Introduction to Data Analysis in R (Day 1)

Dr. Behnam Yousefi
Institute of medical systems biology, UKE
March 2024

24% blue

20% orange

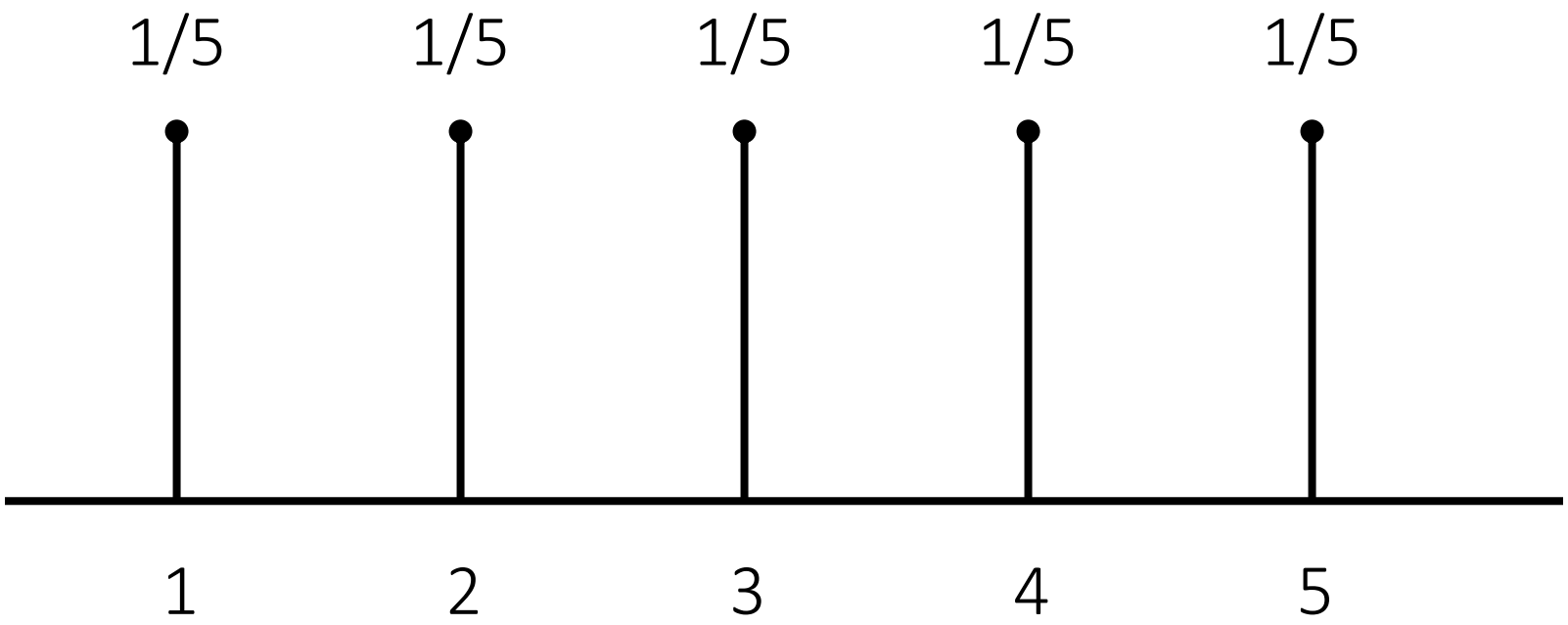
16% green

