.1 + 0.01 \times 0.9} \approx 0.9$$

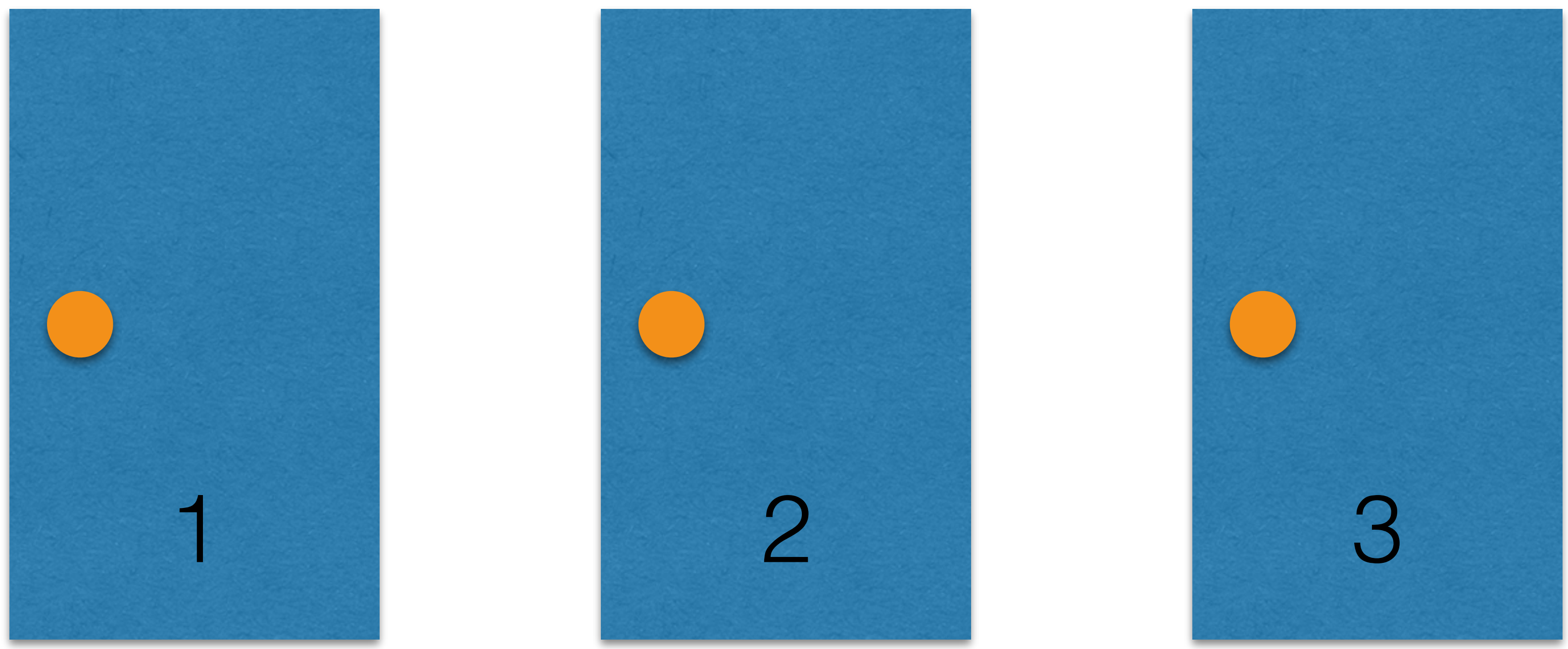
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(I|\text{negative}) = \frac{P(\text{negative}|I)P(I)}{P(\text{negative})}$$

$$= \frac{P(\text{negative}|I)P(I)}{P(\text{negative}|I)P(I) + P(\text{negative}|nI)P(nI)}$$

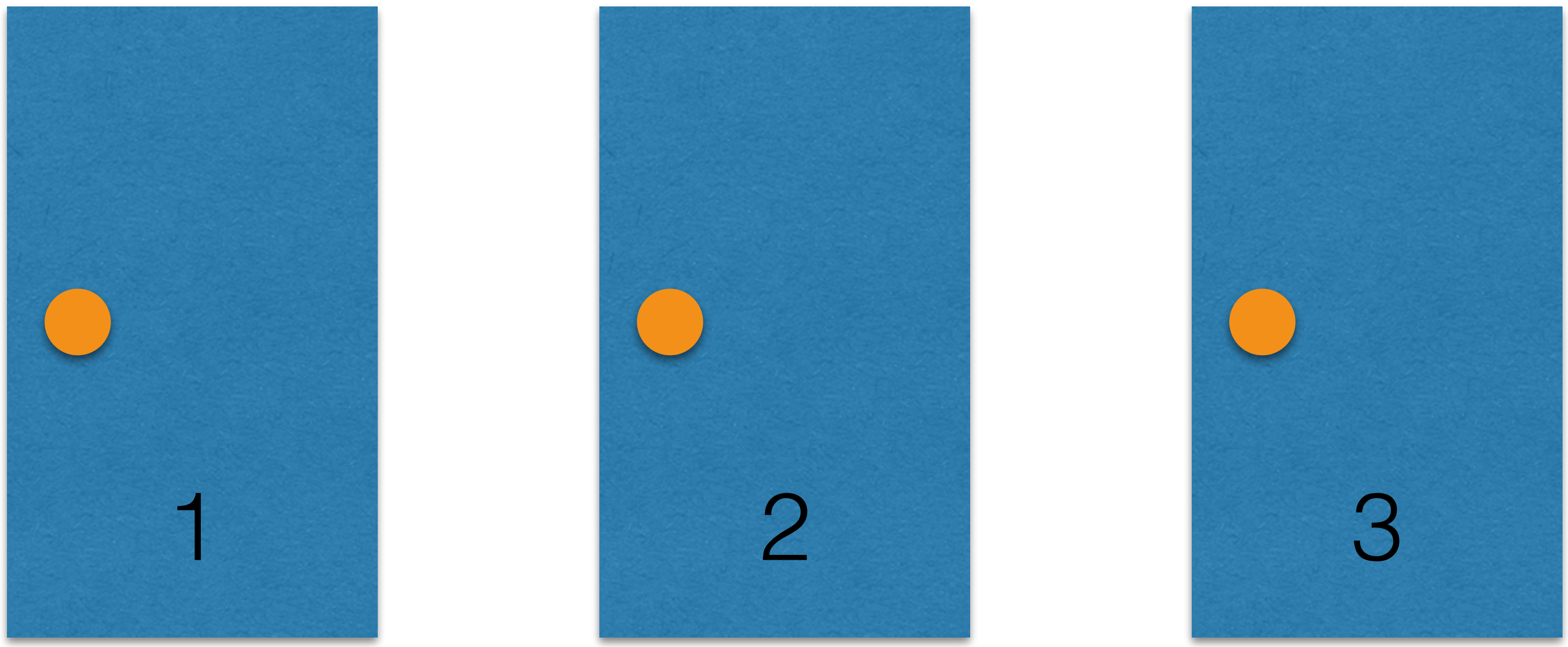
$$= \frac{0.16 \times 0.1}{0.16 \times 0.1 + 0.99 \times 0.9} \approx 0.018$$

Example 2: Monty Hall problem



Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car and behind the other doors is a goat. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat behind. The host then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

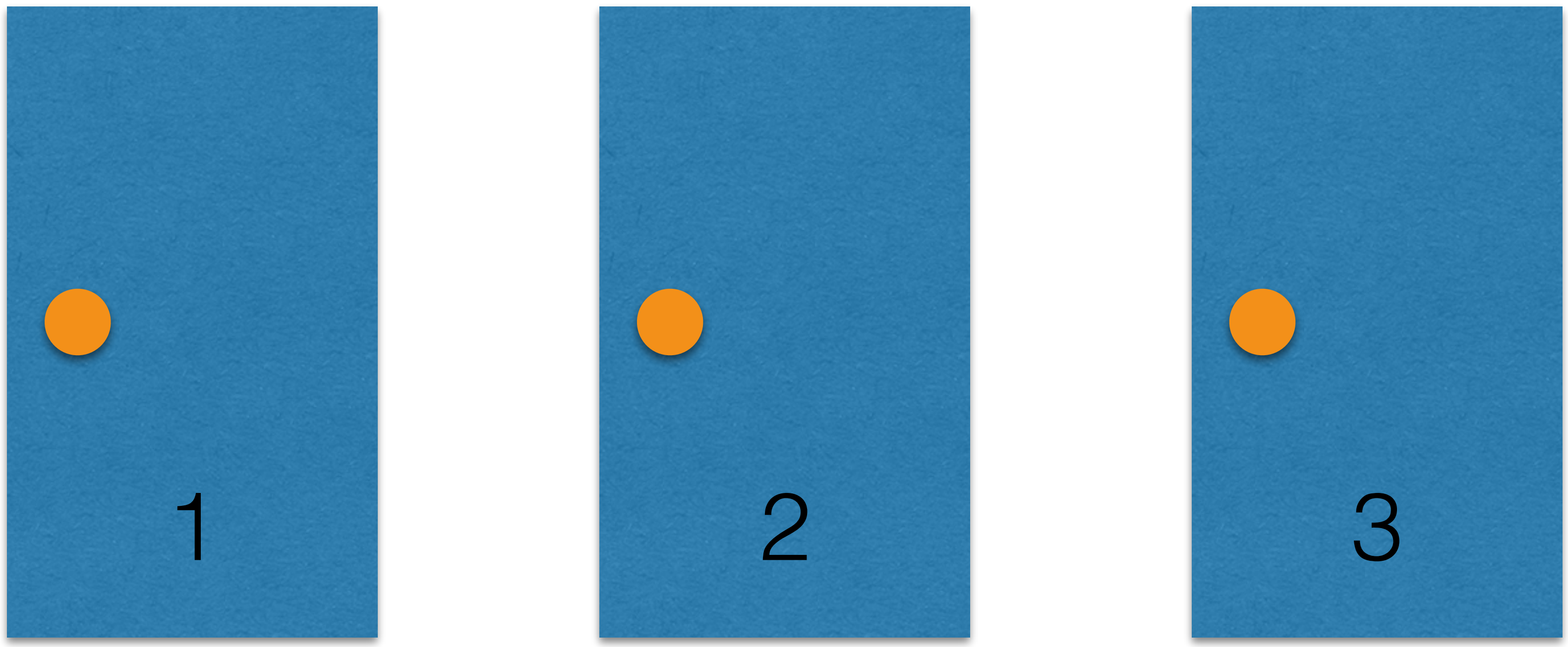
Example 2: Monty Hall problem



With 1 million trials:

