Normal distribution

Tech Lead Data Science

Master en Data Science
2022-2023

Assembler
Institute of Technology

ÍNDICE

Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

How could we find the probability that a leaf is L mm?

Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

$62.5 \leq l < 63.5$

How could we find the probability that a leaf is L mm?

Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

$62.5 \leq l < 63.5$

How could we find the probability that a leaf is L mm?

How could we find the probability that the length L of a leaf lies in the interval $x_1 \leq L < x_2$?

Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

$62.5 \leq l < 63.5$

Let's make a histogram!

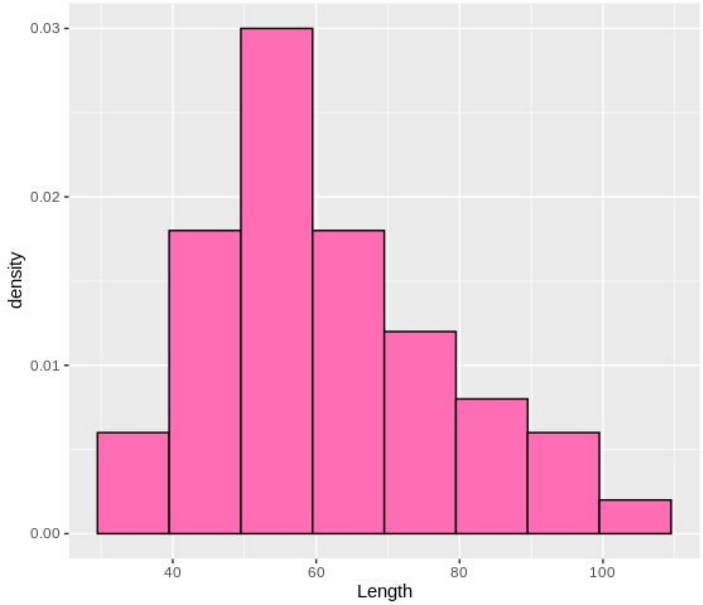
How could we find the probability that the length L of a leaf lies in the interval $x_1 \leq L < x_2$?

Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

$62.5 \leq l < 63.5$

How could we find the probability that the length L of a leaf lies in the interval $x_1 \leq L < x_2$?



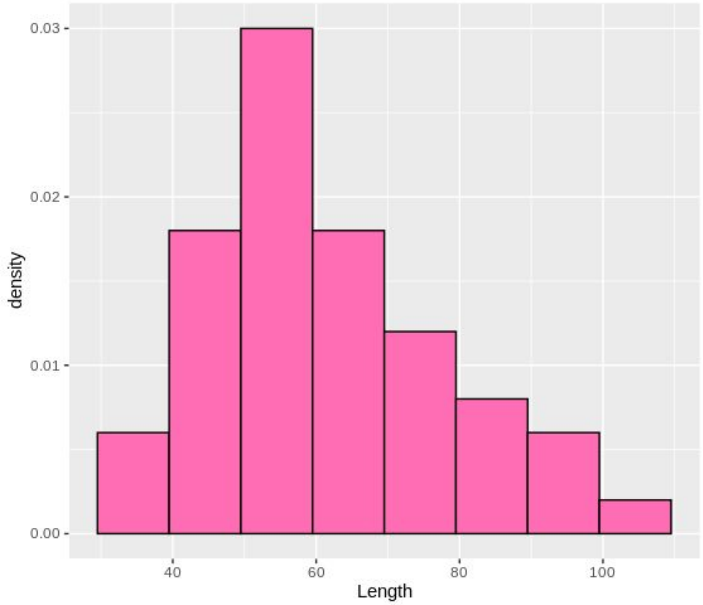
Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

→ $62.5 \leq l < 63.5$

How could we find the probability that the length L of a leaf lies in the interval $x_1 \leq L < x_2$?

$$R\ Freq\ density(x_i) = \frac{Rel\ Freq}{Class\ width}$$



Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

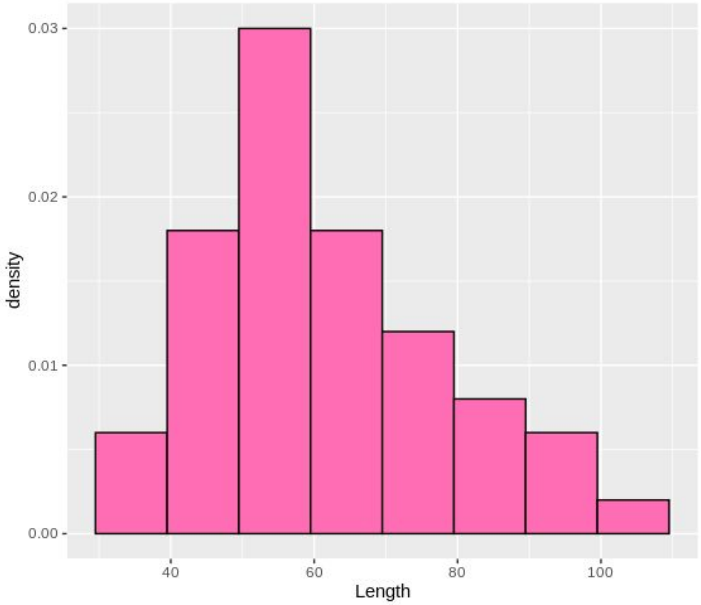
$\rightarrow 62.5 \leq l < 63.5$

How could we find the probability that the length L of a leaf lies in the interval $x_1 \leq L < x_2$?

$$R\ Freq\ density(x_i) = \frac{Rel\ Freq}{Class\ width}$$

Each area represents relative frequency.

$$P(x_i) \sim Relative\ Frequency(x_i)$$



Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

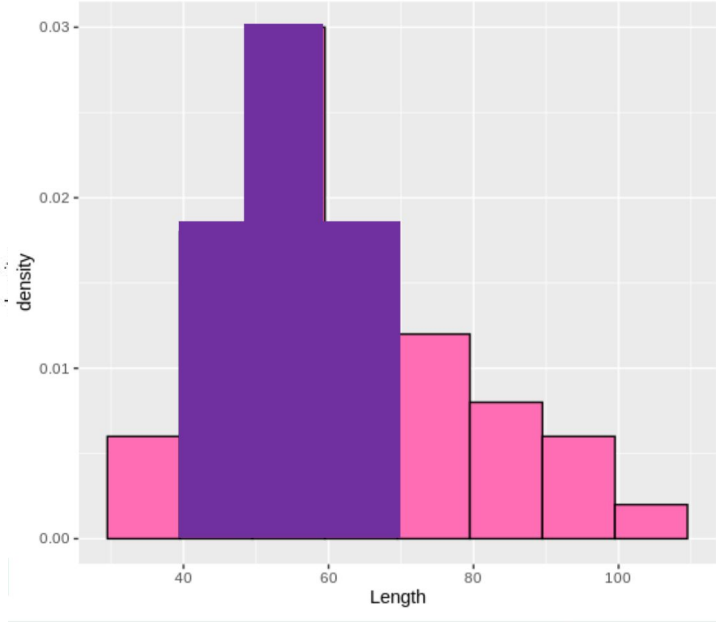
$62.5 \leq l < 63.5$

How could we find the probability that the length L of a leaf lies in the interval $x_1 \leq L < x_2$?

$$R\ Freq\ density(x_i) = \frac{Rel\ Freq}{Class\ width}$$

Each area represents relative frequency.

$$P(x_i) \sim Relative\ Frequency(x_i)$$



What is the probability that the length L lies in the interval $39.5 \leq L < 70.5$?

Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

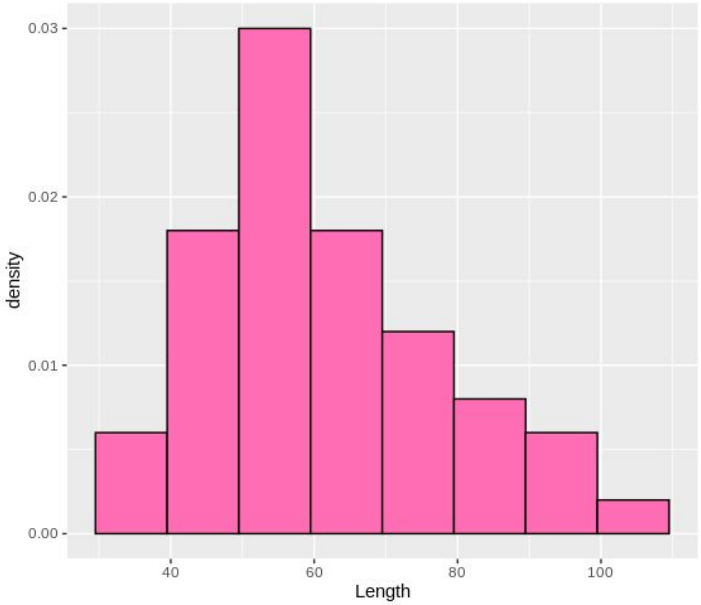
$\rightarrow 62.5 \leq l < 63.5$

How could we find the probability that the length L of a leaf lies in the interval $x_1 \leq L < x_2$?

$$R\ Freq\ density(x_i) = \frac{Rel\ Freq}{Class\ width}$$

Each area represents relative frequency.

$$P(x_i) \sim Relative\ Frequency(x_i)$$



What is the probability that the length L is equal to 39.5?

Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

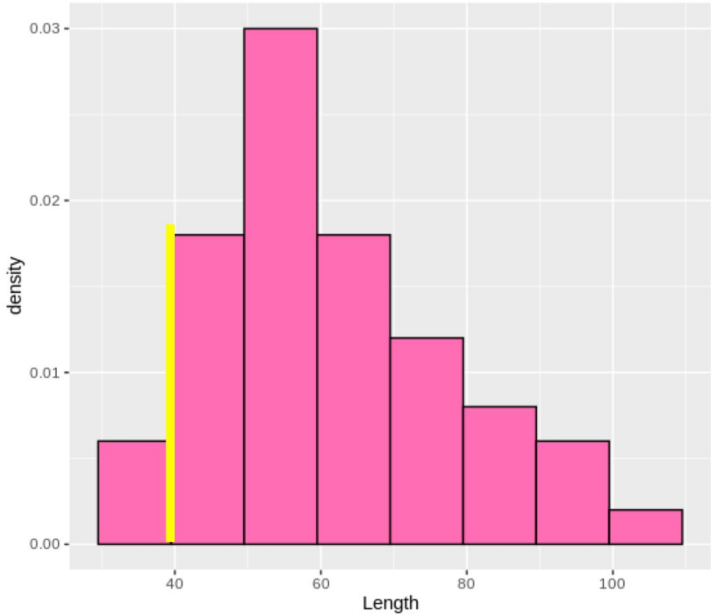
$62.5 \leq l < 63.5$

How could we find the probability that the length L of a leaf lies in the interval $x_1 \leq L < x_2$?

$$R\ Freq\ density(x_i) = \frac{Rel\ Freq}{Class\ width}$$

Each area represents relative frequency.

$$P(x_i) \sim Relative\ Frequency(x_i)$$



What is the probability that the length L is equal to 39.5?

Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

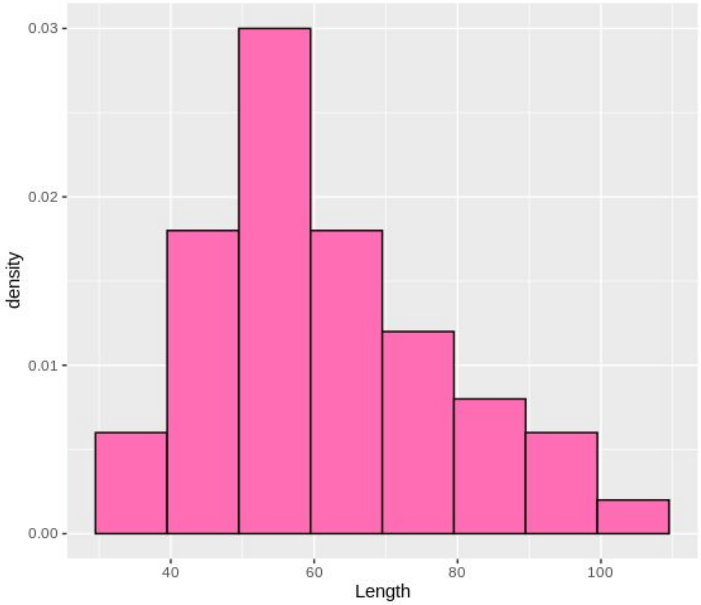
$\rightarrow 62.5 \leq l < 63.5$

How could we find the probability that the length L of a leaf lies in the interval $x_1 \leq L < x_2$?

$$R\ Freq\ density(x_i) = \frac{Rel\ Freq}{Class\ width}$$

Each area represents relative frequency.

$$P(x_i) \sim Relative\ Frequency(x_i)$$



$$P(39.5 \leq L < 70.5) =$$
$$P(39.5 < L < 70.5) =$$
$$P(39.5 < L \leq 70.5)$$

Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

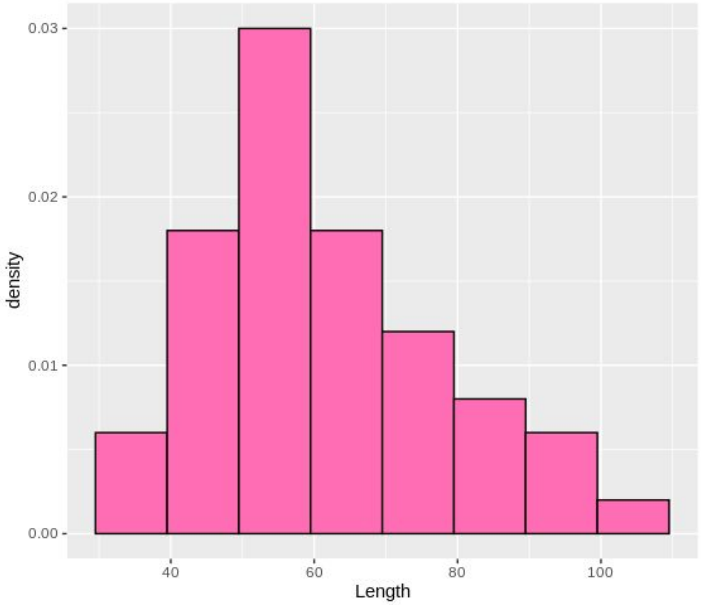
$62.5 \leq l < 63.5$

How could we find the probability that the length L of a leaf lies in the interval $x_1 \leq L < x_2$?

$$R\ Freq\ density(x_i) = \frac{Rel\ Freq}{Class\ width}$$

Each area represents relative frequency.

$$P(x_i) \sim Relative\ Frequency(x_i)$$



$$P(x_1 \leq L < x_2) =$$
$$P(x_1 < L < x_2) =$$
$$P(x_1 < L \leq x_2)$$

$$P(x_i) = 0$$

Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

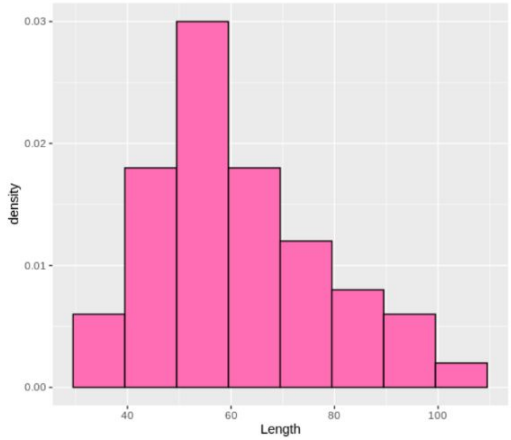
60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

What happens if we reduce the length of the class width ?

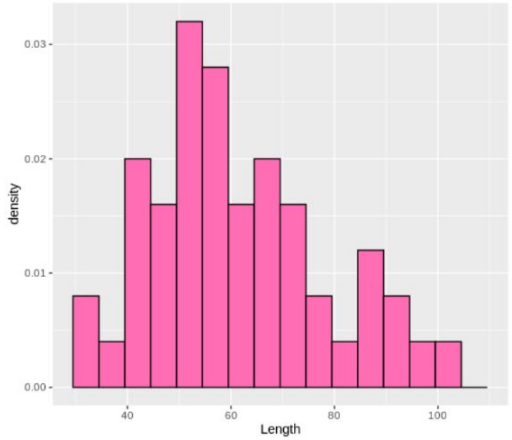
Consider the lengths, in mm, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

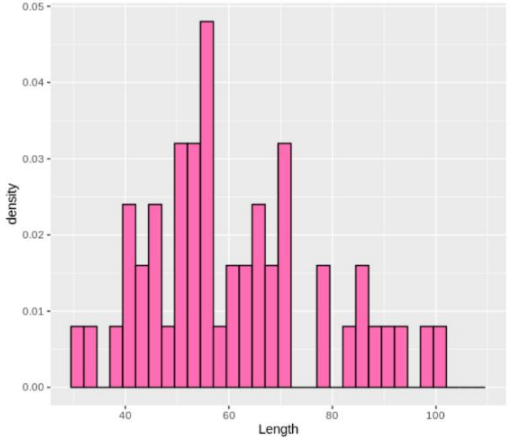
What happens if we reduce the length of the class width ?



Class width: 10

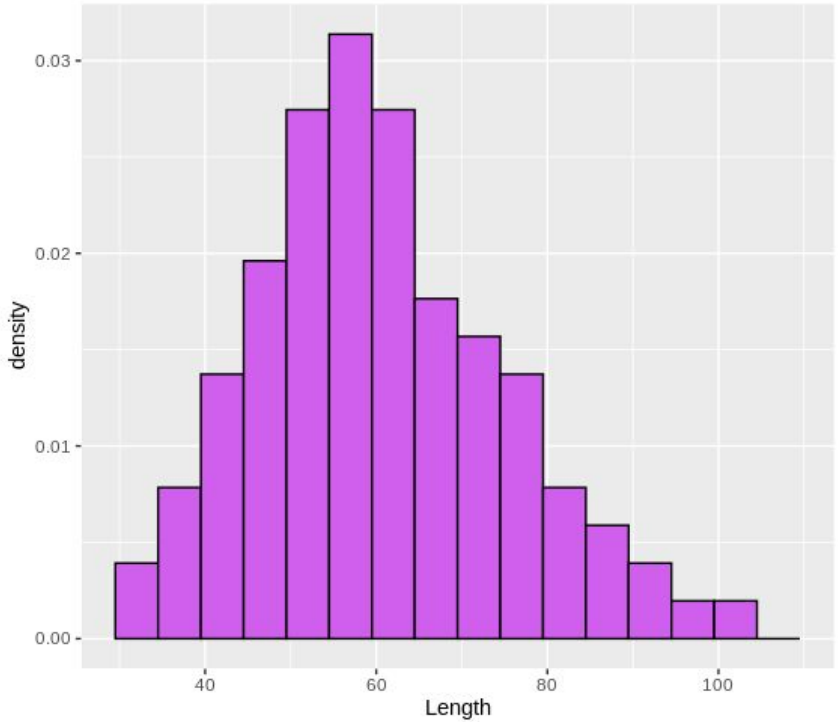


Class width: 5



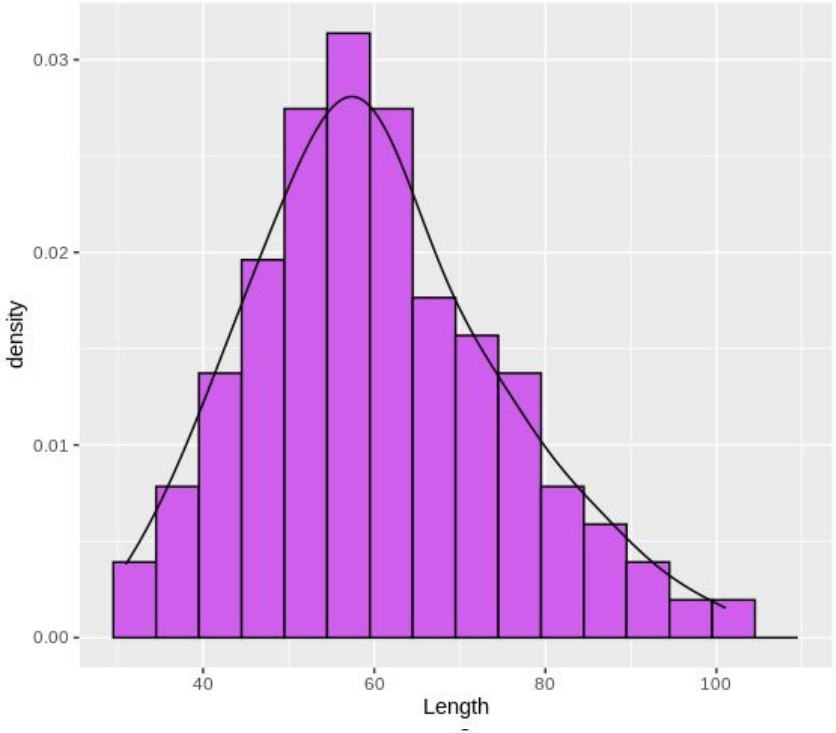
Class width: 2.5

We'll consider a better model with more data and smaller class widths. Imagine we have a dataset of size 100 and class widths of 5.



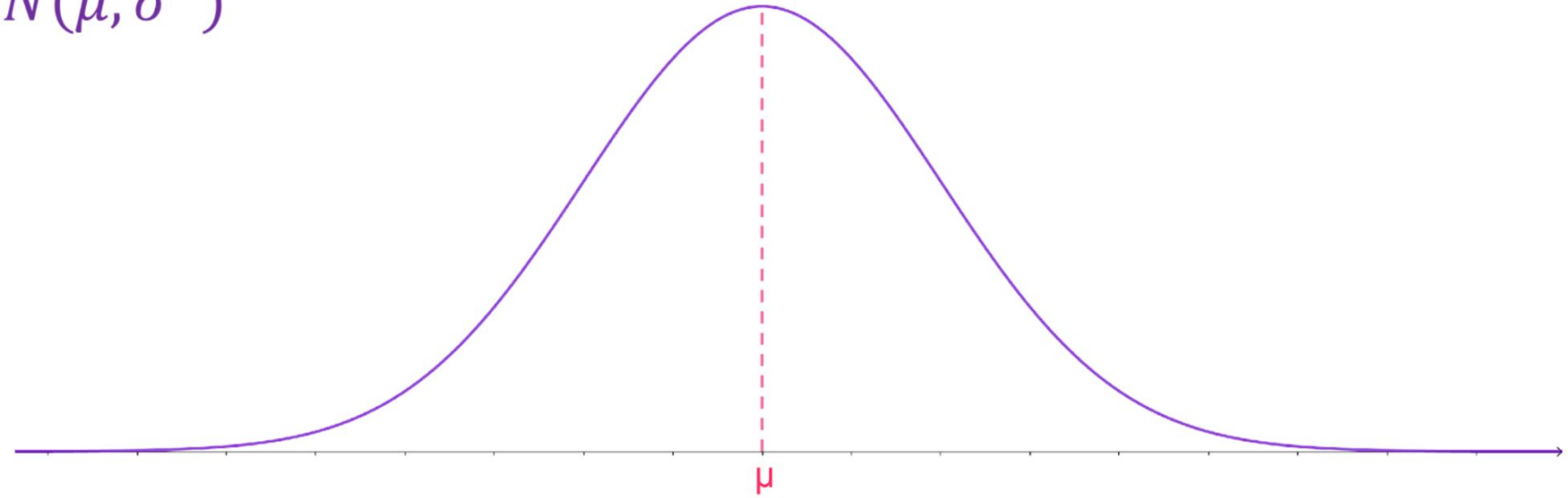
We'll consider a better model with more data and smaller class widths. Imagine we have a dataset of size 100 and class widths of 5.