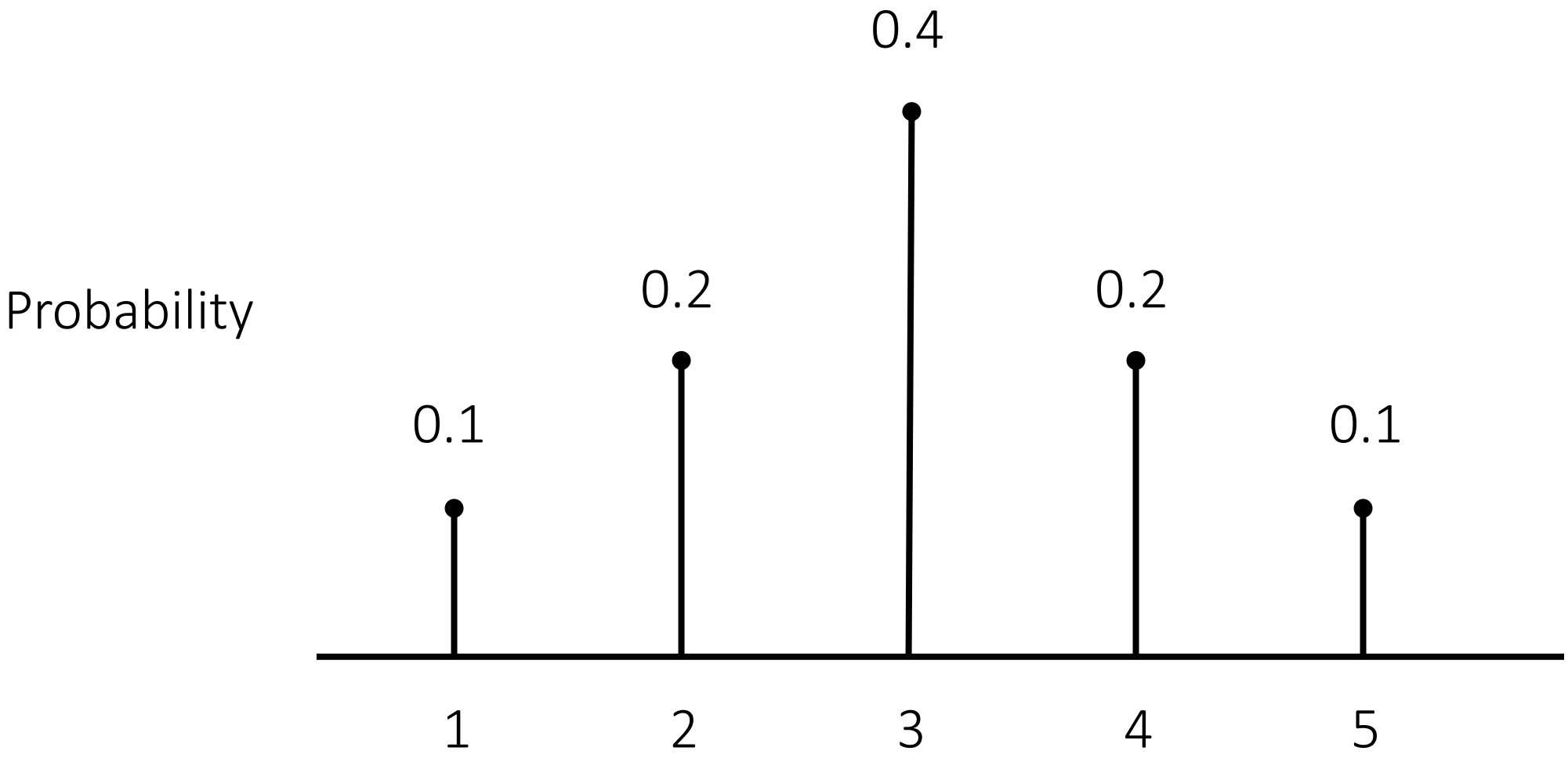
14% yellow

13% red

13% brown

Probability

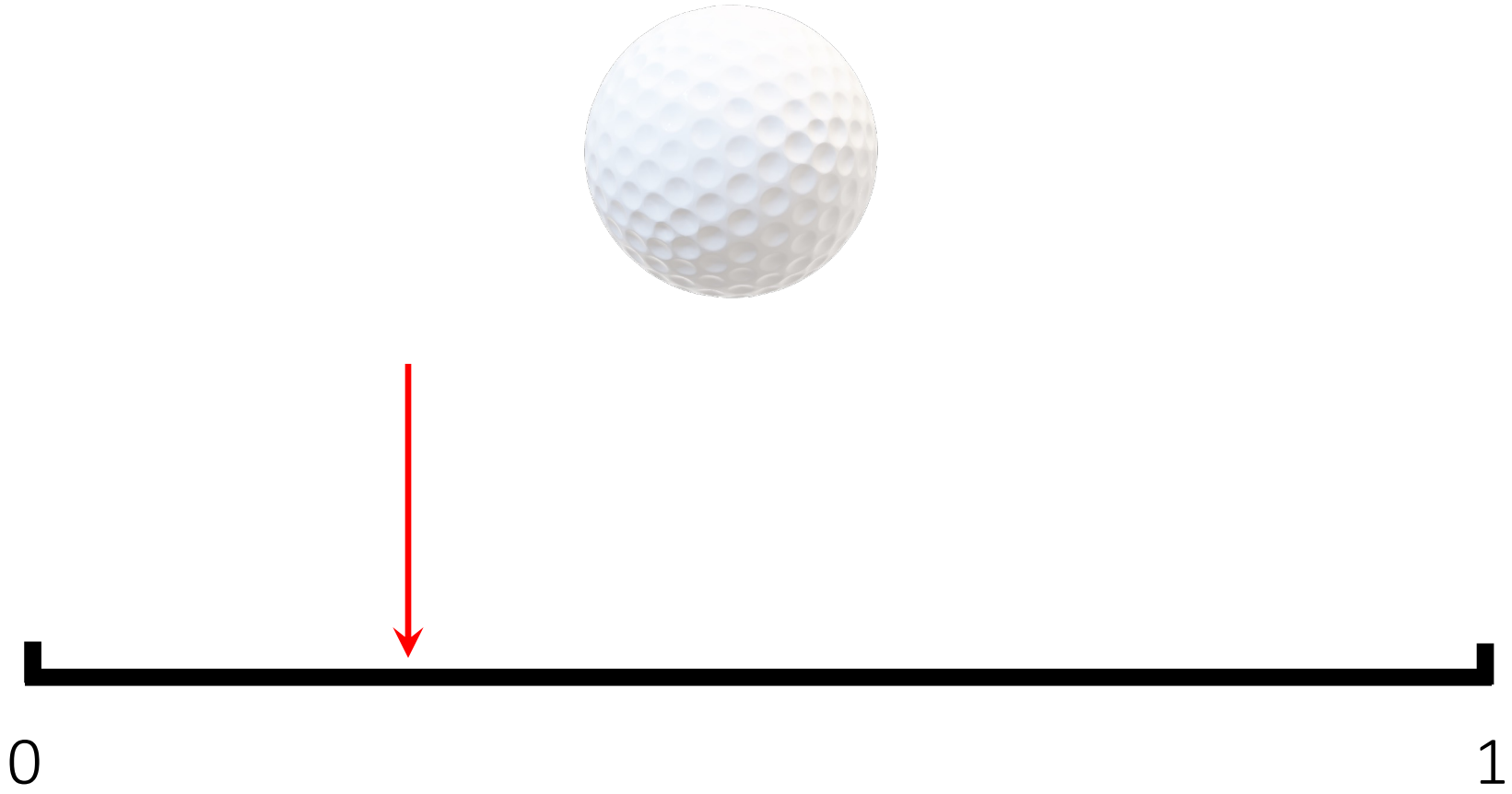


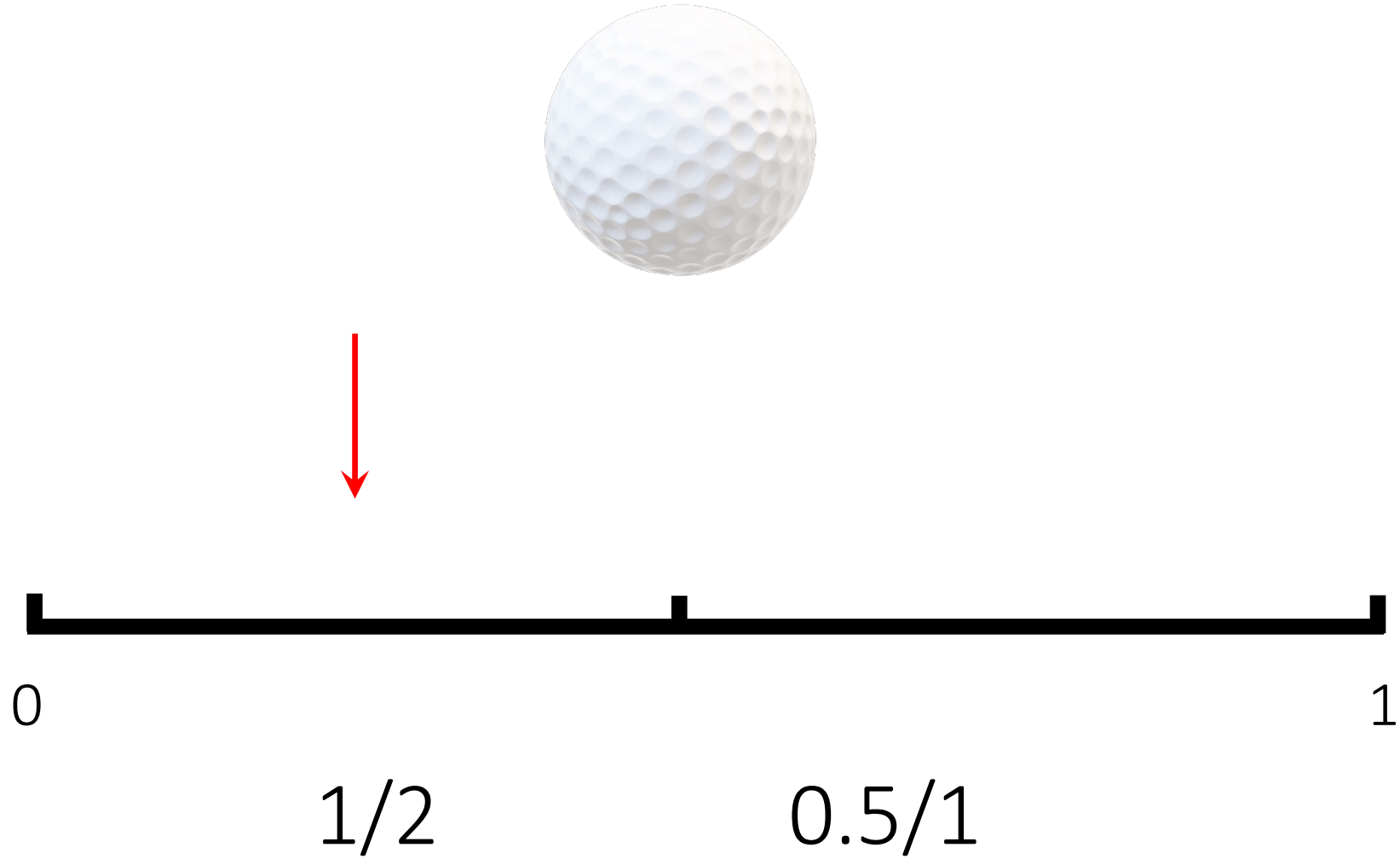

Probability
0.4