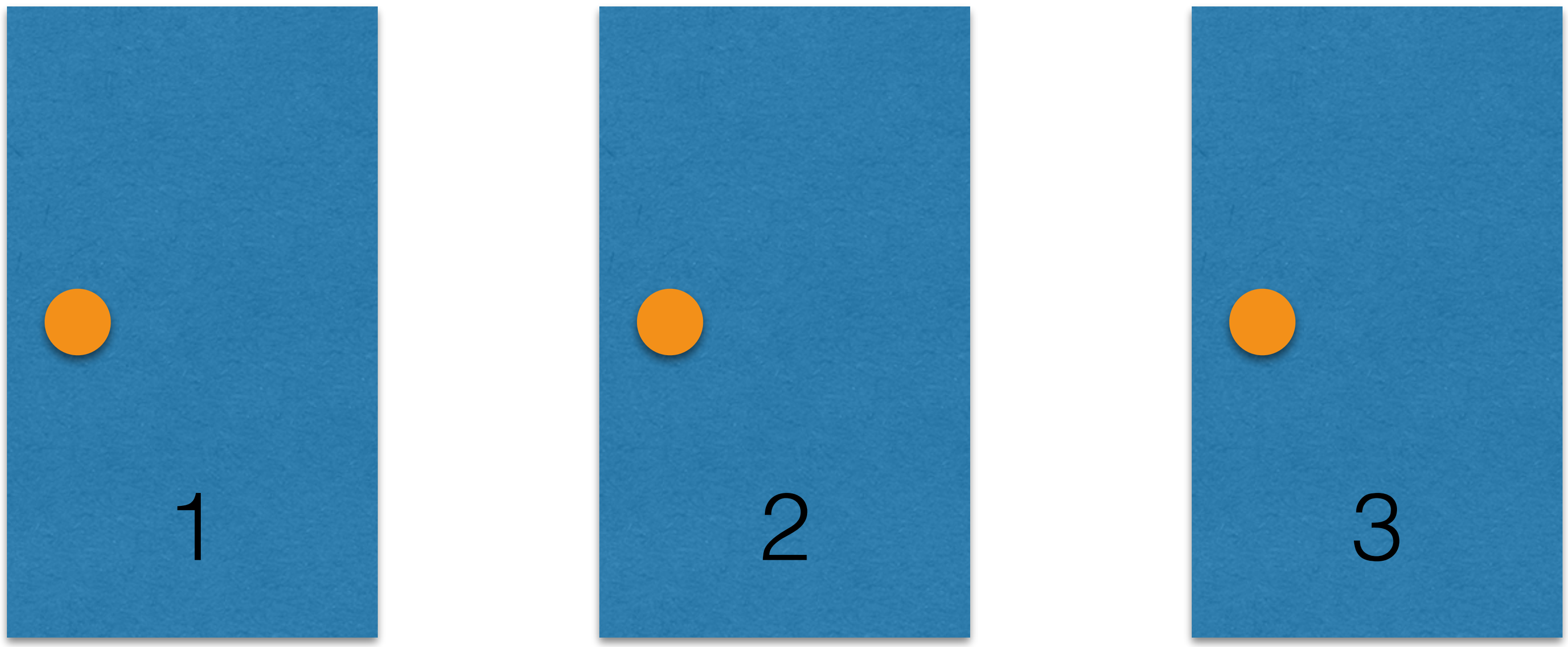
	# Goats	# Cars	Win probability
Stay	666446	333554	0.333
Switch	332870	667130	0.667

Example 2: Monty Hall problem



Behind door 1	Behind door 2	Behind door 3	Result of staying at door 1	Result if switching doors
Goat	Goat	**Car**	Goat	**Car**
Goat	**Car**	Goat	Goat	**Car**
Car	Goat	Goat	**Car**	Goat

Example 2: Monty Hall problem



Behind door 1	Behind door 2	Behind door 3	Result of staying at door 1	Result if switching doors
Goat	Goat	**Car**	Goat	**Car**
Goat	**Car**	Goat	Goat	**Car**
Car	Goat	Goat	**Car**	Goat

Switching wins the car 2/3 of the time!

The Normal Distribution

What does the weight of penguins in Antarctica have to do with average SAT scores at NC high schools in 2018?

What does the weight of penguins in Antarctica have to do with average SAT scores at NC high schools in 2018?

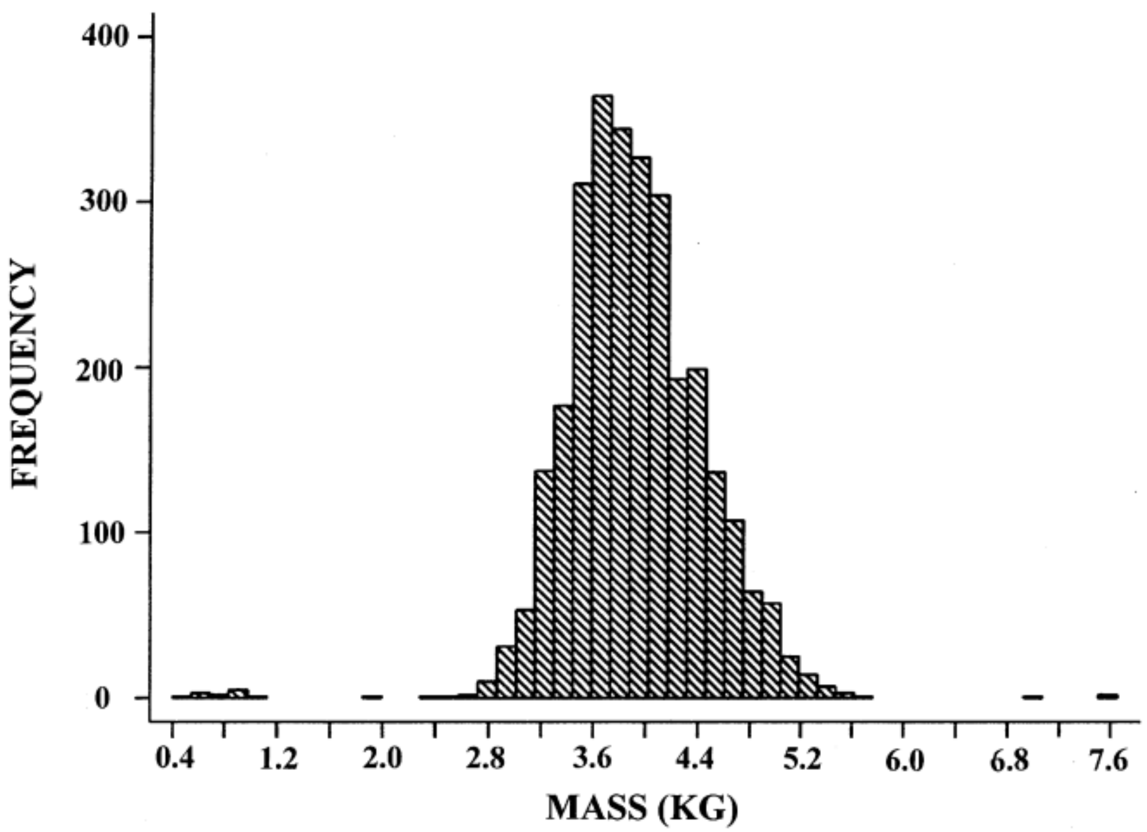
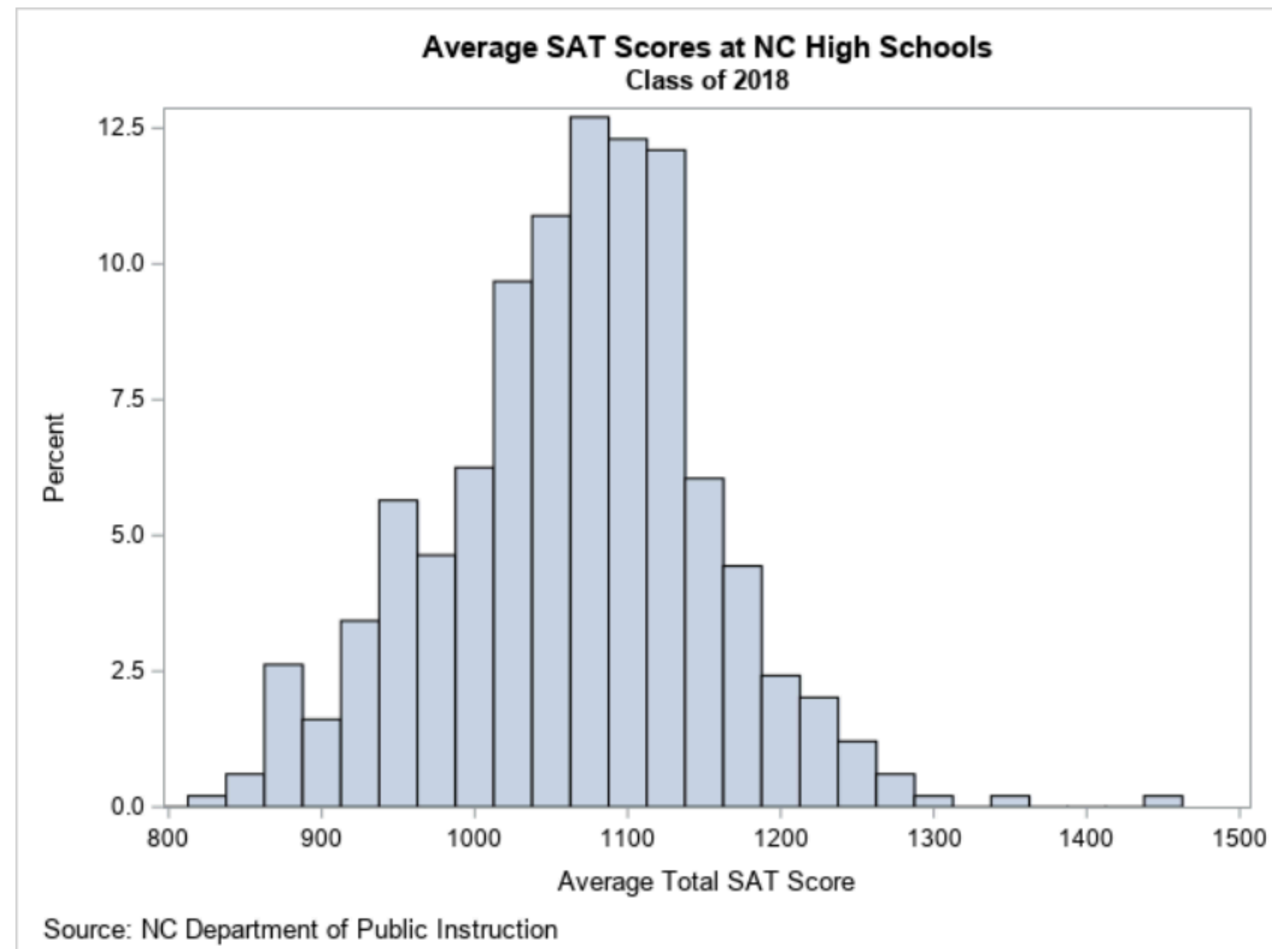
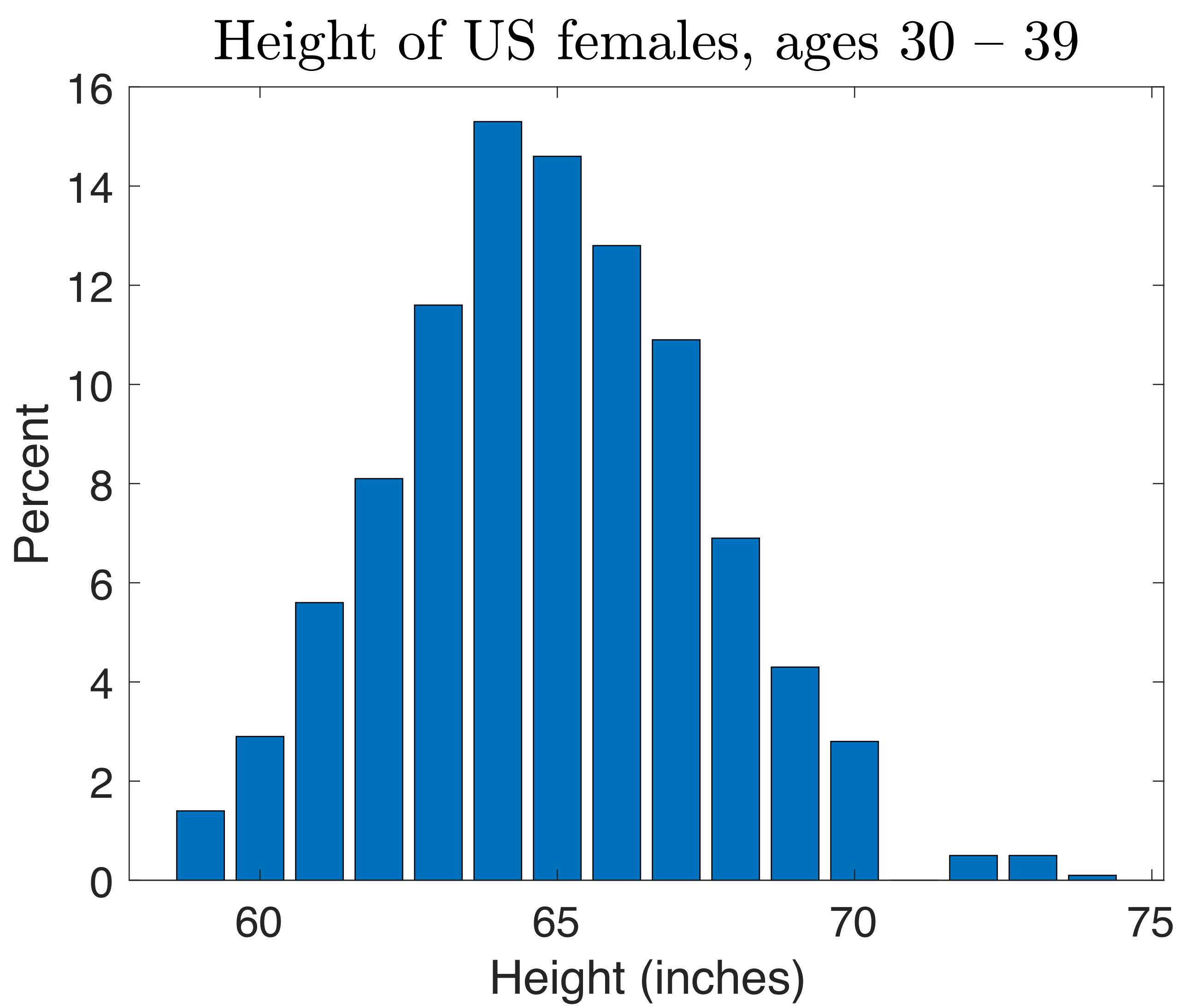