The length of the coffee leaves follows a **normal distribution**



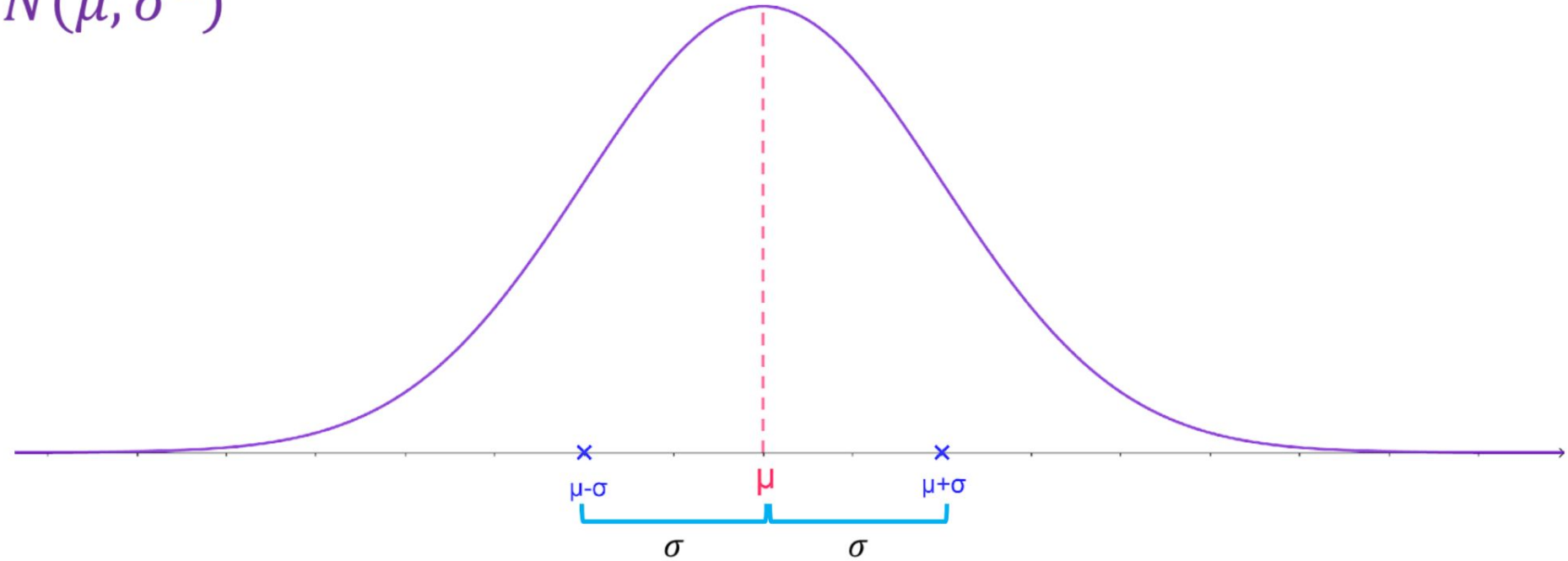
NORMAL DISTRIBUTION

$N(\mu, \sigma^2)$



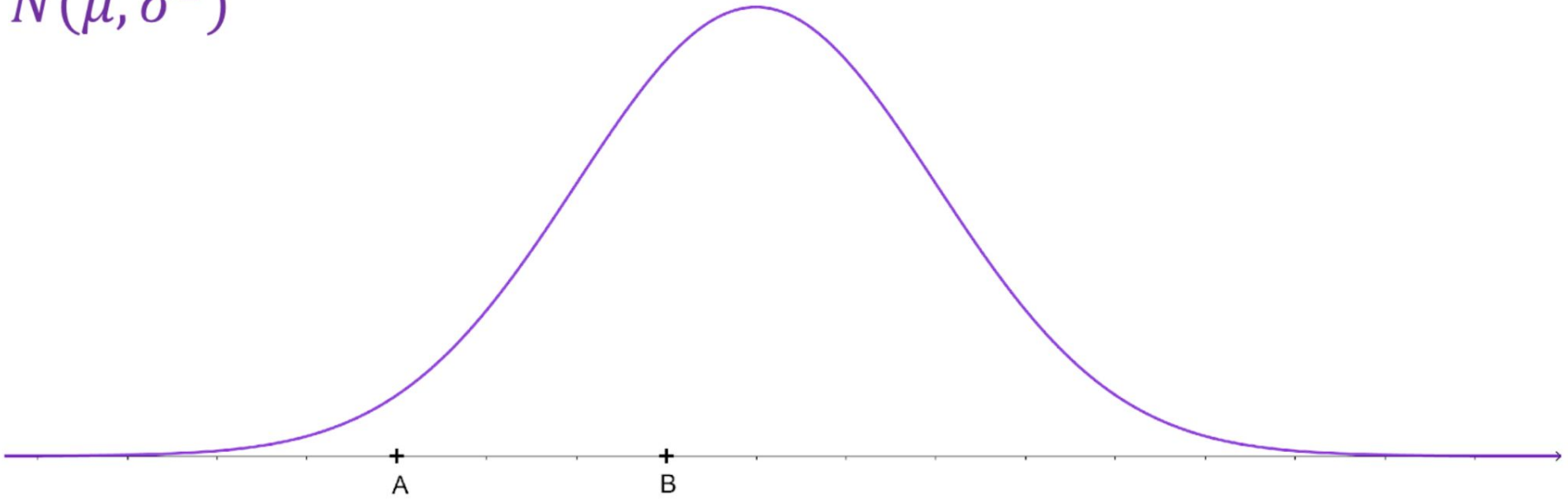
NORMAL DISTRIBUTION

$N(\mu, \sigma^2)$



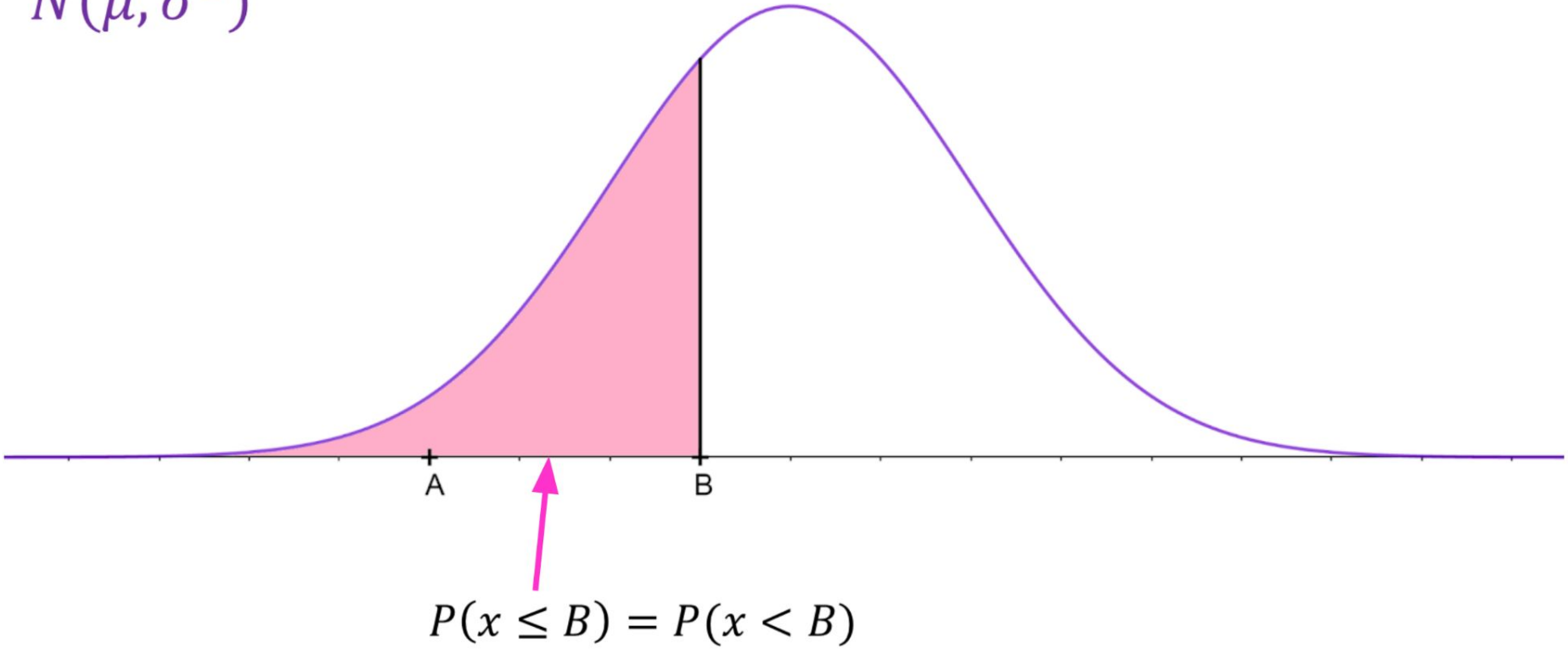
NORMAL DISTRIBUTION

$N(\mu, \sigma^2)$



NORMAL DISTRIBUTION

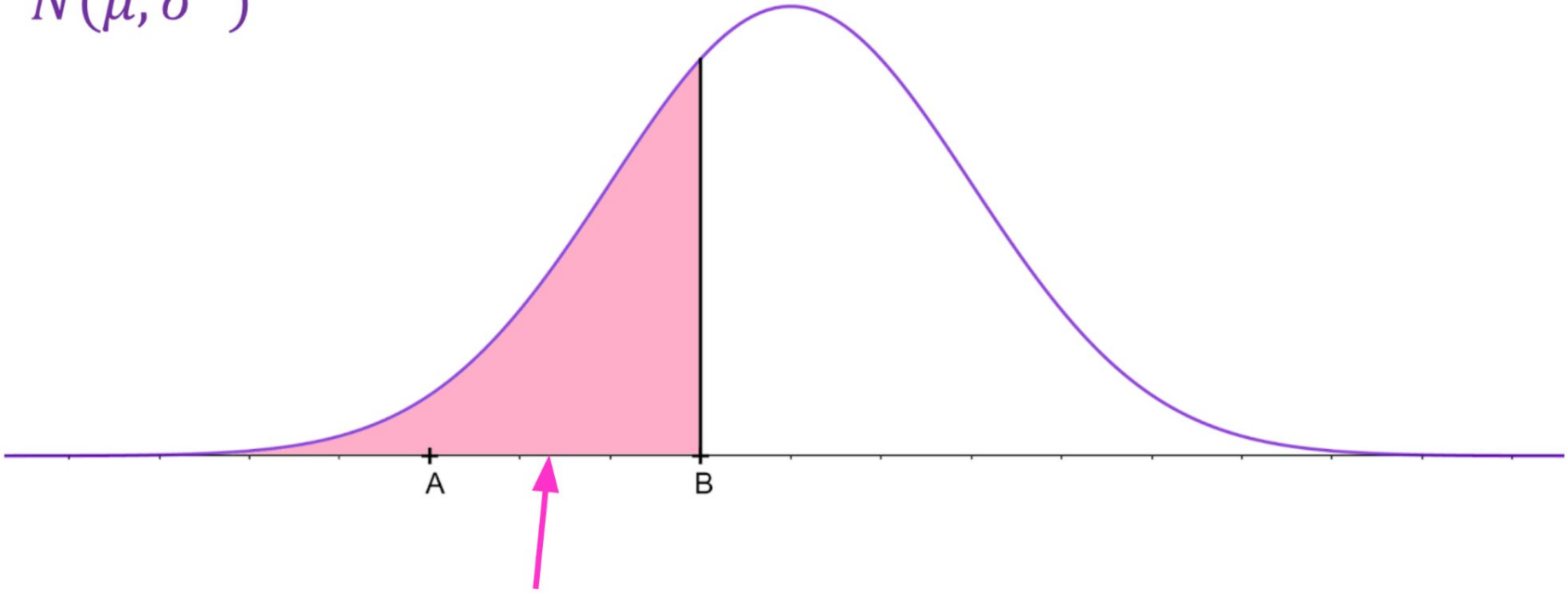
$N(\mu, \sigma^2)$



$$P(x \leq B) = P(x < B)$$

NORMAL DISTRIBUTION

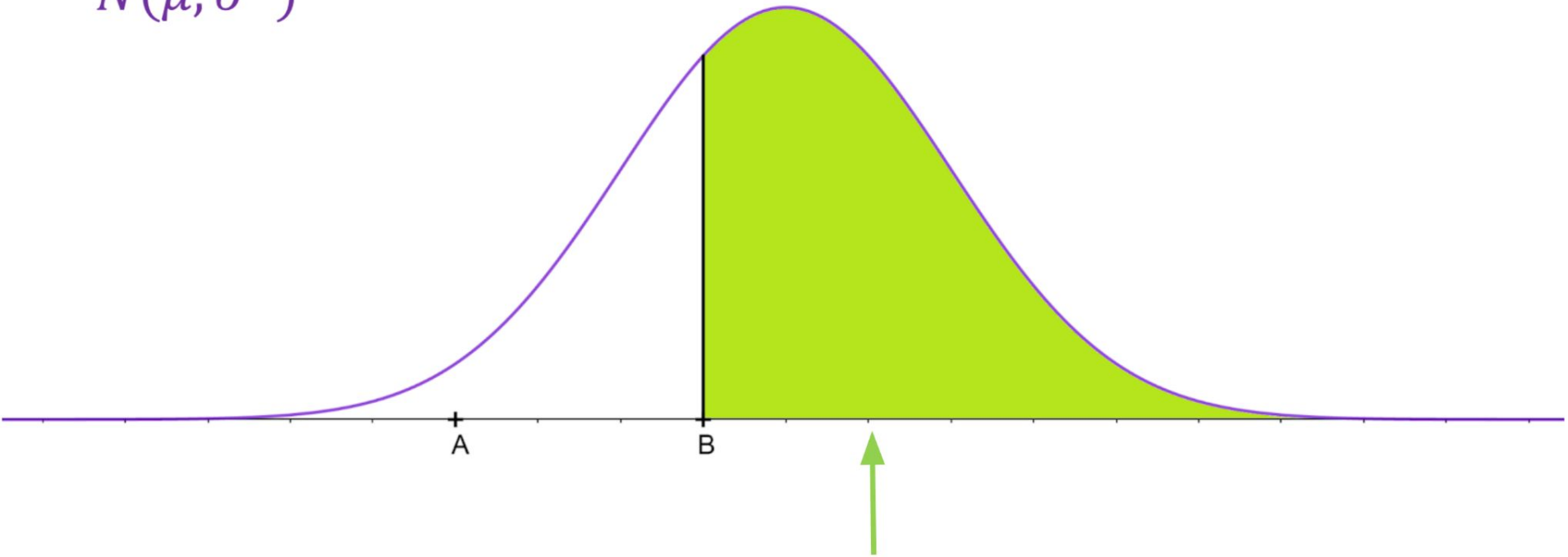
$N(\mu, \sigma^2)$



$$F(B) = P(x \leq B) = P(x < B)$$

NORMAL DISTRIBUTION