0.2
0.2
0.1
0.1
1
2
3
4
5

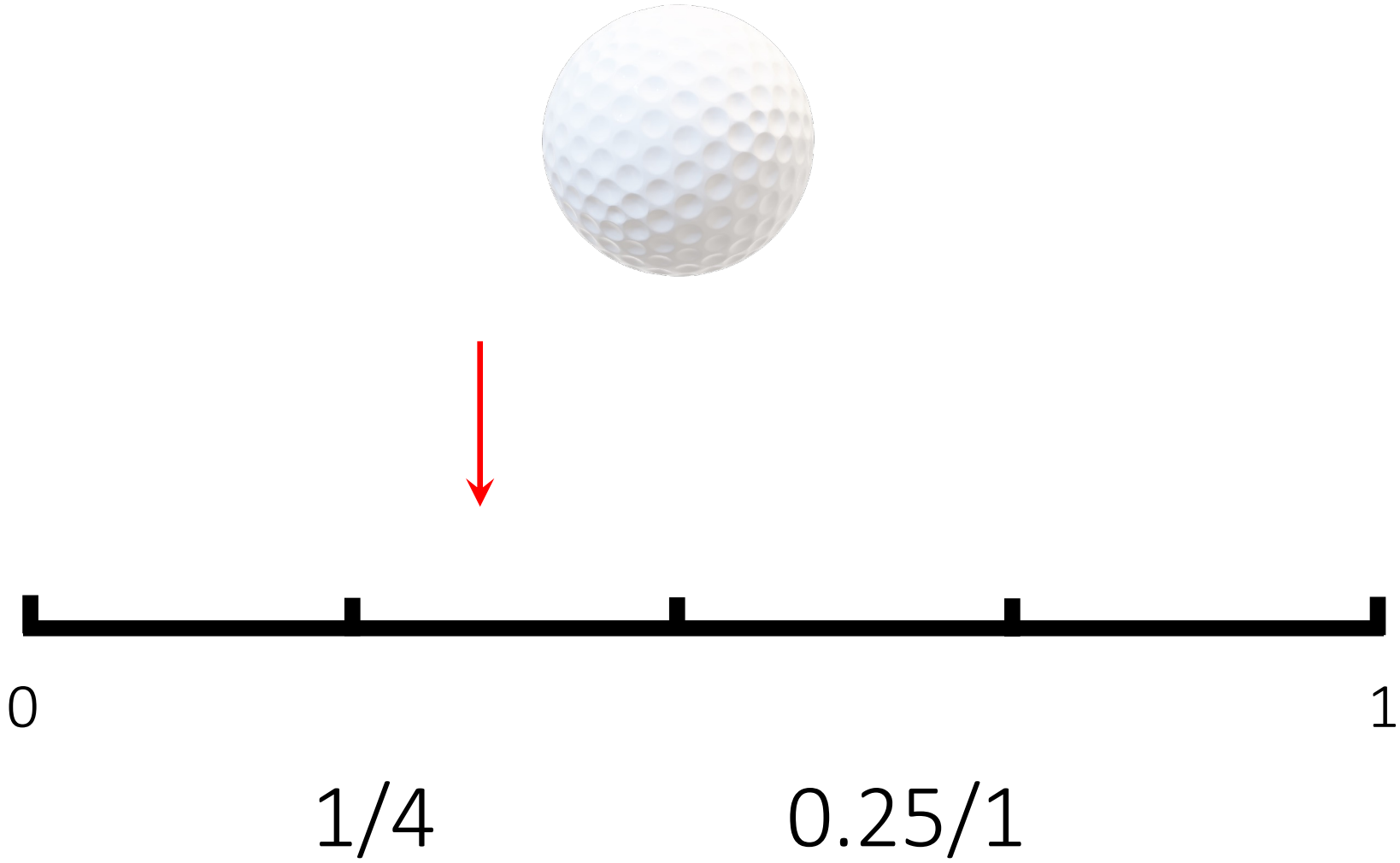
p_i : the probably of the event *i*

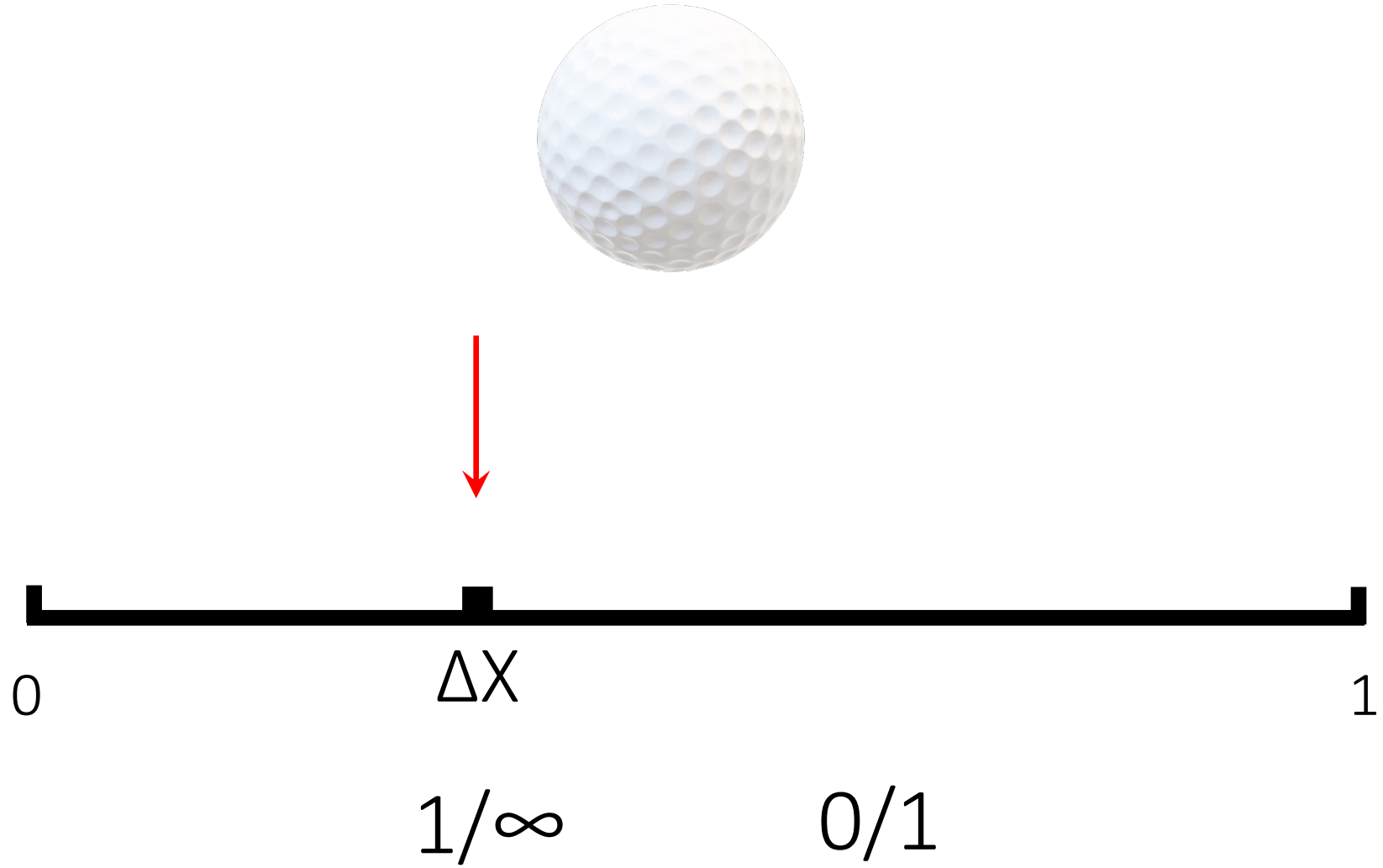
$$0 \leq p_i \leq 1$$