Fig. 2 A frequency histogram showing the masses recorded by the automatic scale, with 5.1 selected as the maximum mass allowed to represent one penguin

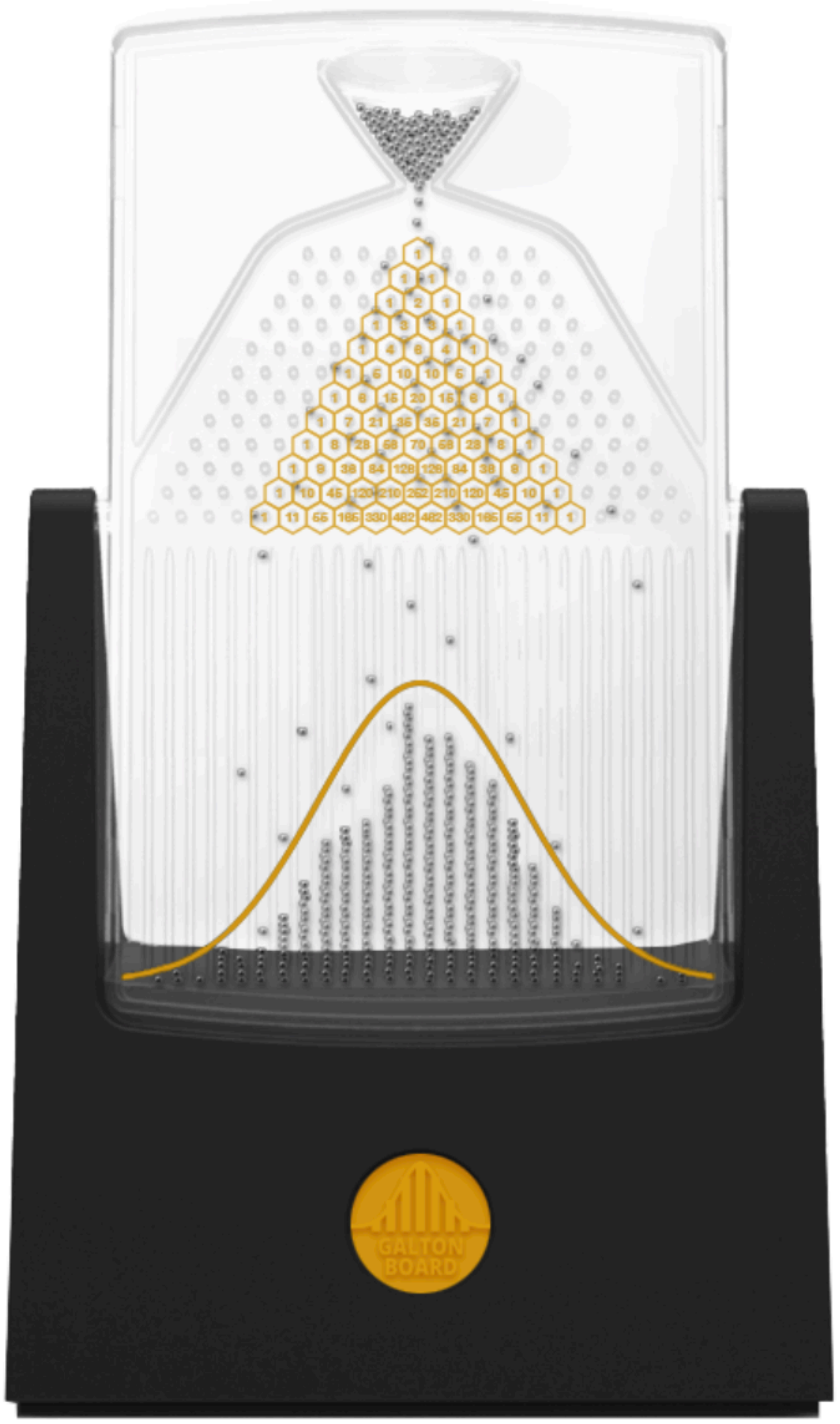


Ainley et al Polar Biol (1998) 20: 311–319

… how about heights in the US?