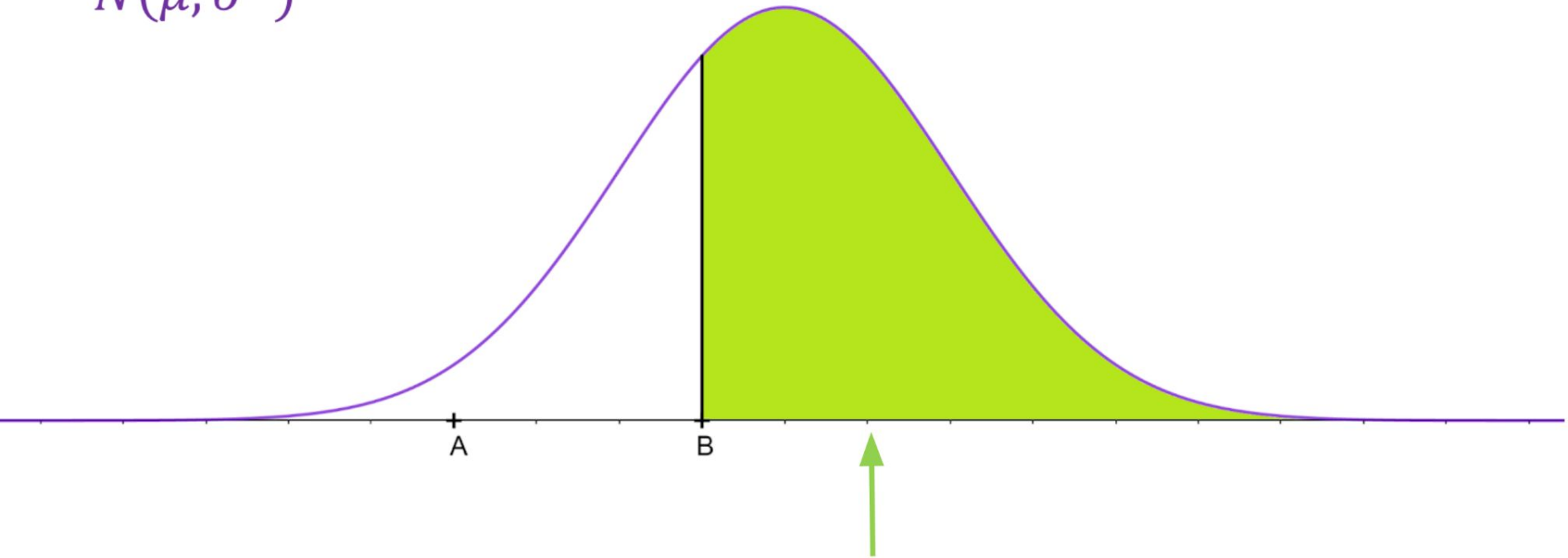
$N(\mu, \sigma^2)$



$P(x \geq B) = P(x > B)$

NORMAL DISTRIBUTION

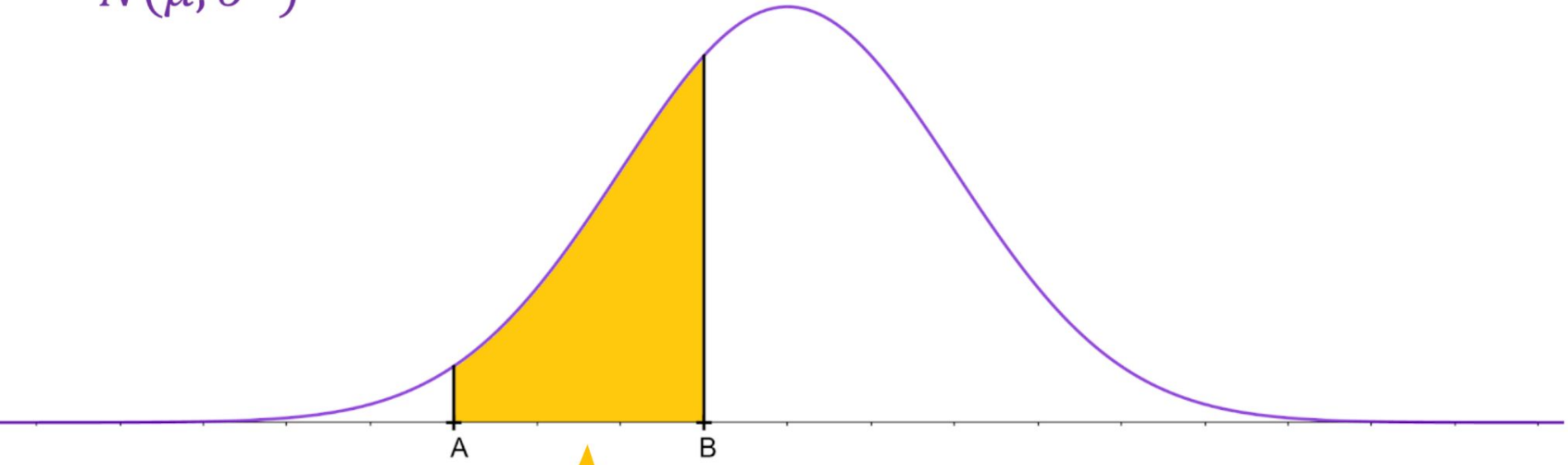
$N(\mu, \sigma^2)$



$$1 - F(B) = P(x \geq B) = P(x > B)$$

NORMAL DISTRIBUTION

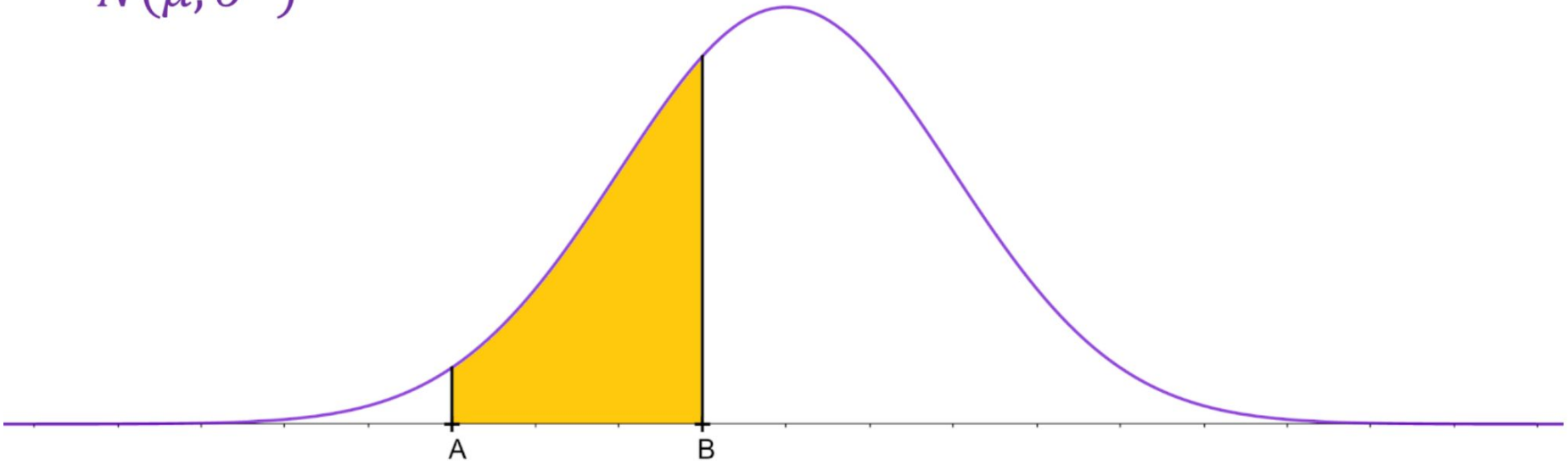
$N(\mu, \sigma^2)$



$$P(A < x < B) = P(A < x \leq B) = P(A \leq x < B) = P(A \leq x \leq B)$$

NORMAL DISTRIBUTION

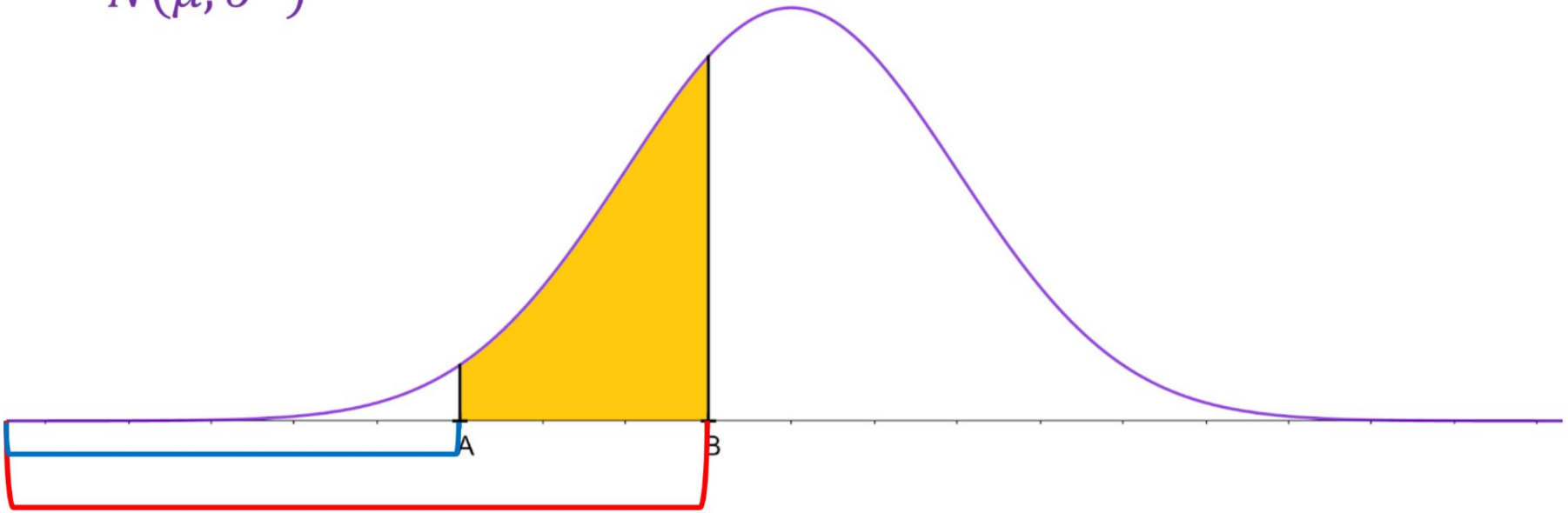
$N(\mu, \sigma^2)$



$P(A \leq x \leq B) =$

NORMAL DISTRIBUTION

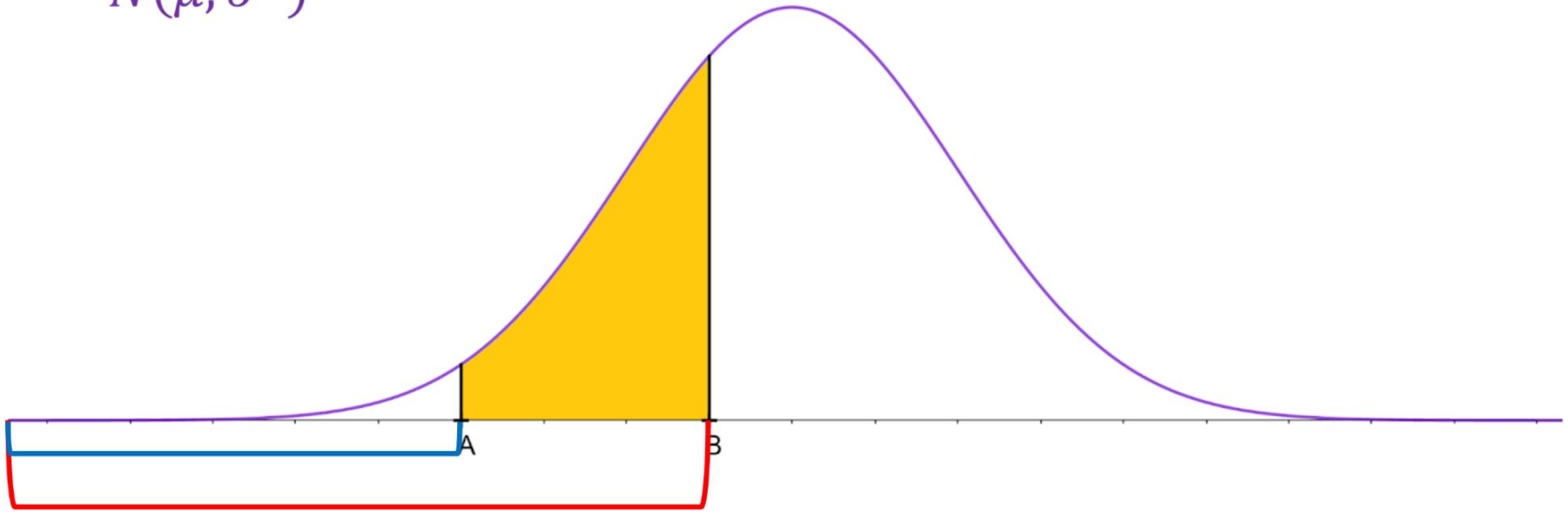
$N(\mu, \sigma^2)$



$$P(A \leq x \leq B) = P(x \leq B) - P(x \leq A)$$