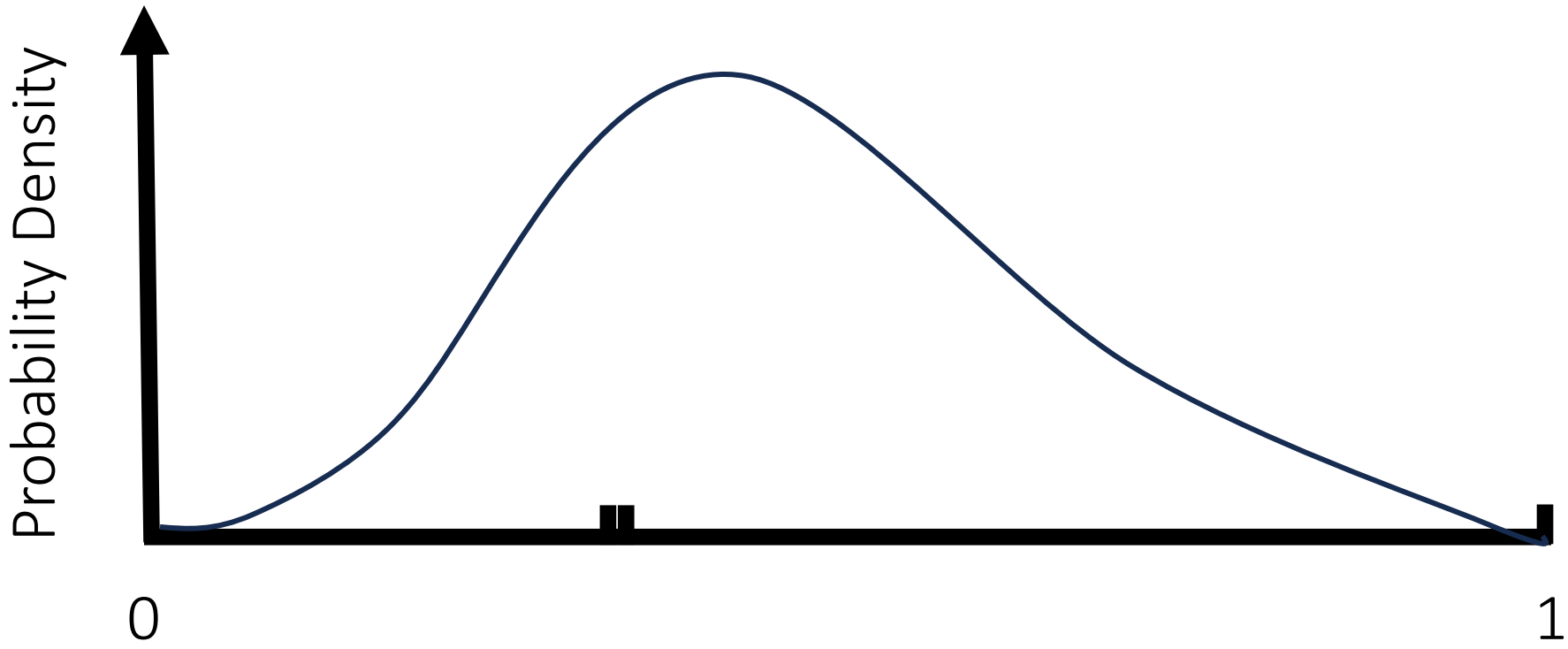
$$\sum_{all\ i} p_i = 1$$


0
1

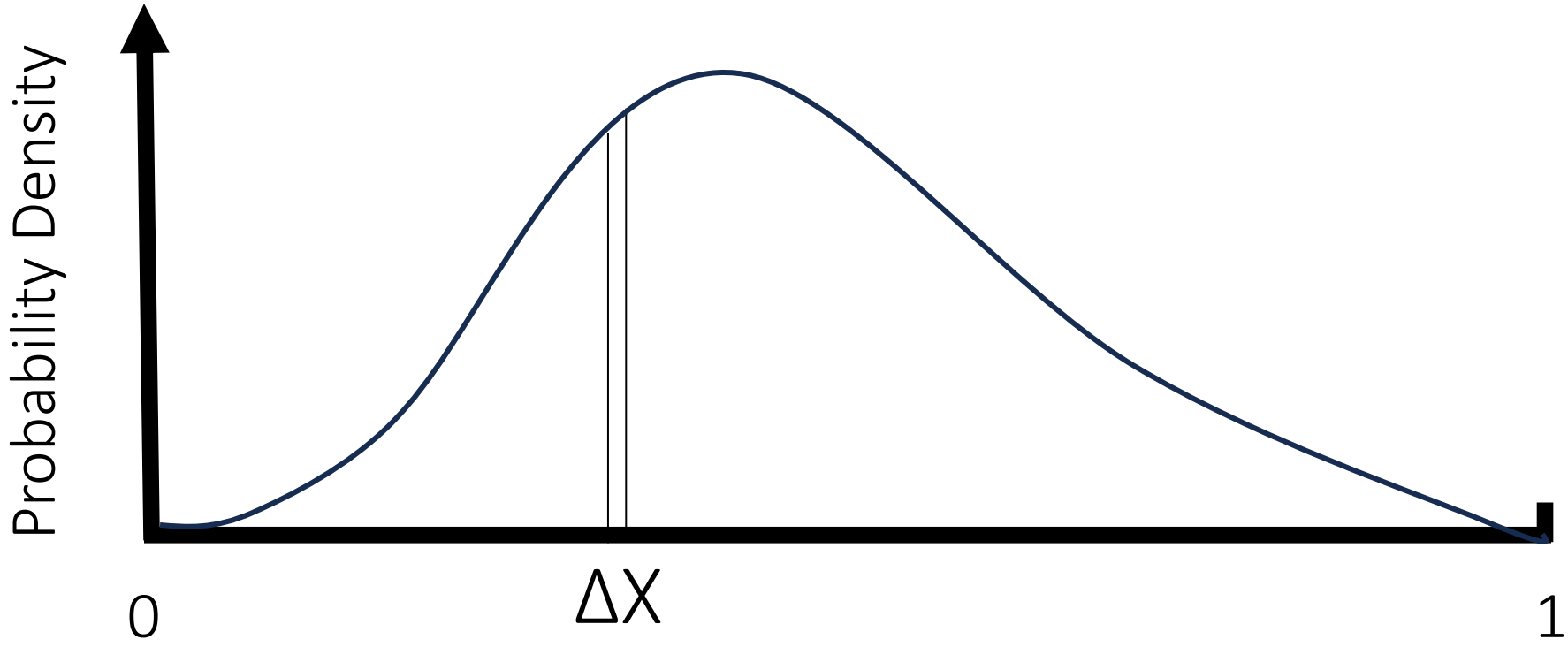

0
1
1/2
0.5/1


0
1
1/4
0.25/1

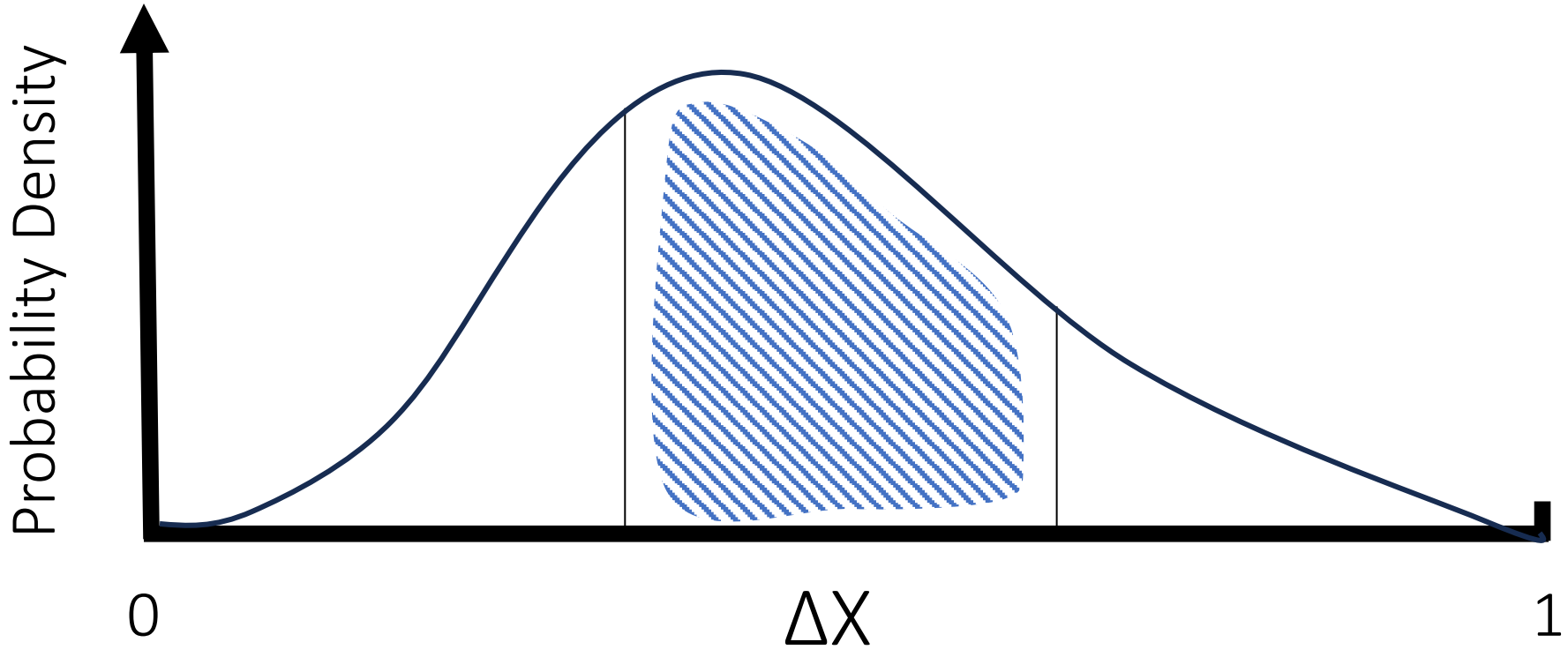

0
ΔX
1
1/∞
0/1