Galton board



https://commons.wikimedia.org/wiki/File:GaltonBoard.png

For the rest of the class: explain where this phenomenon comes from, using coin tosses as example

Suppose you toss a coin 100 times, and you count the number of heads obtained, what is the probability to get 60 heads?

$$P(A) = \frac{\#\text{ of outcomes in A}}{\#\text{ of total possibilities}}$$

Suppose you toss a coin 100 times, and you count the number of heads obtained, what is the probability to get 60 heads?

$$P(A) = \frac{\#\text{ of outcomes in A}}{\#\text{ of total possibilities}}$$

$$\#\text{ of total possibilities} = \quad ?$$

Flipping 1 coins gives 2 possibilities

Suppose you toss a coin 100 times, and you count the number of heads obtained, what is the probability to get 60 heads?

$$P(A) = \frac{\text{\# of outcomes in A}}{\text{\# of total possibilities}}$$

$$\text{\# of total possibilities} = \quad ?$$

Flipping 1 coins gives 2 possibilities

Flipping 2 coins gives $2^2 = 4$ possibilities

Suppose you toss a coin 100 times, and you count the number of heads obtained, what is the probability to get 60 heads?

$$P(A) = \frac{\#\text{ of outcomes in A}}{\#\text{ of total possibilities}}$$

$$\#\text{ of total possibilities} = \quad ?$$

Flipping 1 coins gives 2 possibilities

Flipping 2 coins gives $2^2 = 4$ possibilities

$$\vdots$$

Flipping 100 coins gives 2^{100} possibilities

Suppose you toss a coin 100 times, and you count the number of heads obtained, what is the probability to get 60 heads?

$$P(A) = \frac{\text{\# of outcomes in A}}{\text{\# of total possibilities}}$$

$$\text{\# of total possibilities} = 2^{100}$$

Suppose you toss a coin 100 times, and you count the number of heads obtained, what is the probability to get 60 heads?

$$P(A) = \frac{\text{\# of outcomes in A}}{\text{\# of total possibilities}}$$

$$\text{\# of total possibilities} = 2^{100}$$

$$\text{\# of outcomes with 60 heads} = ?$$

Suppose you toss a coin 100 times, and you count the number of heads obtained, what is the probability to get 60 heads?

$$P(A) = \frac{\text{\# of outcomes in A}}{\text{\# of total possibilities}}$$

$$\text{\# of total possibilities} = 2^{100}$$

$$\text{\# of outcomes with 60 heads} = \binom{100}{60}$$

Suppose you toss a coin 100 times, and you count the number of heads obtained, what is the probability to get 60 heads?

$$P(A) = \frac{\text{\# of outcomes in A}}{\text{\# of total possibilities}}$$

$$\text{\# of total possibilities} = 2^{100}$$

$$\text{\# of outcomes with 60 heads} = \binom{100}{60}$$

$$P(60\text{ H}) = \binom{100}{60} \Big/ 2^{100} \approx 0.011$$

Here, the symbol $\binom{n}{k}$ is defined by:

$\binom{n}{k}$ = the number of ways of choosing exactly k elements from a collection of n elements

Here, the symbol $\binom{n}{k}$ is defined by:

$\binom{n}{k}$ = the number of ways of choosing exactly k elements from a collection of n elements

How do we compute it? Choose 'y' from exactly k factors:

$$
\begin{aligned}
(x+y)^n &= (x+y)(x+y)\cdots(x+y) \\
&= x^n + nyx^{n-1} + \cdots + \binom{n}{k} y^k x^{n-k} \\
&\quad + \cdots + ny^{n-1}x + y^n
\end{aligned}
$$

Pascal's triangle

(Rows and columns starting from zero)

$\binom{n}{k}$

*k*th entry ↓

1

1 1

1 2 1

*n*th row ⟶ 1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

1 6 15 20 15 6 1

1 7 21 35 35 21 7 1

1 8 28 56 70 56 28 8 1

1 9 36 84 126 126 84 36 9 1

1 10 45 120 210 252 210 120 45 10 1

⋮

Pascal's triangle

(Rows and columns starting from zero)