NORMAL DISTRIBUTION

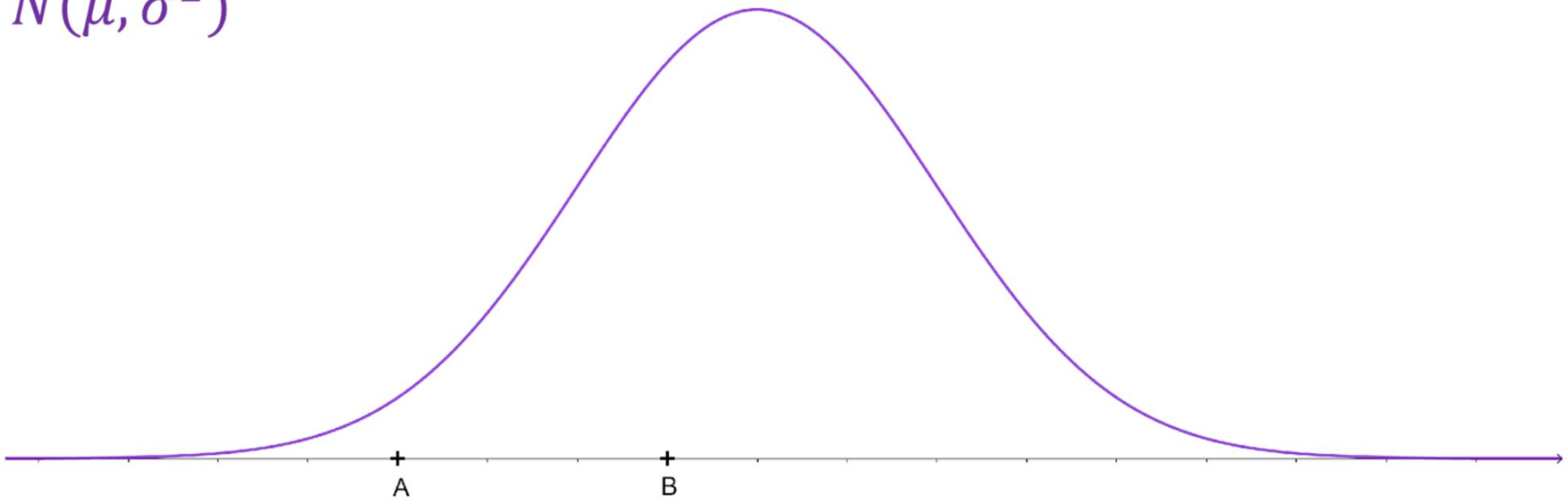
$N(\mu, \sigma^2)$



$$P(A \leq x \leq B) = P(x \leq B) - P(x \leq A) = F(B) - F(A)$$

NORMAL DISTRIBUTION

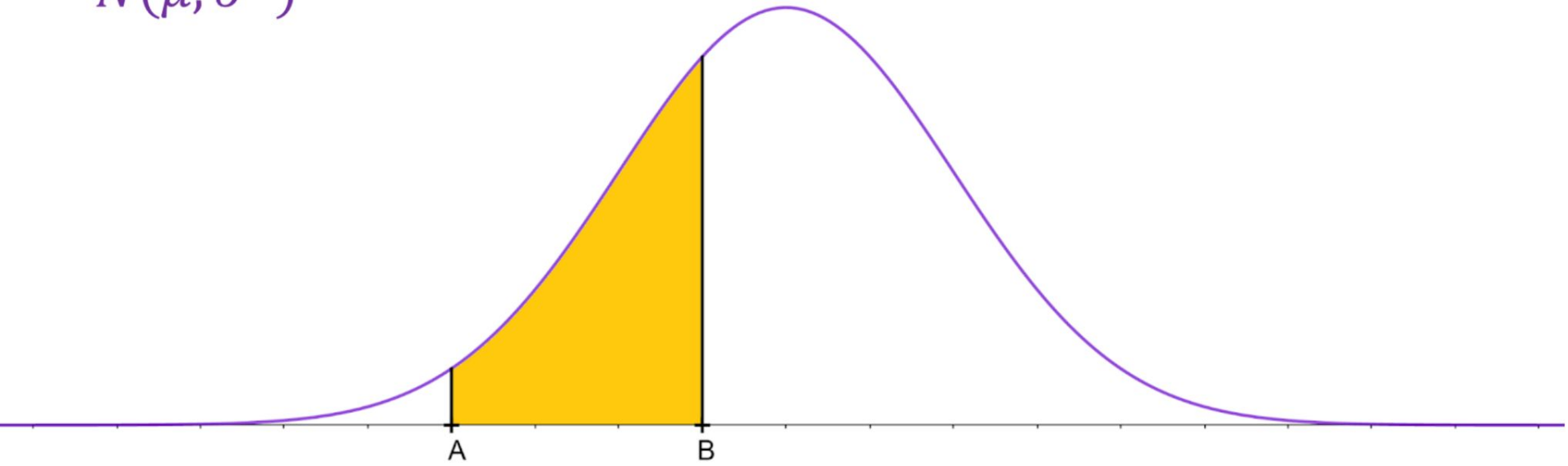
$N(\mu, \sigma^2)$



How do we calculate the areas?🤔

NORMAL DISTRIBUTION

$N(\mu, \sigma^2)$

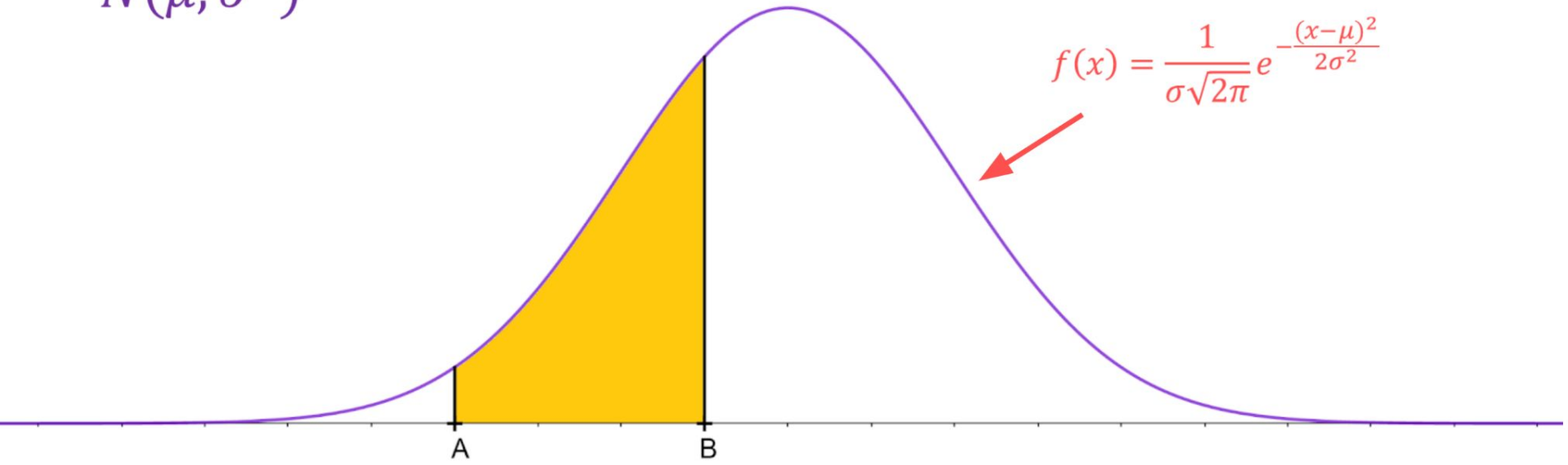


$P(A \leq x \leq B) =$

NORMAL DISTRIBUTION

$N(\mu, \sigma^2)$

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

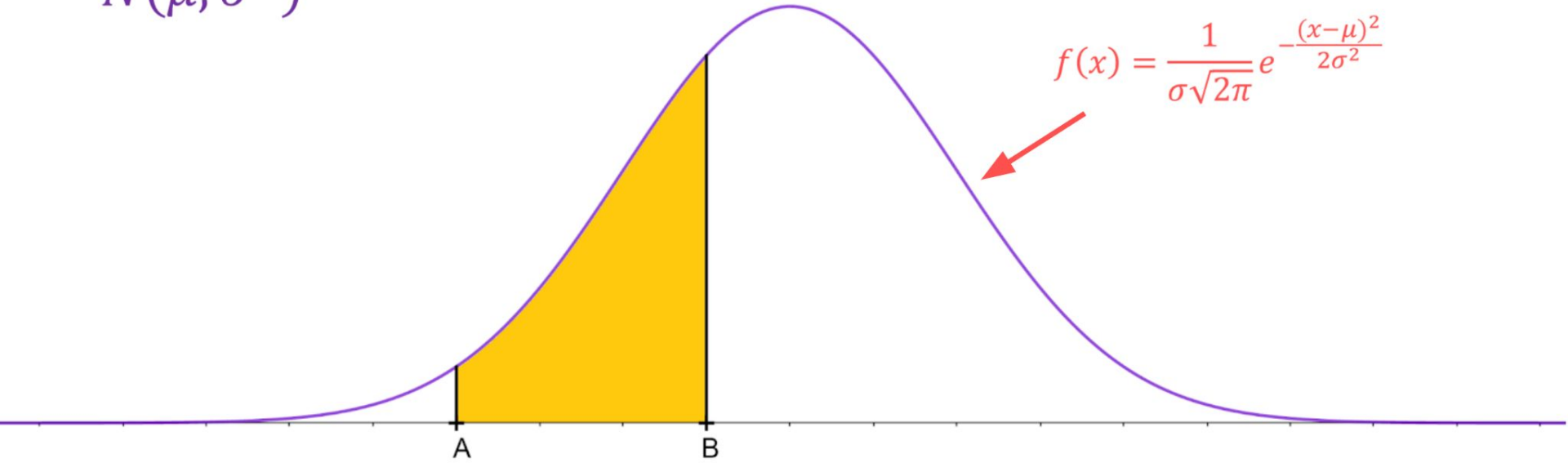


$$P(A \leq x \leq B) =$$

NORMAL DISTRIBUTION