Probability Density
0
1

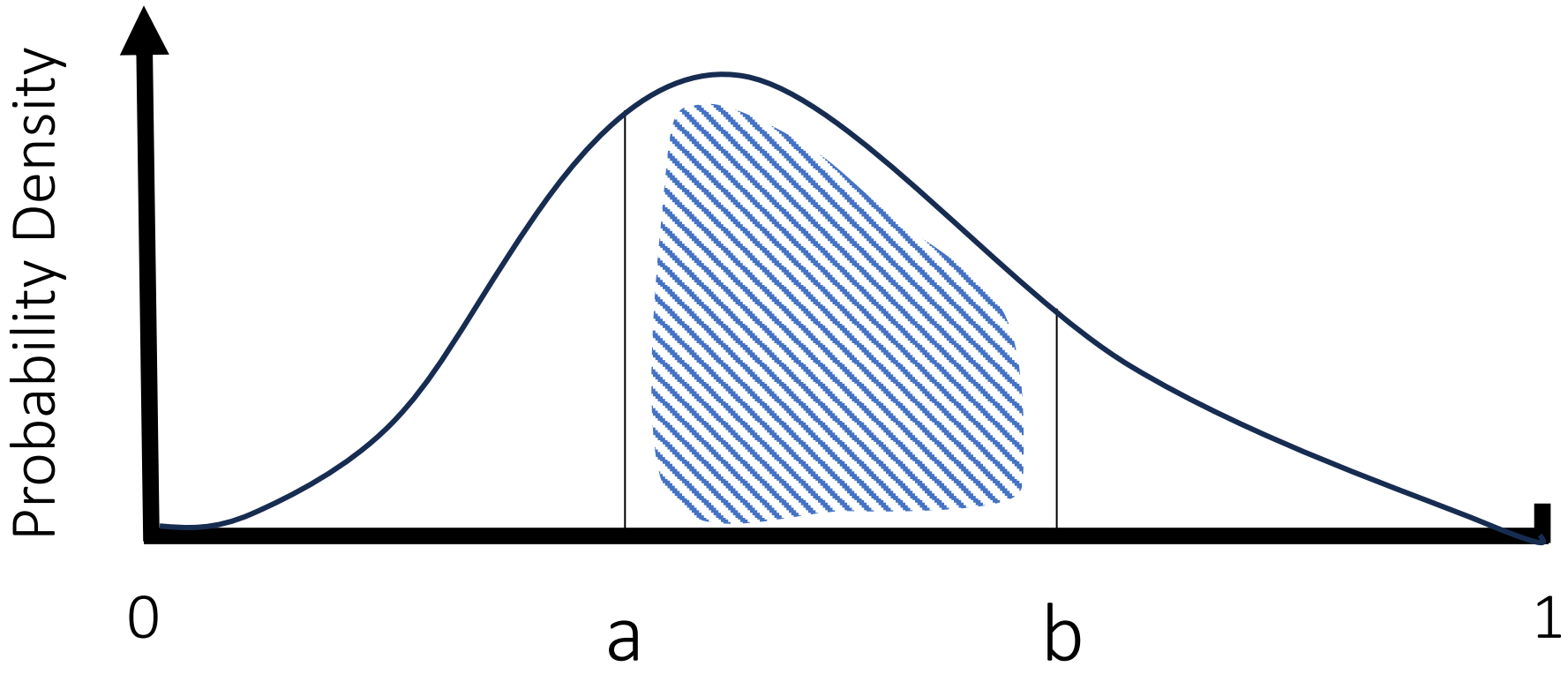
Probability = Density × ΔX



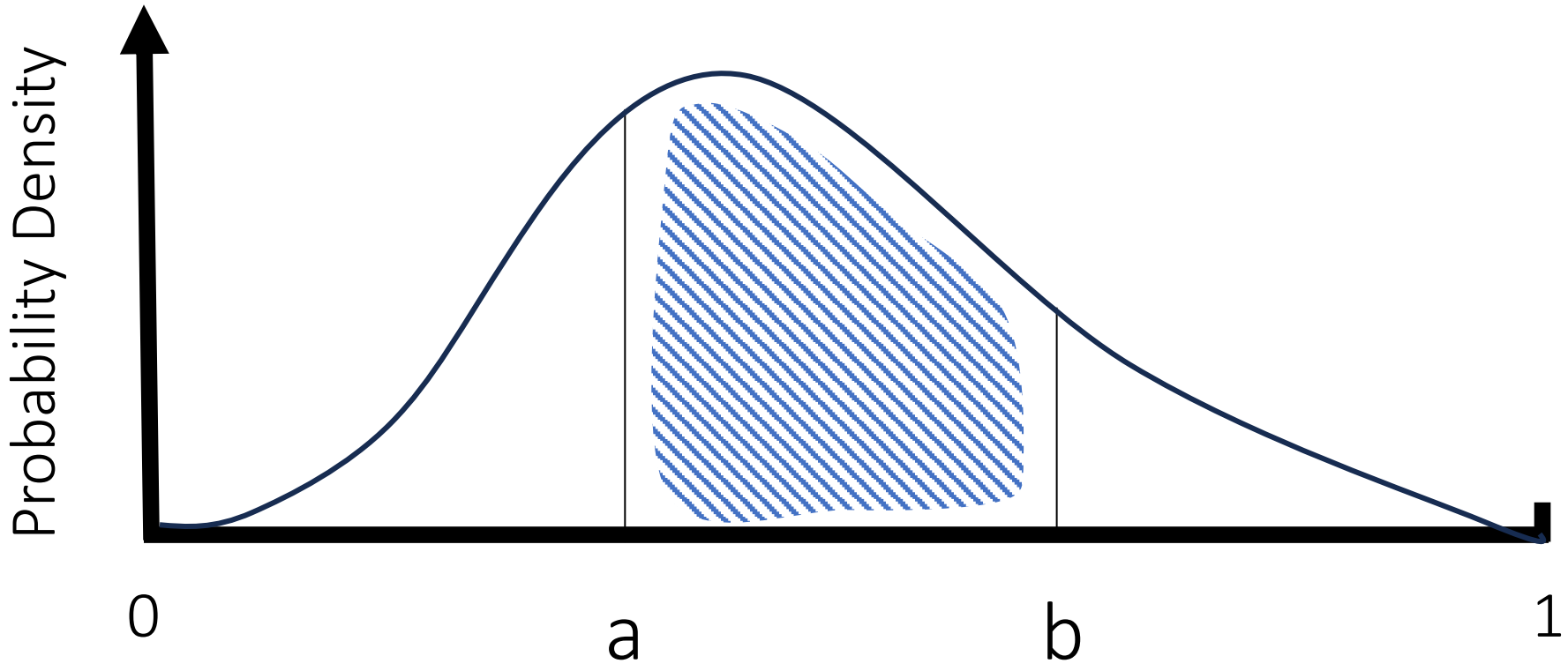
Probability = area under the density curve over ΔX