kth entry

$$\binom{n}{k}$$

$$\binom{3}{1} = 3$$

1

1 1

1 2 1

nth row → 1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

1 6 15 20 15 6 1

1 7 21 35 35 21 7 1

1 8 28 56 70 56 28 8 1

1 9 36 84 126 126 84 36 9 1

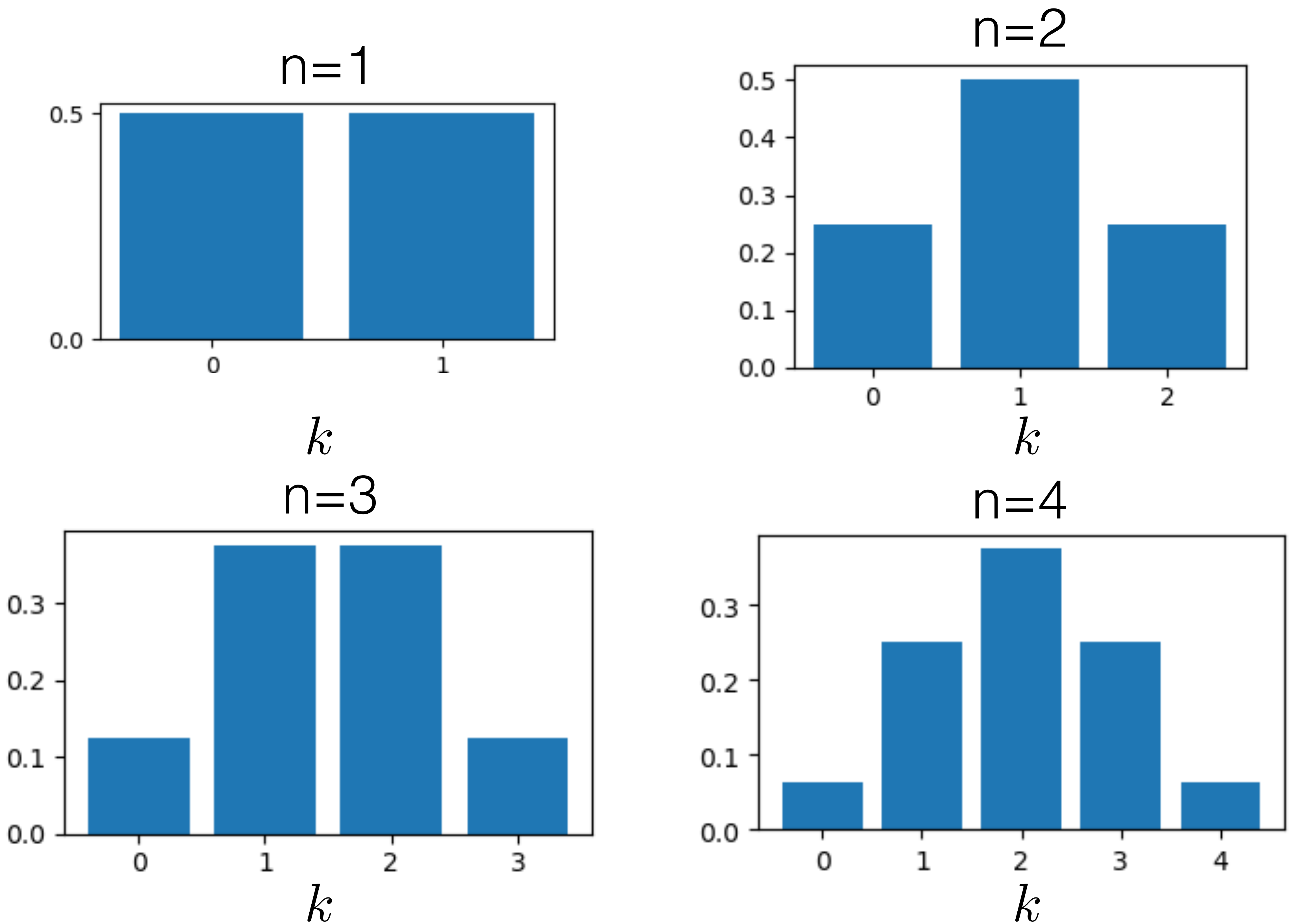
1 10 45 120 210 252 210 120 45 10 1

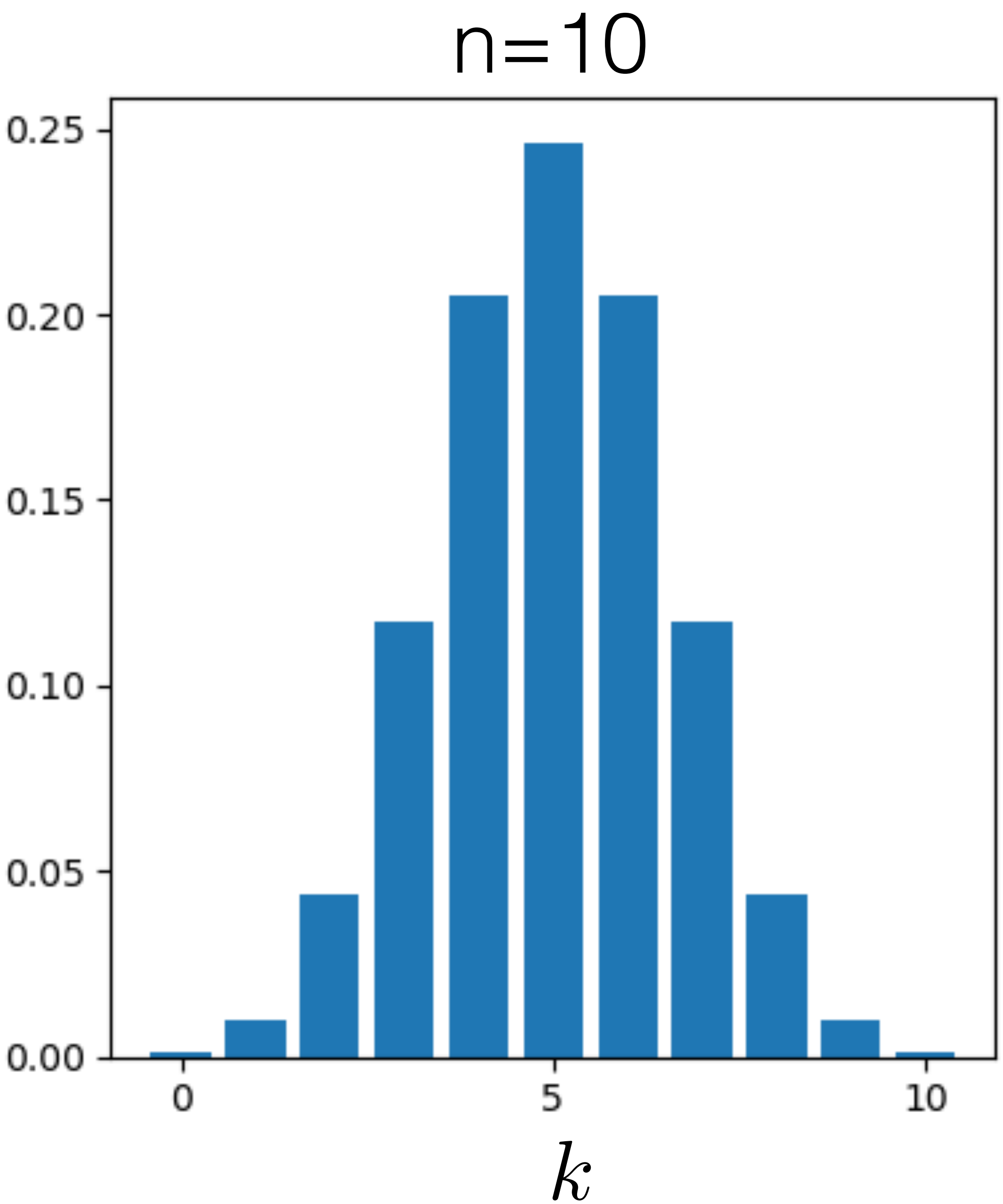
⋮

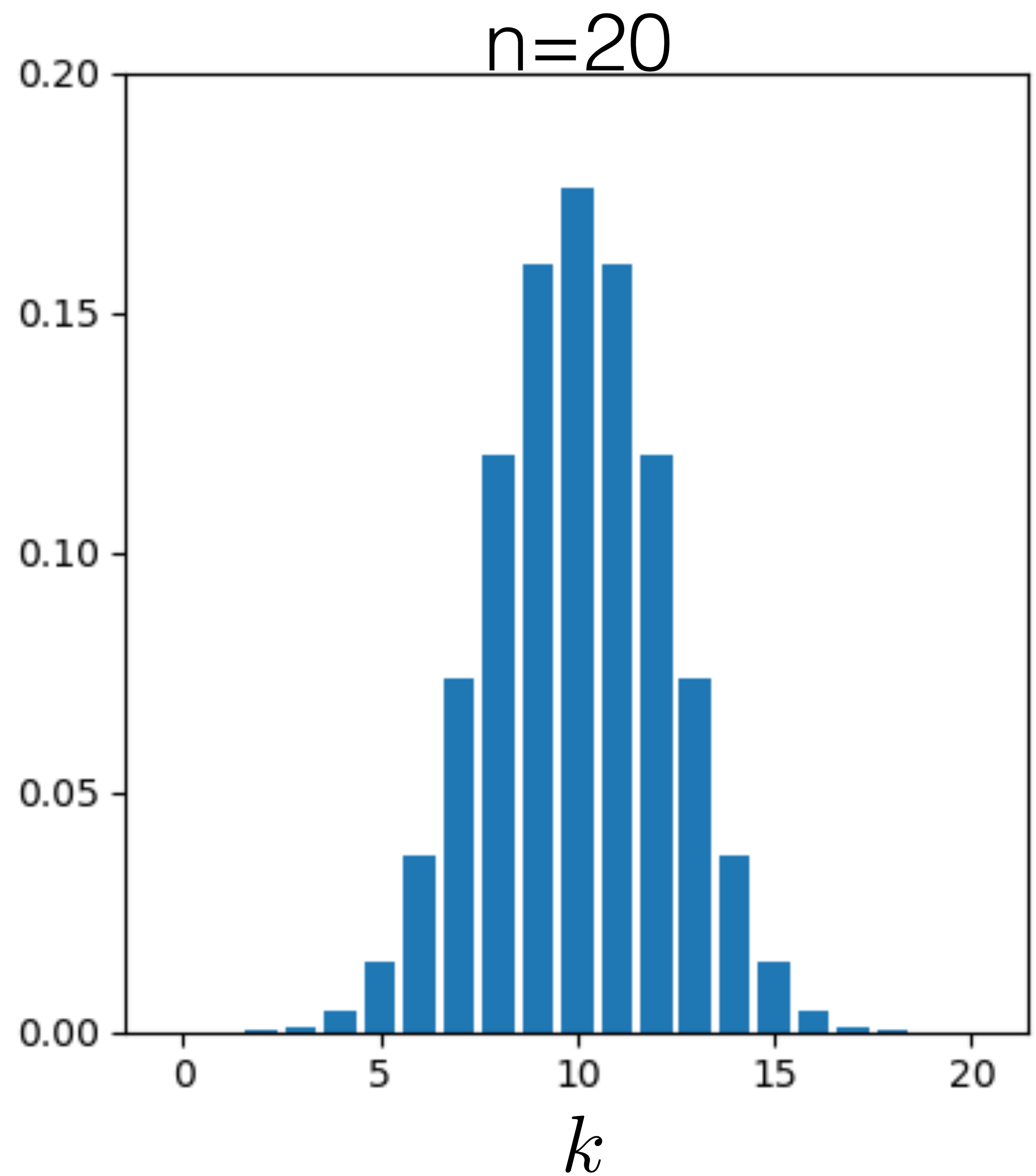
The probability of getting k heads in n flips is therefore

$$\binom{n}{k} \cdot \frac{1}{2^n}$$

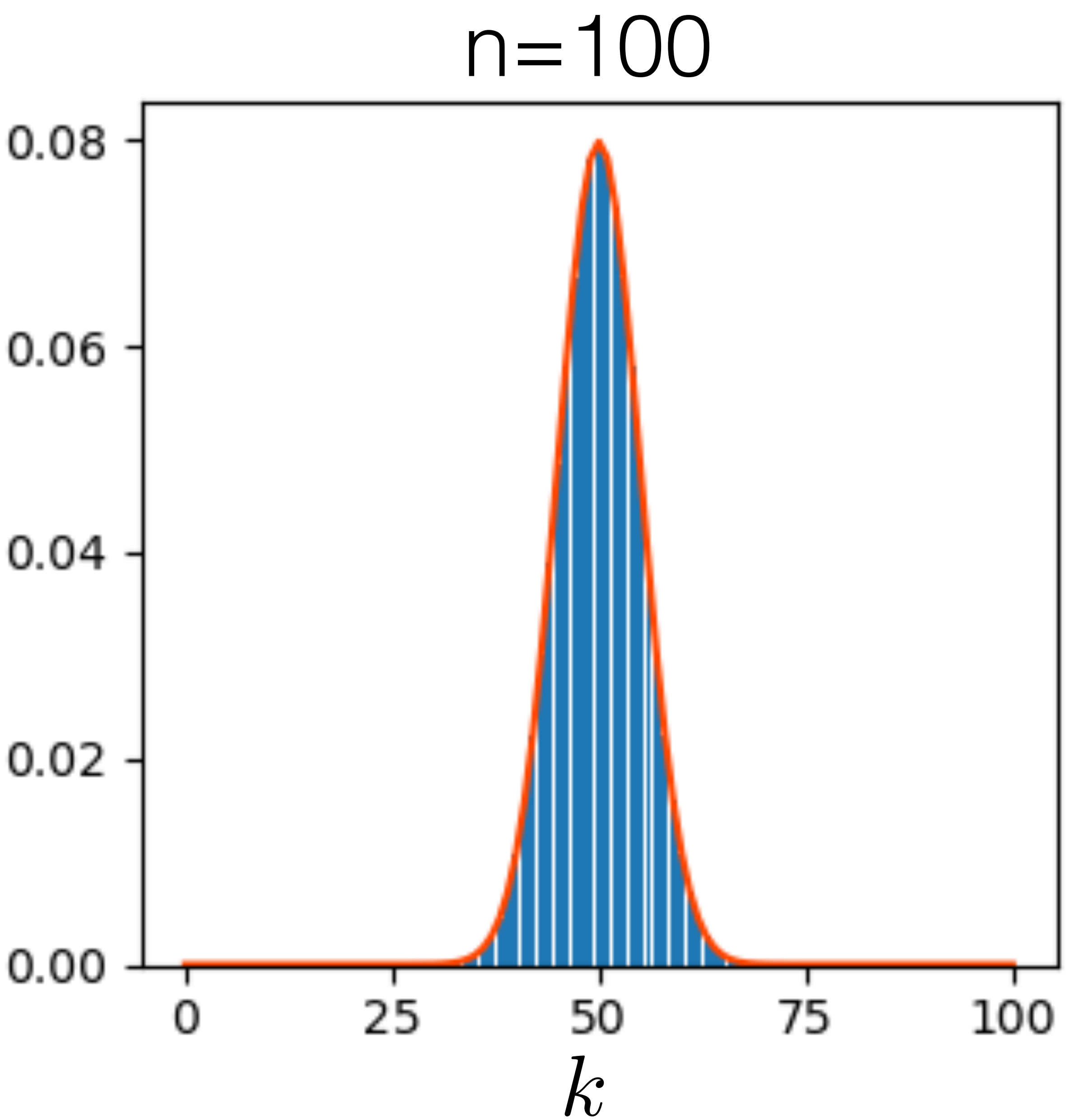
Let's plot this for different n to understand it better!