$N(\mu, \sigma^2)$

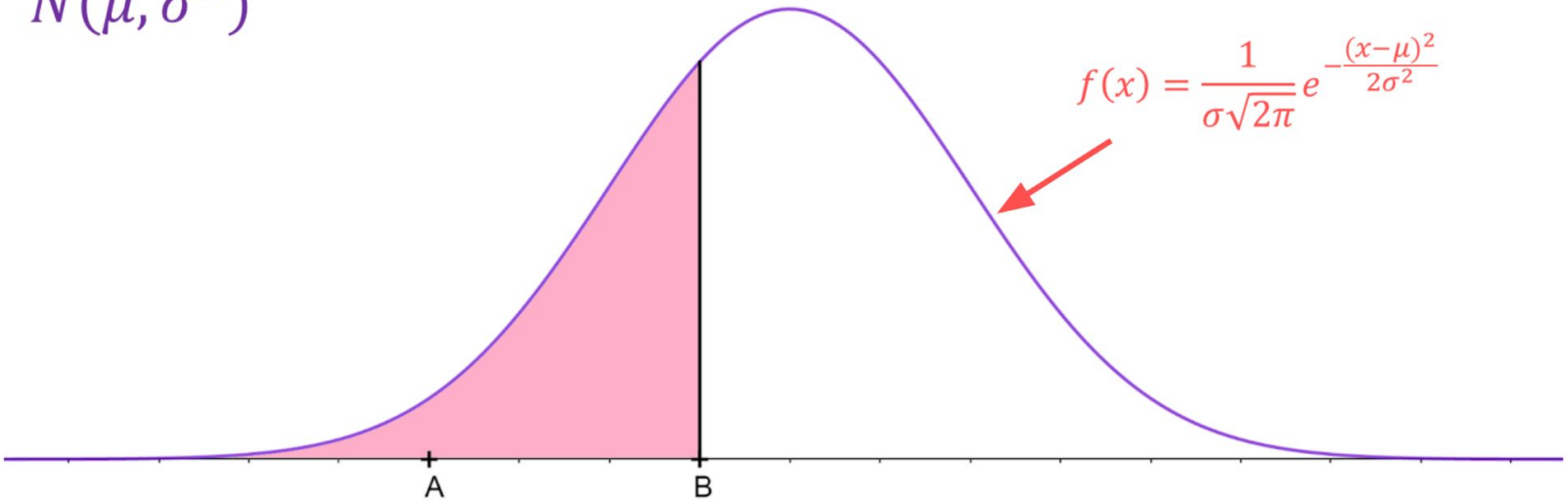
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$P(A \leq x \leq B) = \int_A^B \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \cdots$$

NORMAL DISTRIBUTION

$N(\mu, \sigma^2)$

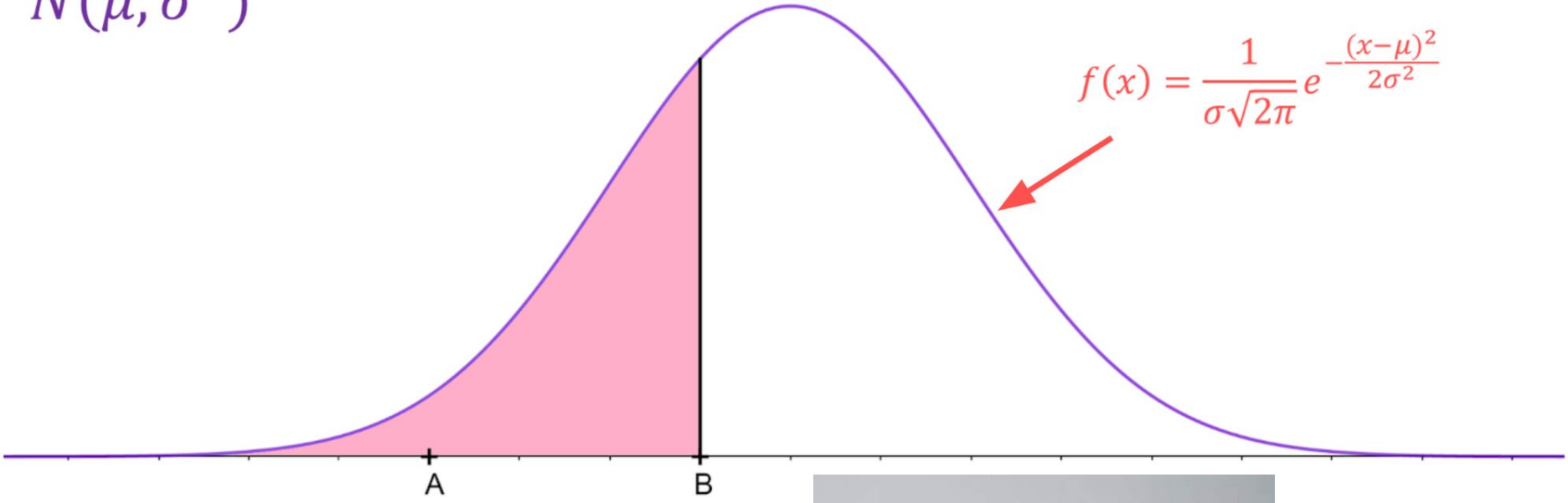


$P(x \leq B) =$

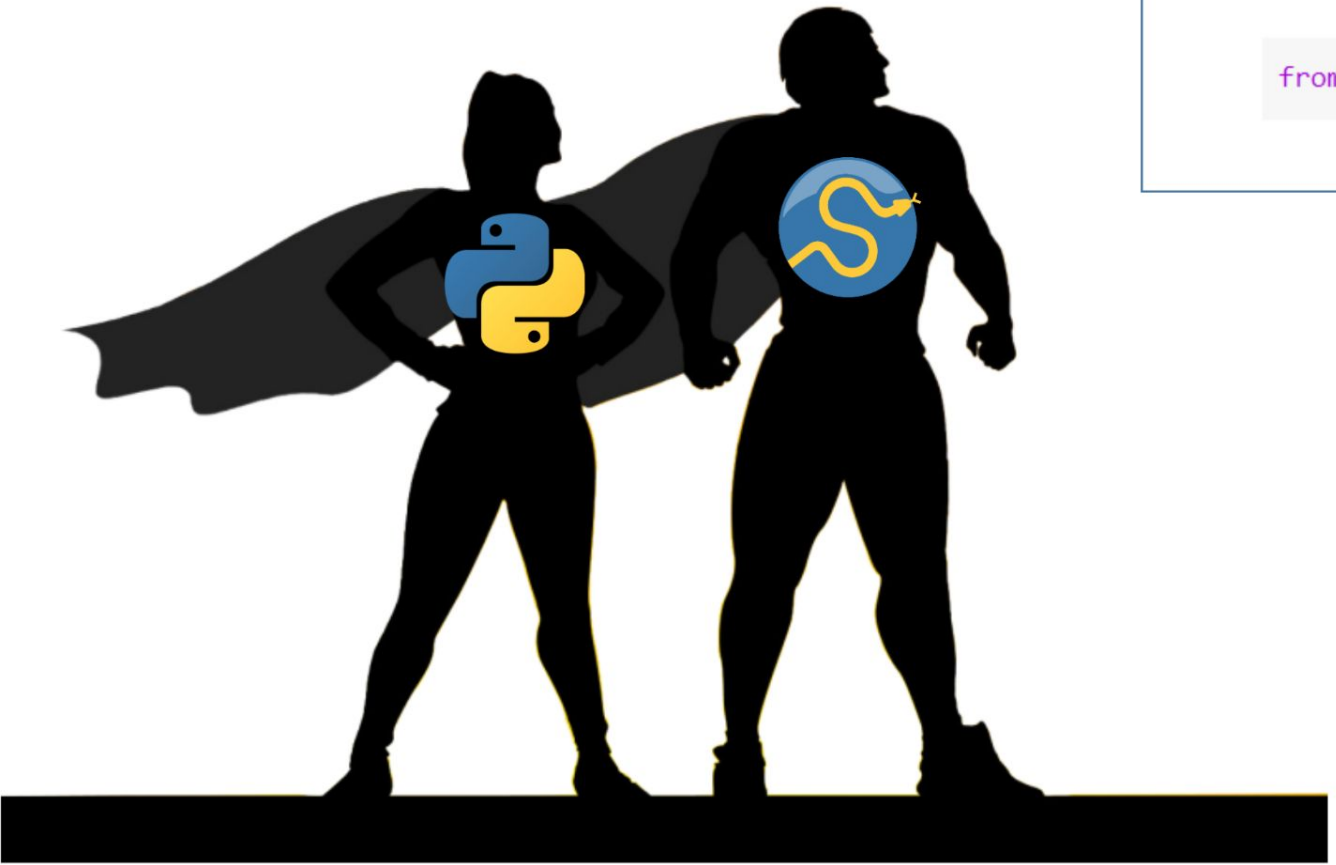
NORMAL DISTRIBUTION

$N(\mu, \sigma^2)$

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

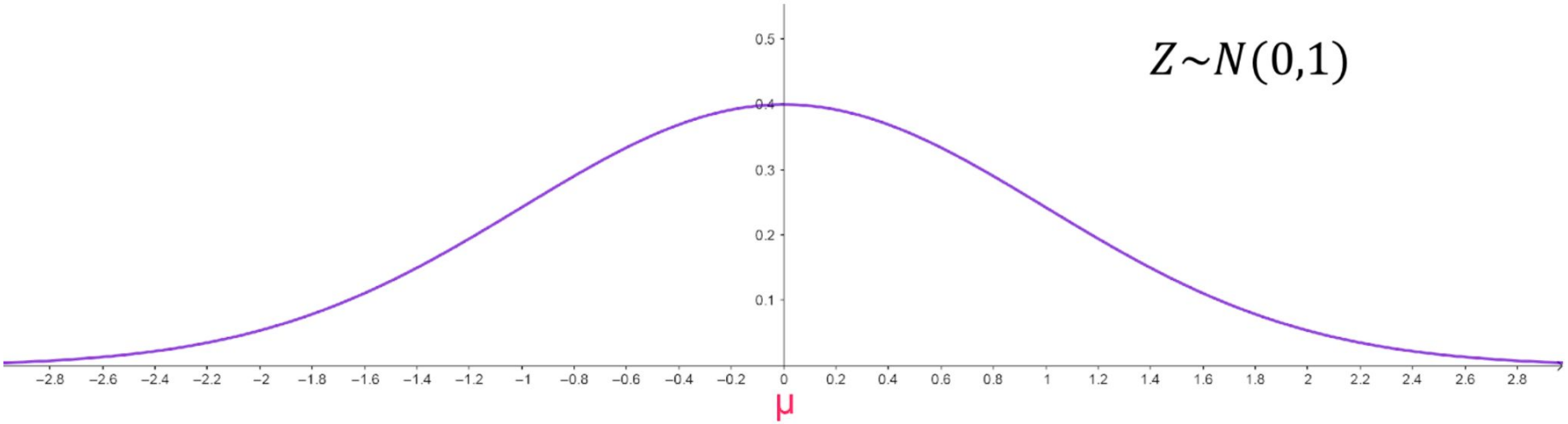


$$P(x \leq B) = \int_{-\infty}^{B} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