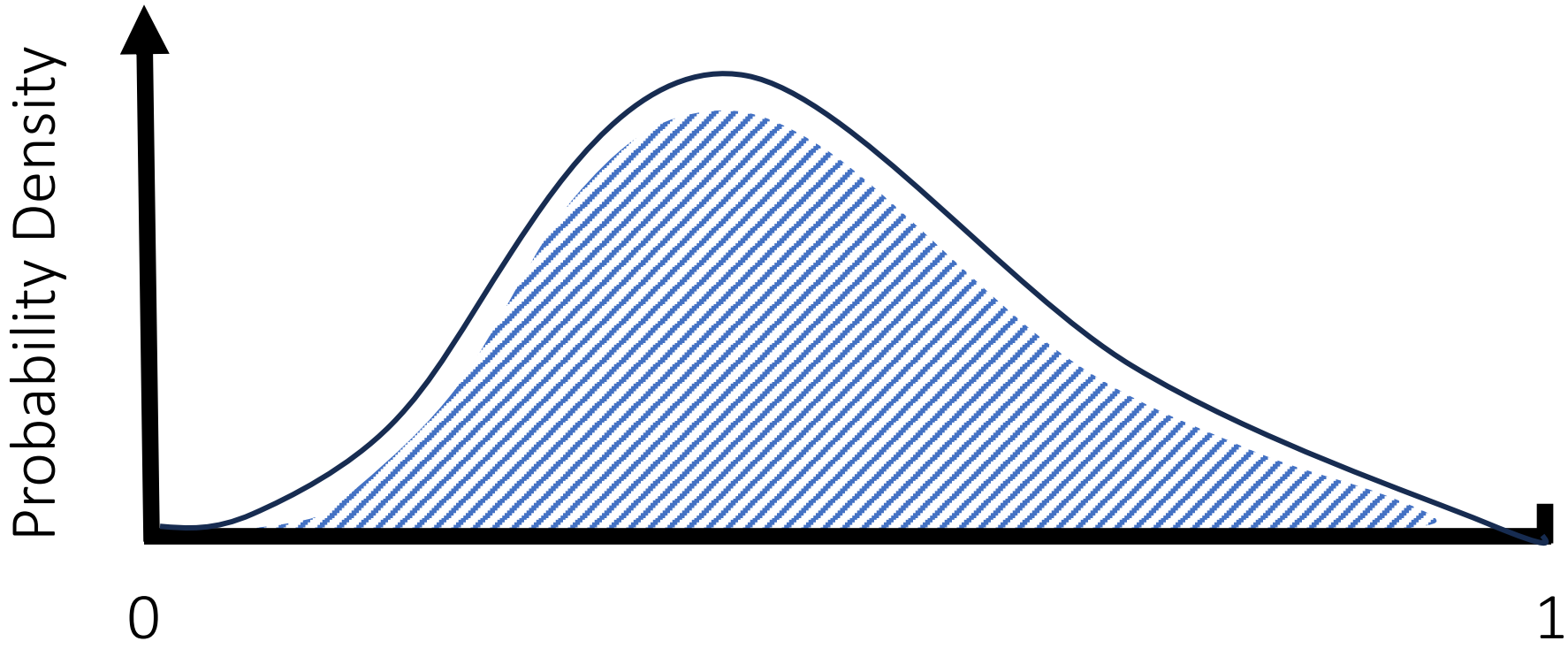
$$P(a < x < b)$$



$$P(a < x < b) = \int_a^b p(x)dx$$



$$P(-\infty < x < +\infty) = \int_{-\infty}^{+\infty} p(x)dx$$




Congratulations!