n=1
0.5
0.0
0
1
k
n=2
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
k
n=3
0.3
0.2
0.1
0.0
0
1
2
3
k
n=4
0.3
0.2
0.1
0.0
0
1
2
3
4
k


n=10
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
k


n=20
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
k

Normal distribution



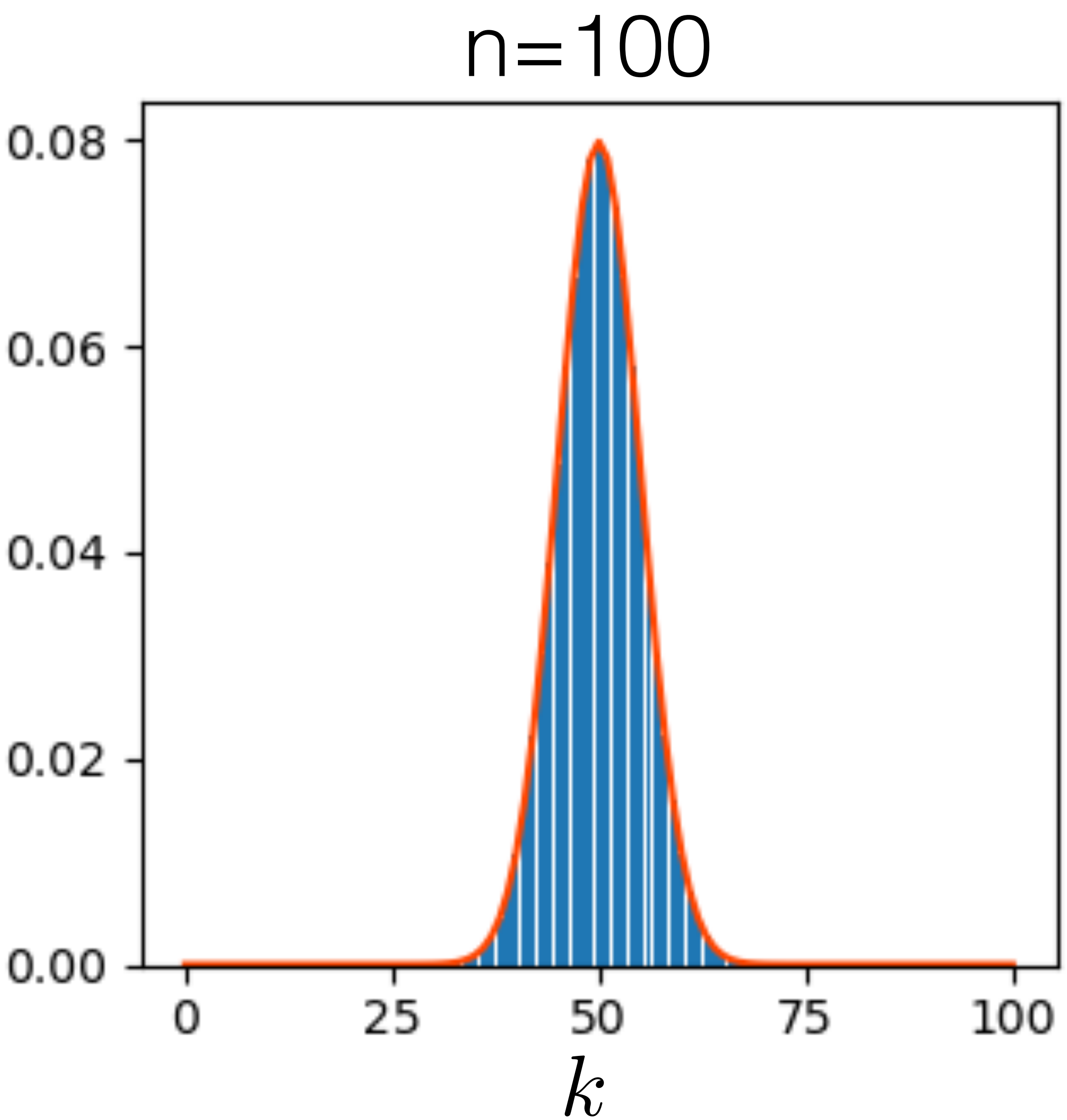
$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ : mean

σ : standard deviation

a.k.a. Gaussian distribution or Bell curve

Normal distribution



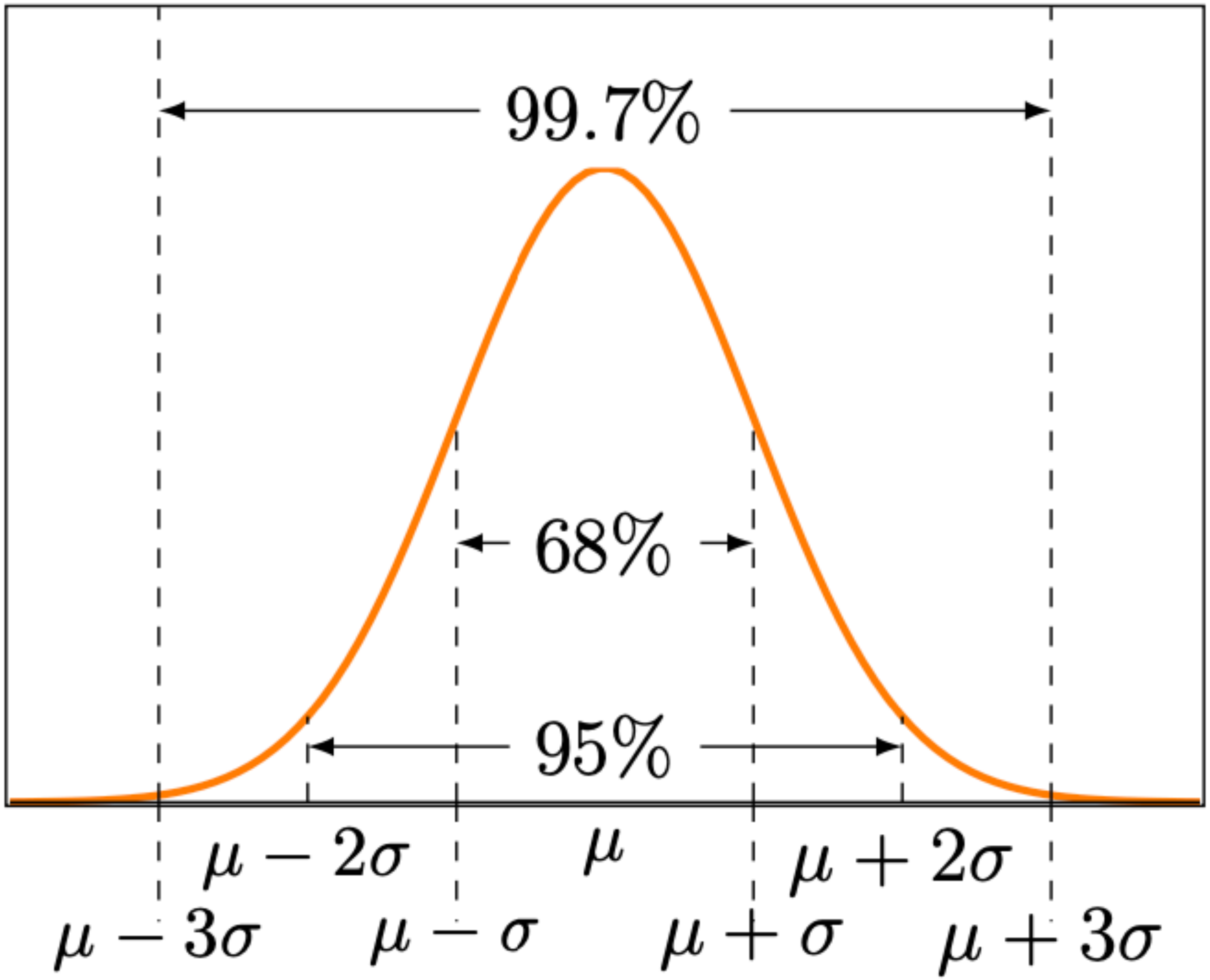
$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

For the coin flips:

$$\mu = \frac{n}{2}, \quad \sigma = \frac{\sqrt{n}}{2}$$

a.k.a. Gaussian distribution or Bell curve

3-sigma rule

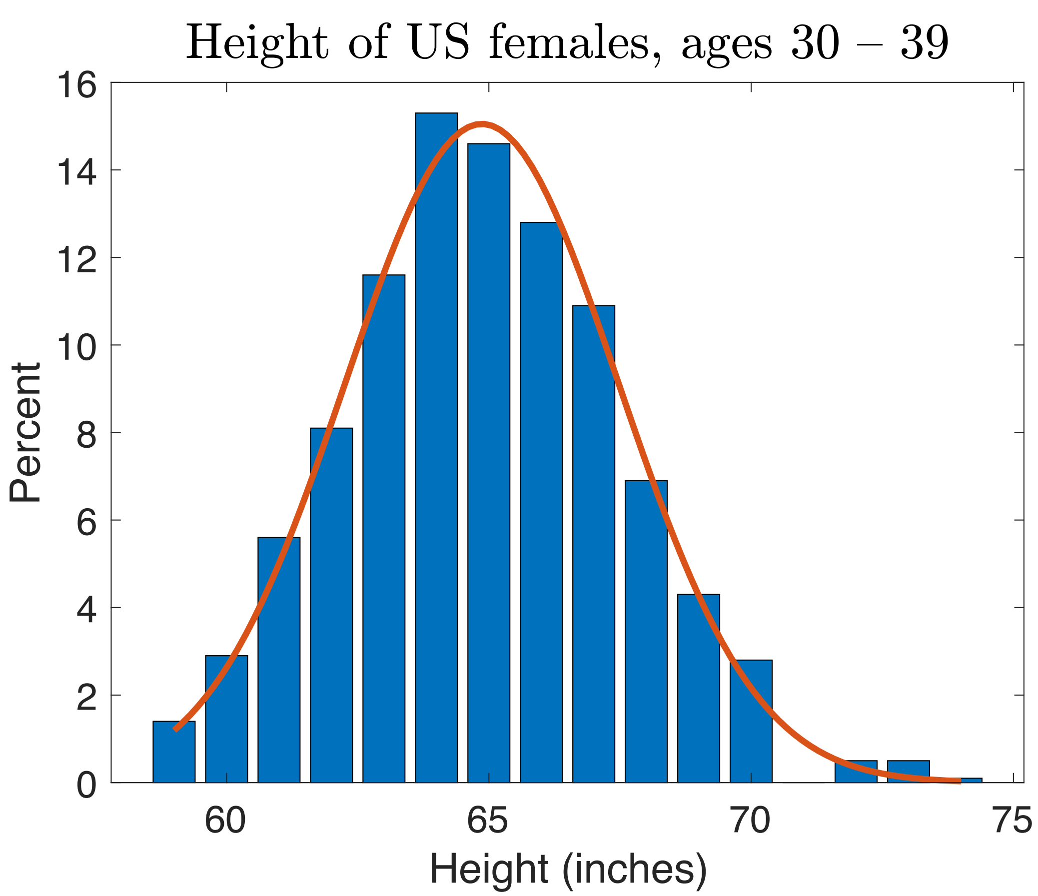


99.7%
68%
95%
$\mu - 2\sigma$
μ
$\mu + 2\sigma$
$\mu - 3\sigma$
$\mu - \sigma$
$\mu + \sigma$
$\mu + 3\sigma$

If the distribution of the data points $\{x_1, x_2, \ldots, x_N\}$ is close to normal,

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots (x_N - \mu)^2}{N}$$



$\mu \approx 64.9$

$\sigma \approx 2.6$

Central limit theorem

"Sums (or means) are normally distributed"

When independent random variables are added, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed.