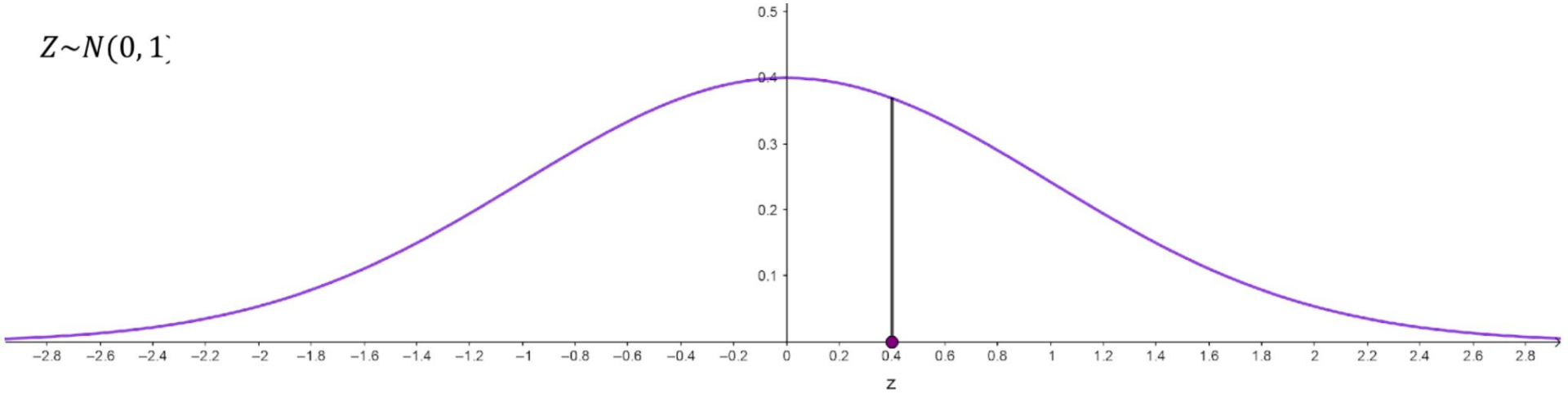
```
from scipy.stats import norm
```

STANDARD NORMAL DISTRIBUTION



STANDARD NORMAL DISTRIBUTION

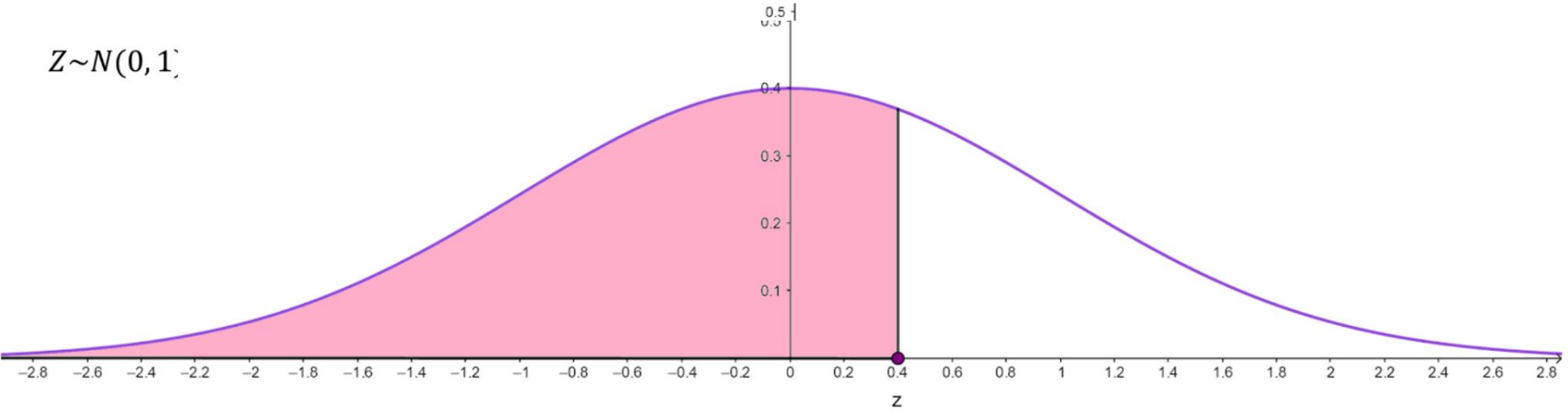
$Z \sim N(0, 1)$



¿$P(Z \leq z)$?

STANDARD NORMAL DISTRIBUTION

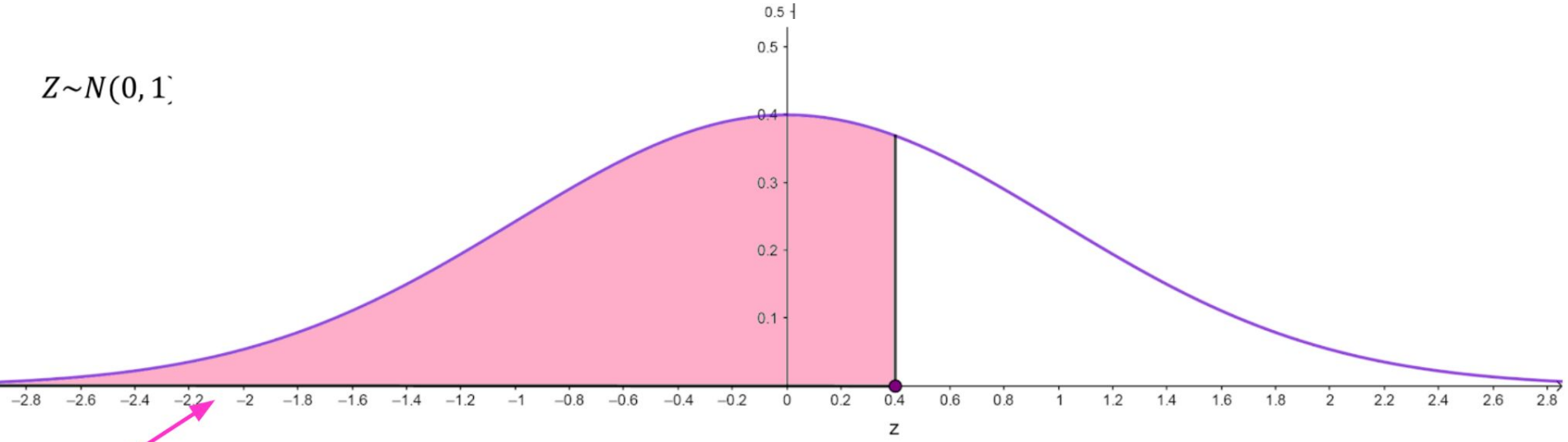
$Z \sim N(0,1)$



¿ $P(Z \leq z)$?

STANDARD NORMAL DISTRIBUTION

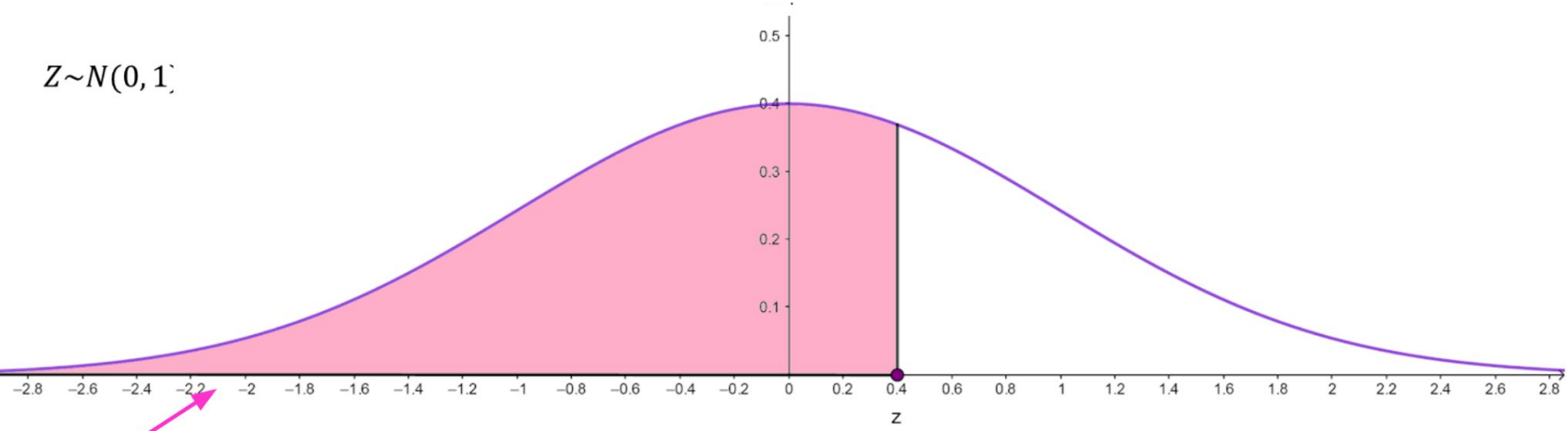
$Z \sim N(0,1)$



$P(Z \leq z) = \Phi(z)$

STANDARD NORMAL DISTRIBUTION

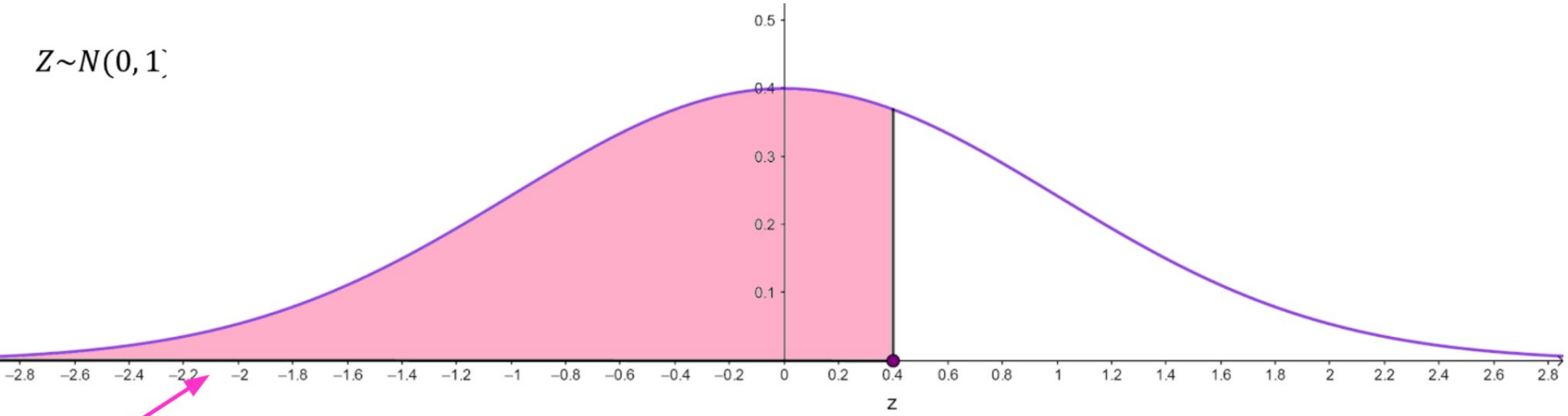
$Z \sim N(0,1)$



$P(Z \leq 0.4) = \Phi(0.4)$

STANDARD NORMAL DISTRIBUTION

$Z \sim N(0, 1)$



$P(Z \leq 0.4) = \Phi(0.4)$

cdf(x, loc=0, scale=1)

Cumulative distribution function.