RStudio

Programming Languages

R	Python
Pros: Free and open source Broad community Designed for data analysis Easier for non-programmers	**Pros:** Free and open source Broad community Broader application
Cons: Not suitable for advanced deep learning	**Cons:** Steeper learning curve for non-specialists

Basic math

Add	+
Subtract	–
Multiply	*
Divide	/
remainder	%%
Power	^

Variables

```
> a = 1
```

```
> a <- 1
```

```
> a = a + b
```

```
> print(a)
```

```
> a = “Hello”
```

```
> a = “TRUE”
```

```
> a = TRUE
```

```
> a = NA
```

Logical operations

Is equal?	==
Is not equal?	!=
Is greater?	>
Is less?	<
Is greater than or equal?	>=
Is less than or equal?	<=

and	&
or	|
Not	!

Conditions

```
> height = 170        # m
> weight = 64         # kg
> bmi = ...
```

Conditions

```
> height = 170        # m
> weight = 64         # kg
> bmi = weight / (height^2)

> if (condition){
>         …
> }
```

Variables: vectors (array)

```
> var1 = c(1,2,3,4)
> var1 = 1:4
> var2 = 5:10

> var1[1]
> var1[2] = 8
```

Variables: matrices (2D array)

```
> mat1 = cbind(var1, var2)
```

```
> mat2 = rbind(var1, var2)
```

Data frame

```
> df <- data.frame(
  patient_id = 101:105,
  age = c(30, 41, 23, 53, 60),
  bmi = c(23, 26, 18, 28, 28),
  gender = c("male", "female", "male", "female", "male"),
  smoker = c("yes", "no", "no", "yes", "no")
)
```

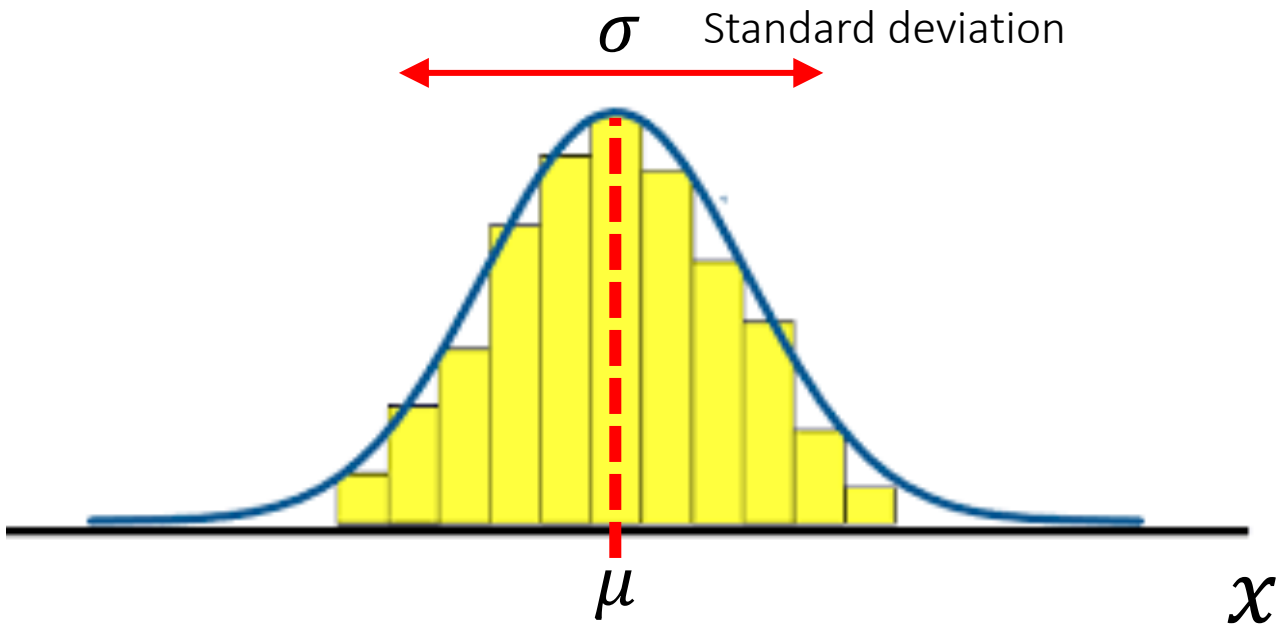
Read data frame

```
> data = read.table("obesity_data.csv", sep = ",", header = TRUE)
```

```
> data = read.csv("obesity_data.csv")
```

Data representation

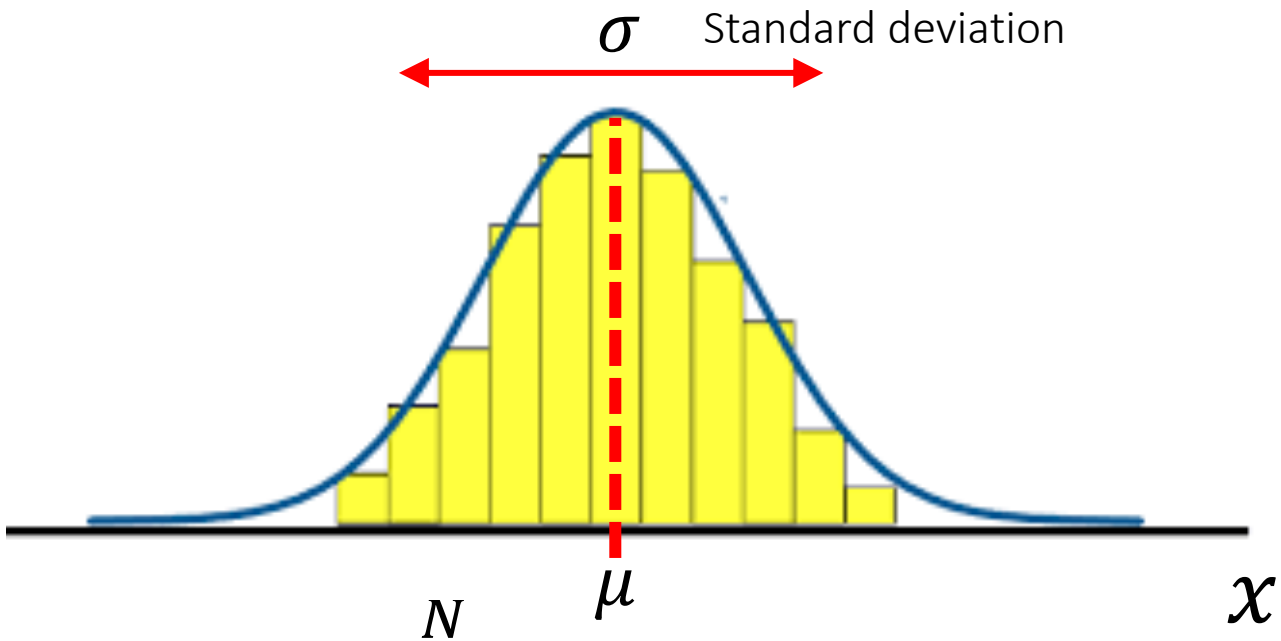
Normal Distribution



$$x \sim \mathcal{N}(\mu, \sigma)$$

$$p(x) = \frac{1}{2\pi\sqrt{\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