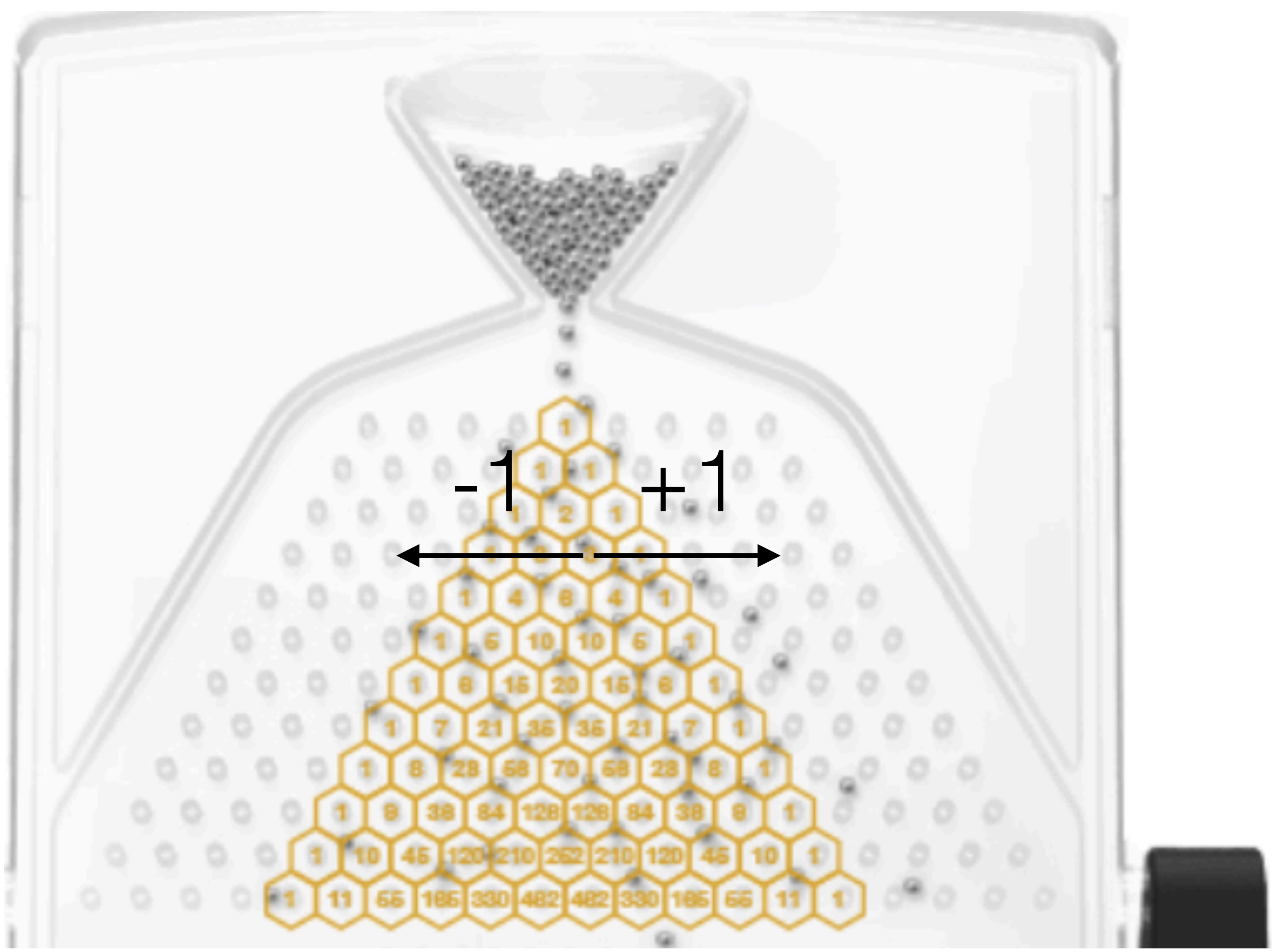
Central limit theorem

"Sums (or means) are normally distributed"

… but how are sums related to weights of the penguins and the SAT scores, heights and the Galton board?

… but how are sums related to weights of the penguins and the SAT scores, heights and the Galton board?

The position of a ball in the Galton board is a sum (to keep track of the final position, add 1 for every peg it bounces to the right, subtract 1 if it bounces to the left)



… but how are sums related to weights of the penguins and the SAT scores, heights and the Galton board?

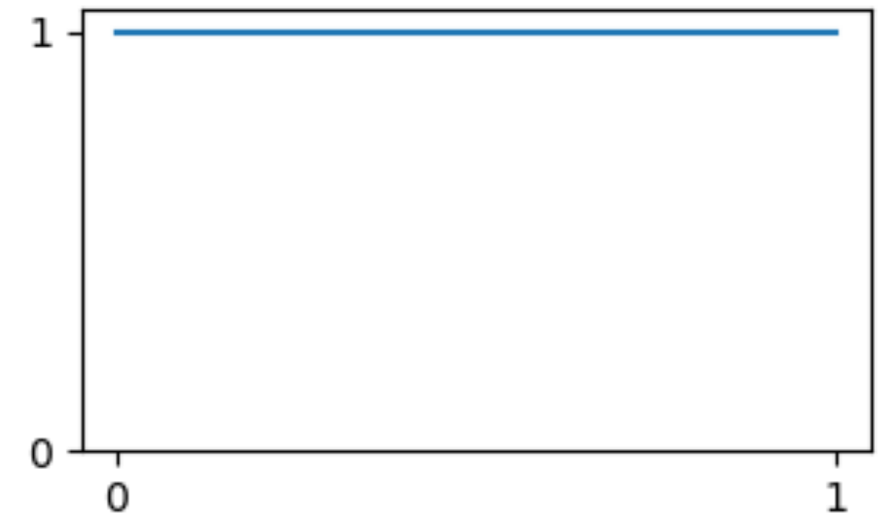
The position of the marble in the Galton board is a sum (to keep track of the final position, add 1 for every peg it bounces to the right, subtract 1 if it bounces to the left)

The average SAT score of a high school is the mean of the scores of the students in the school

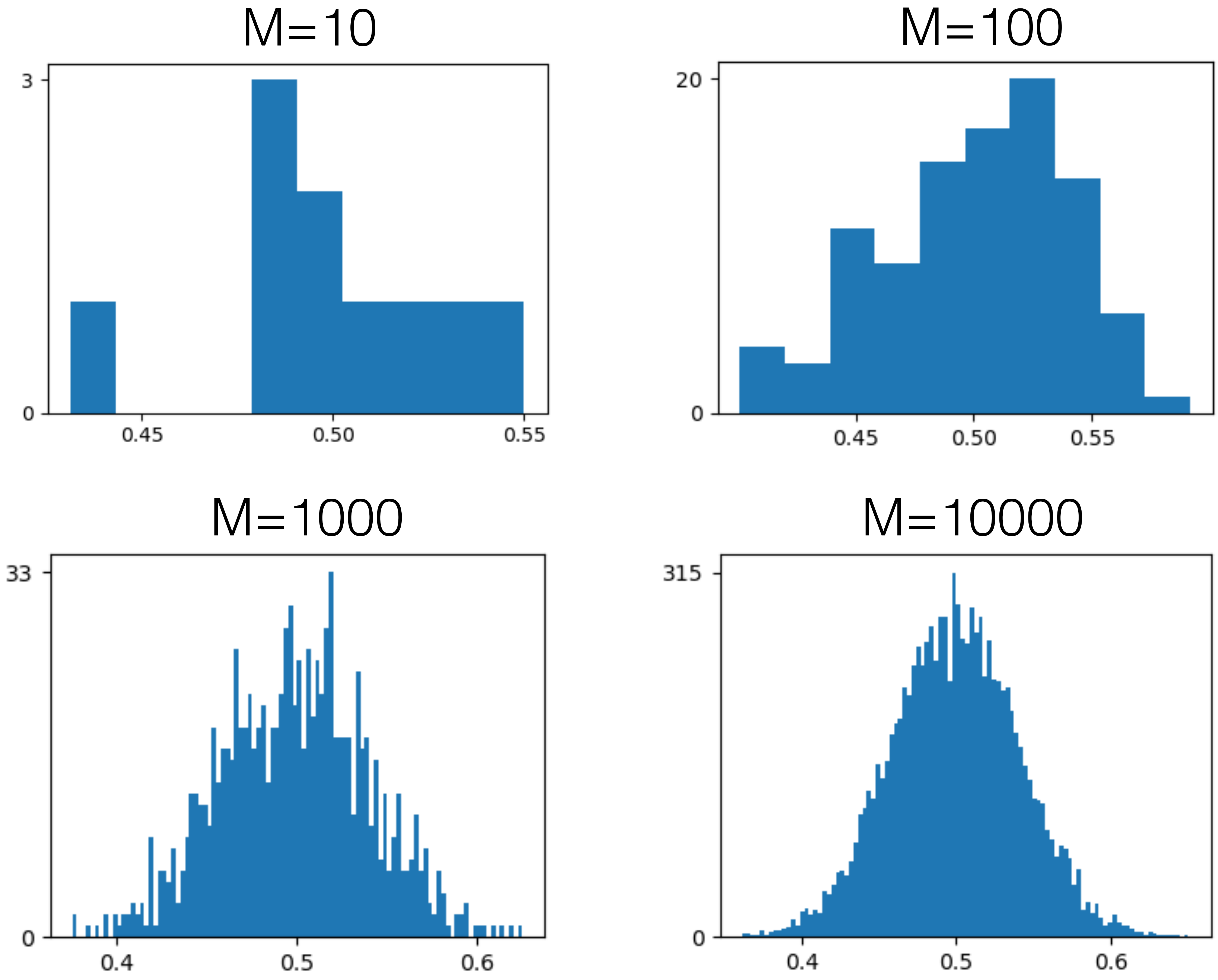
The weight of the penguins is a sum (add say 100 calories when catching a fish, 0 when failing to catch)

Consider the uniform distribution

$$P(x) = 1, \quad 0 \leq x \leq 1$$

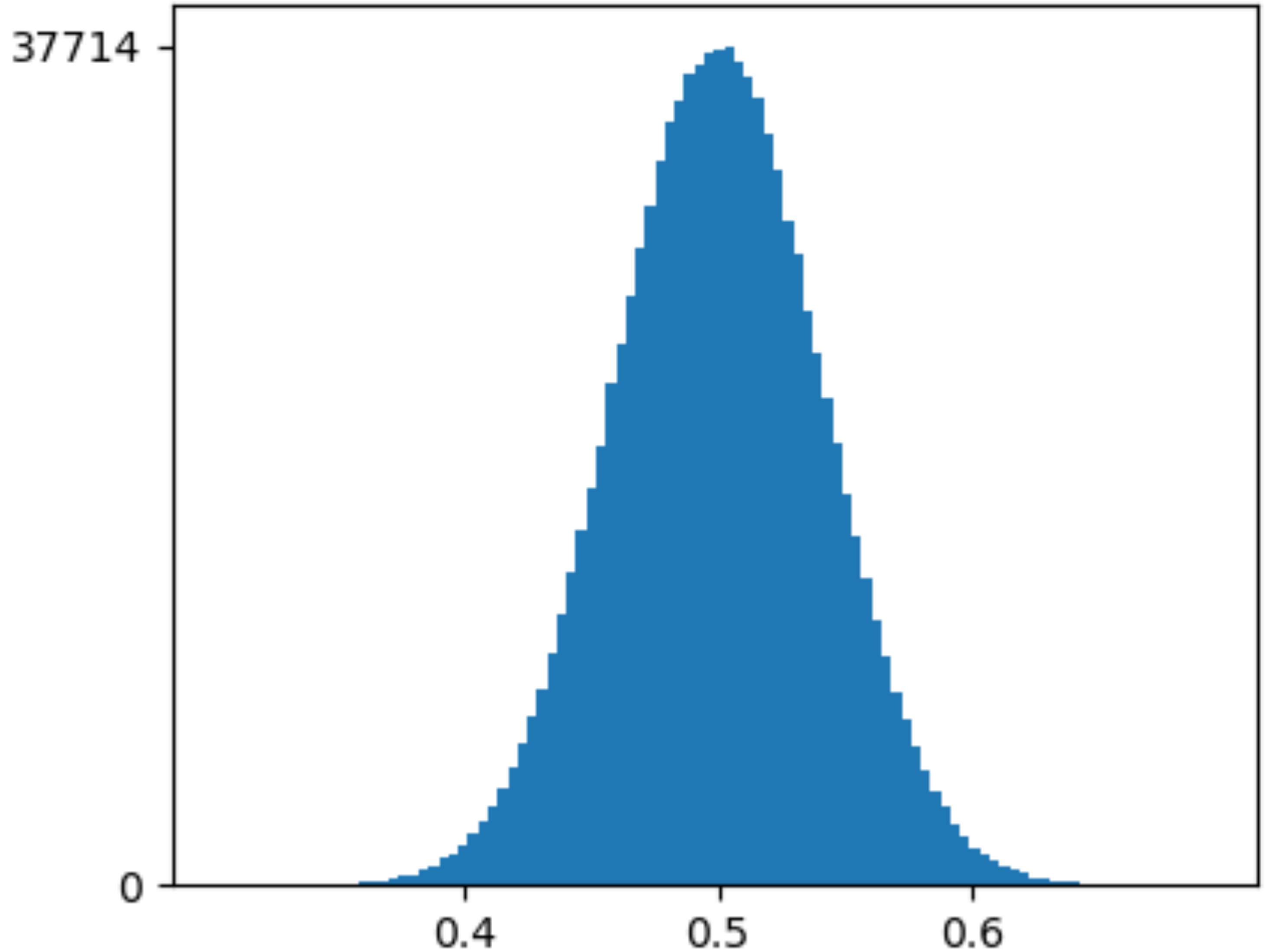


In each trial of an experiment, we randomly draw 50 samples from the distribution and calculate their mean. We then plot the histogram of the means obtained from M trials.



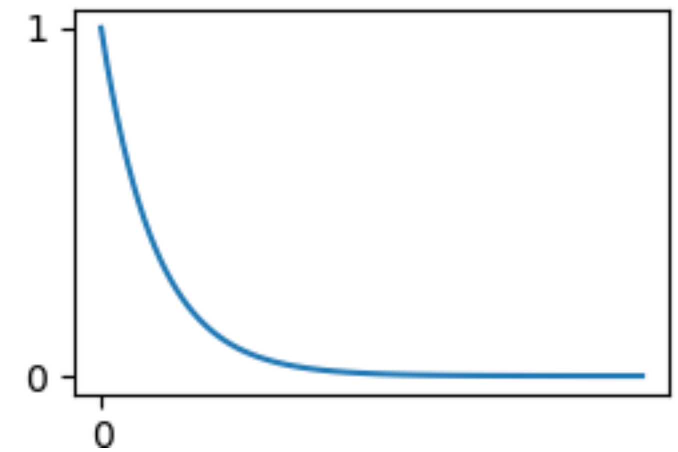
M=10
3
0
0.45
0.50
0.55
M=100
20
0
0.45
0.50
0.55
M=1000
33
0
0.4
0.5
0.6
M=10000
315
0
0.4
0.5
0.6

$M = 10^6$

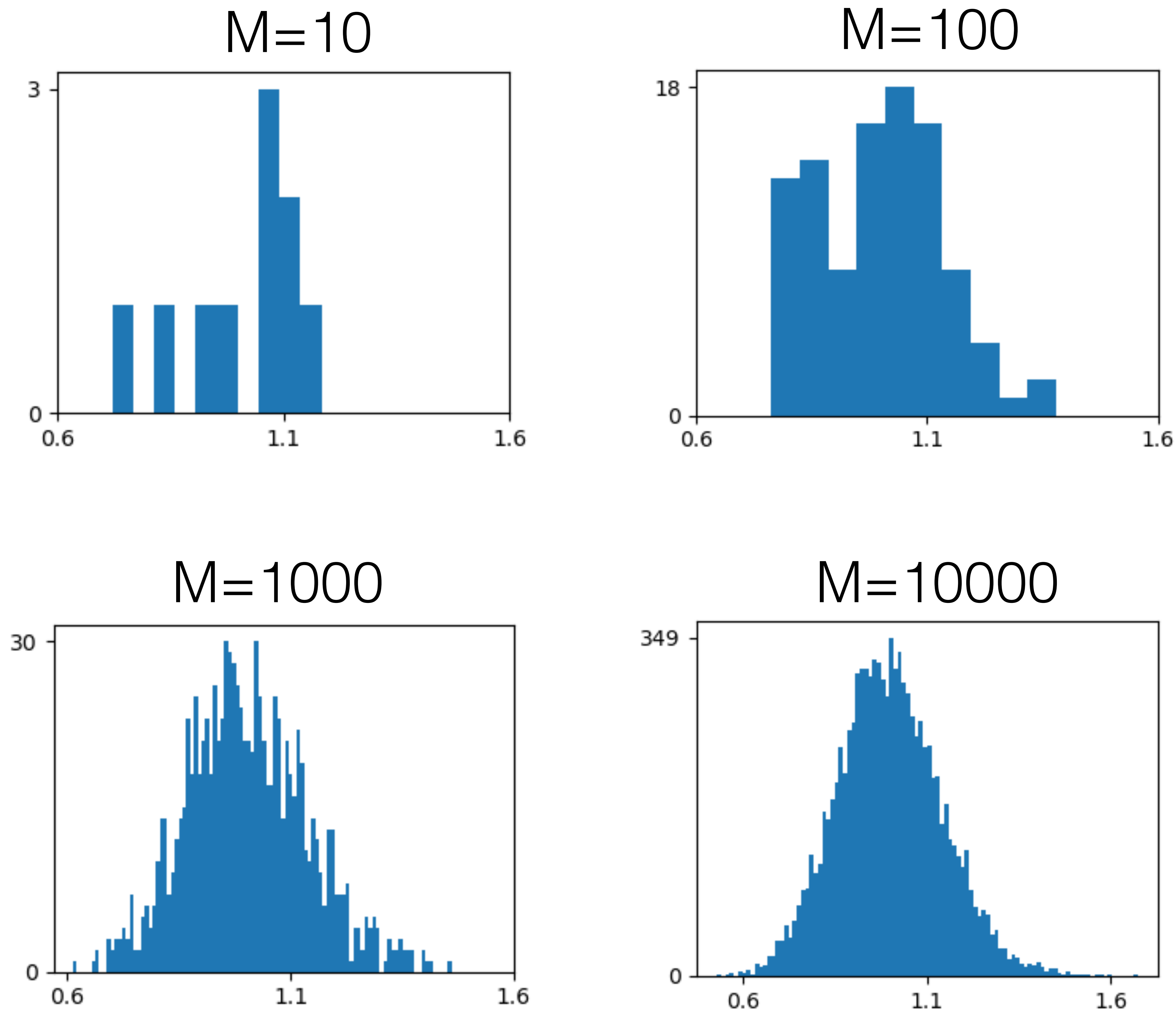


Consider another distribution

$$P(x) = e^{-x}, \quad x \geq 0$$



We repeat the same experiment: In each trial, we randomly draw 50 samples from the distribution and calculate their mean. We then plot the histogram of the means obtained from M trials.


M=10
3
0
0.6
1.1
1.6
M=100
18
0
0.6
1.1
1.6
M=1000
30
0
0.6
1.1
1.6
M=10000
349
0
0.6
1.1
1.6

$$M = 10^6$$