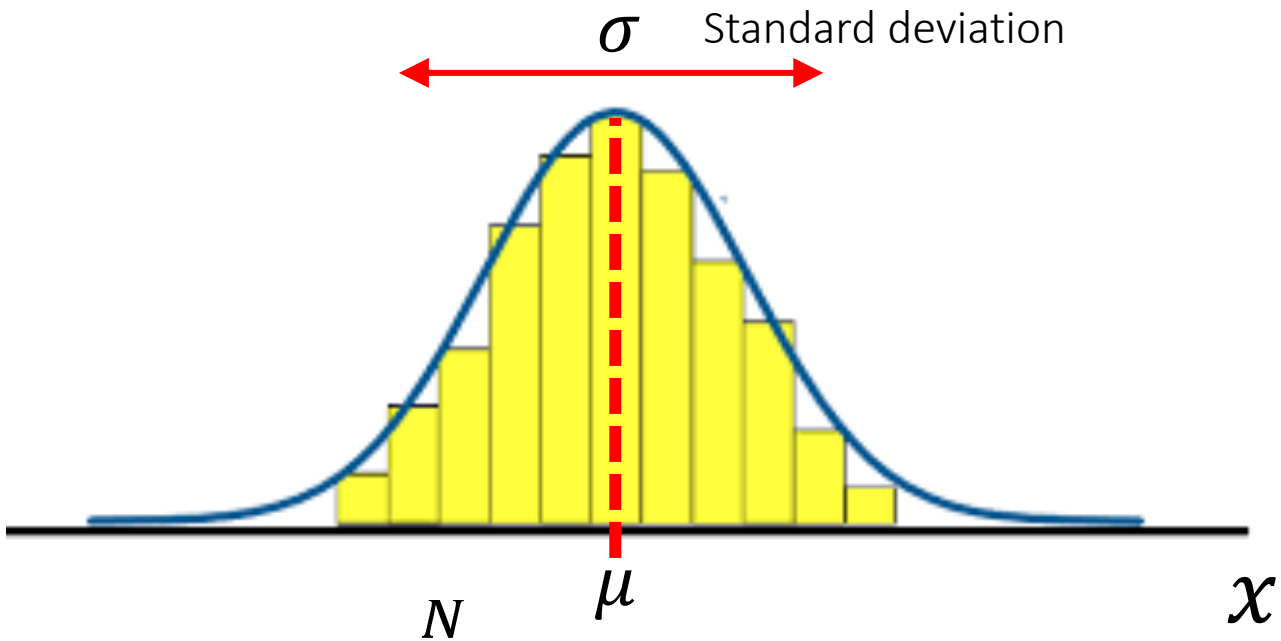
Normal Distribution



$$\mu = \mathcal{E}\{x\} \quad \approx \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2 = \mathcal{E}\{(x-\mu)^2\} \quad \approx s^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2$$

Normal Distribution



$$\mu = \mathcal{E}\{x\} \quad \approx \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2 = \mathcal{E}\{(x-\mu)^2\} \quad \approx s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Normal Distribution

$$\mu \approx \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Mean

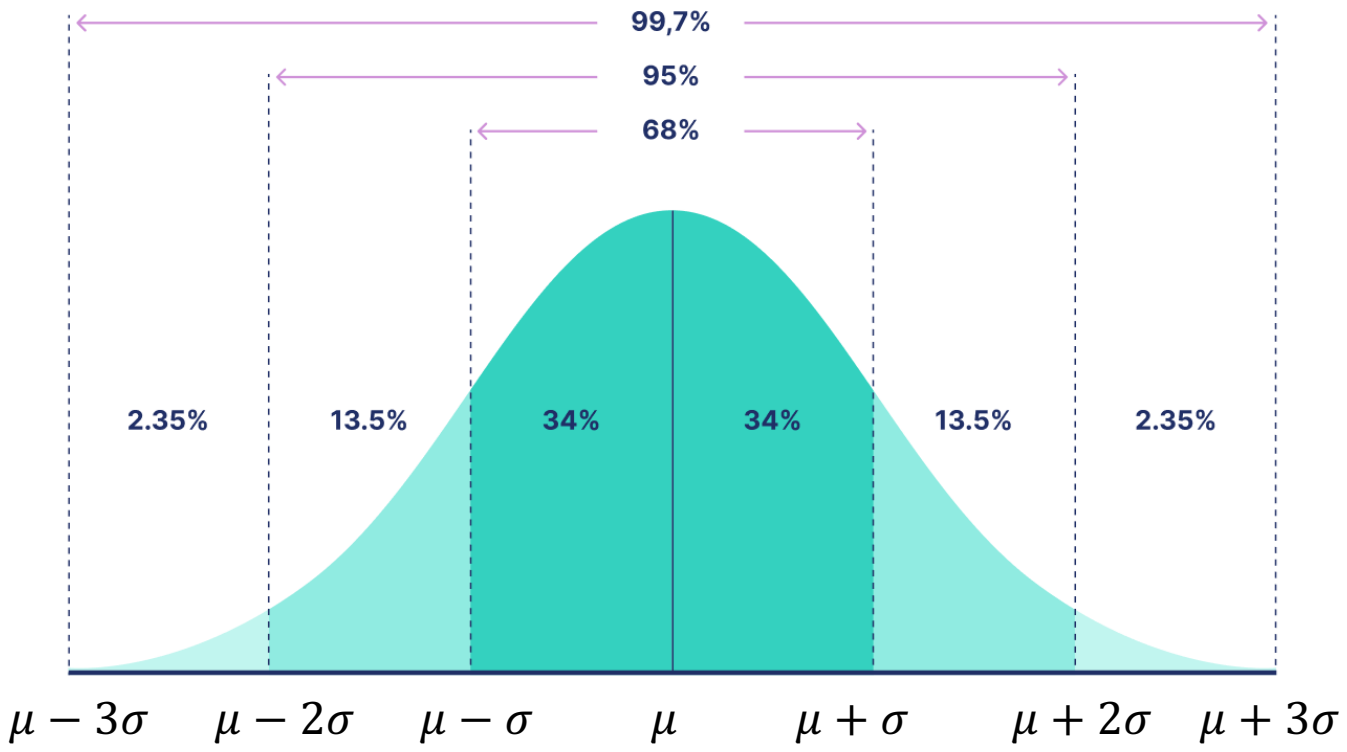
$$\sigma^2 \approx s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2$$

Variance

$$\sigma \approx s$$

Standard deviation

Normal Distribution



Data Representation and Visualization in R

Thank you for your attention!