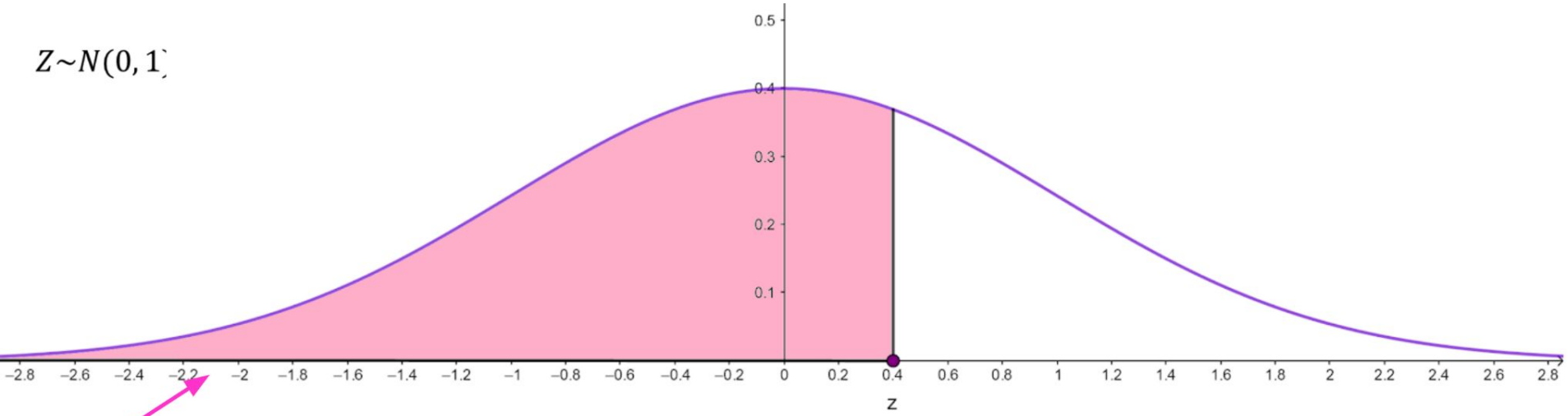
```
mean, var = 0, 1

# P(Z<=0.4)
norm.cdf(0.4, loc=mean, scale=var)
```

```
0.6554217416103242
```

STANDARD NORMAL DISTRIBUTION

$Z \sim N(0, 1)$



$P(Z \leq 0.4) = \Phi(0.4)$

cdf(x, loc=0, scale=1)

Cumulative distribution function.

```
mean, var = 0, 1

# P(Z<=0.4)
norm.cdf(0.4, loc=mean, scale=var)
```

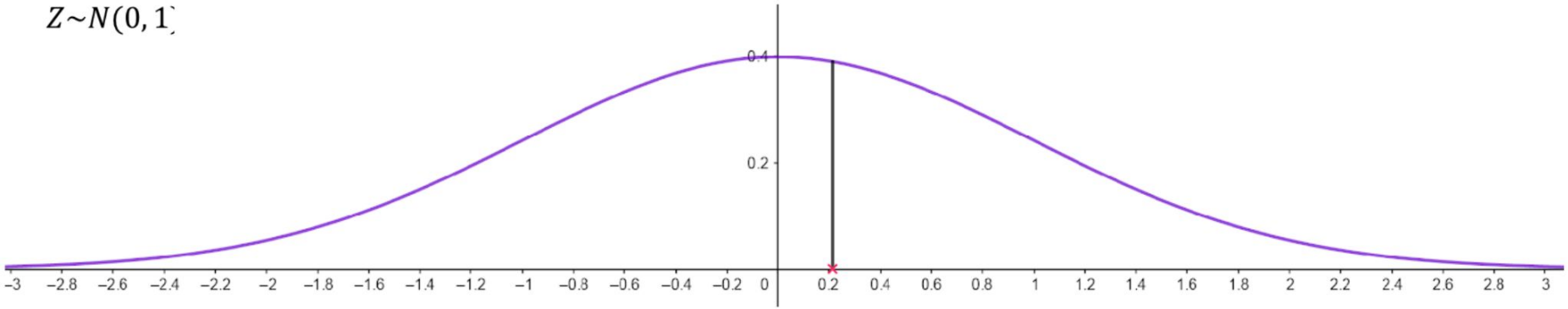
```
0.6554217416103242
```

```
# P(Z<=0.4)
norm.cdf(0.4)
```

```
0.6554217416103242
```

STANDARD NORMAL DISTRIBUTION

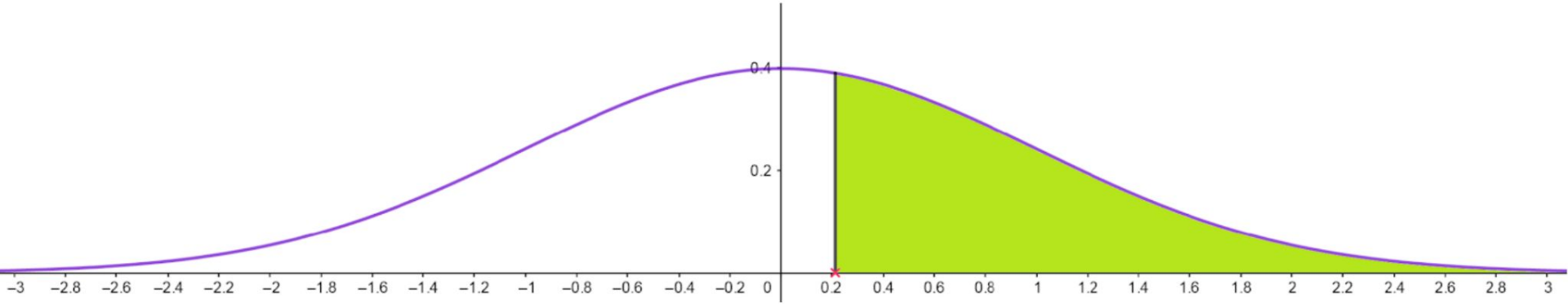
$Z \sim N(0, 1)$



$P(Z \geq 0.213)$

STANDARD NORMAL DISTRIBUTION

$Z \sim N(0,1)$



$$P(Z \geq 0.213) = 1 - \Phi(0.213)$$

```
# P(Z>=0.213)
1-norm.cdf(0.213)
```

```
0.415663481341247
```

STANDARD NORMAL DISTRIBUTION

$$Z \sim N(0, 1)$$

$$P(1.15 \leq Z \leq 1.35) = \Phi(1.35) - \Phi(1.15)$$

```
# P(1.15<=Z<=1.35)
norm.cdf(1.35)-norm.cdf(1.15)
```

```
0.036563944199748 4
```

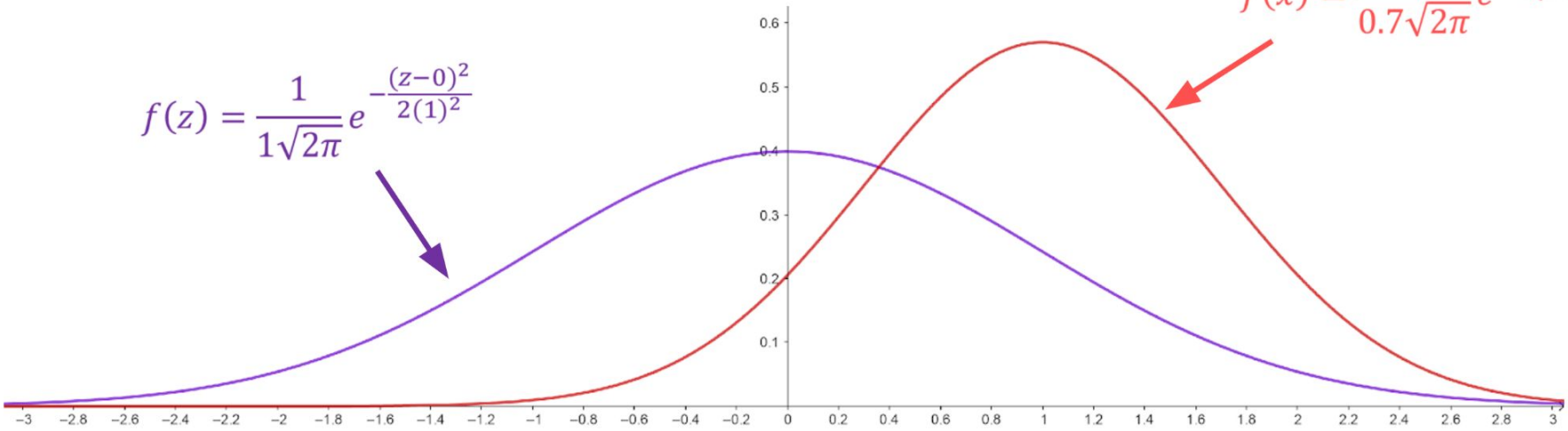
NORMAL DISTRIBUTION

$Z \sim N(0,1)$

$X \sim N(1, 0.7^2)$

$$f(z) = \frac{1}{1\sqrt{2\pi}} e^{-\frac{(z-0)^2}{2(1)^2}}$$

$$f(x) = \frac{1}{0.7\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2(0.7)^2}}$$



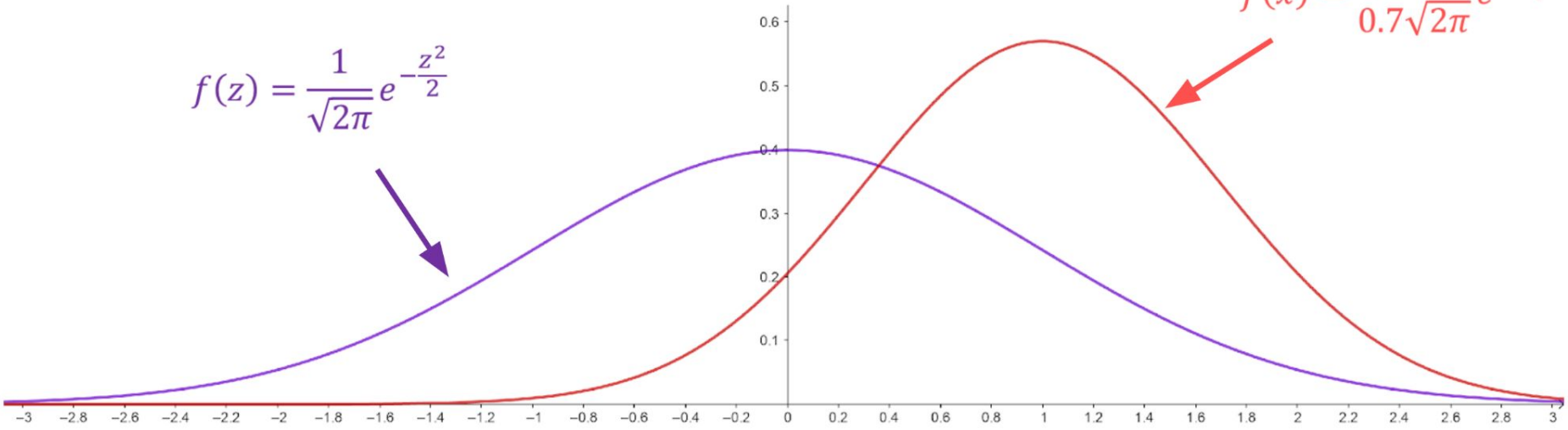
NORMAL DISTRIBUTION

$Z \sim N(0,1)$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

$X \sim N(1, 0.7^2)$

$$f(x) = \frac{1}{0.7\sqrt{2\pi}} e^{-\frac{(x-1)^2}{2(0.7)^2}}$$



NORMAL DISTRIBUTION

$$X \sim N(205, 400)$$

$$P(X \leq 230)$$

```
mean, var = 205, 400
```

```
# P(X<=230)
norm.cdf(230, loc=mean, scale=sqrt(var))
```

```
0.8943502263331446
```

NORMAL DISTRIBUTION

Dado $X \sim N(6,4)$, hallar el valor de s tal que $P(X \leq s) = 0.6500$

```
mean, var = 6, 4
```

ppf(q, loc=0, scale=1)

Percent point function (inverse of cdf — percentiles).

```
[11] norm.ppf(0.65, loc=mean, scale=sqrt(var))

     6.770640932815136
```

Two friends Sarah and Hannah often go to the post office together. They travel on Sarah's scooter. Sarah always drives Hannah to the post office and drops her off there. Sarah then drives around until she is ready to pick Hannah up some time later. Their experience has been that the time Hannah takes in the post office can be approximated by a normal distribution with mean 6 minutes and standard deviation 1.3 minutes. **How many minutes after having dropped Hannah off should Sarah return if she wants to be at least 95% certain that Hannah will not keep her waiting?**

Two friends Sarah and Hannah often go to the post office together. They travel on Sarah's scooter. Sarah always drives Hannah to the post office and drops her off there. Sarah then drives around until she is ready to pick Hannah up some time later. Their experience has been that the time Hannah takes in the post office can be approximated by a normal distribution with mean 6 minutes and standard deviation 1.3 minutes. **How many minutes after having dropped Hannah off should Sarah return if she wants to be at least 95% certain that Hannah will not keep her waiting?**

$$T = \text{"}tiempo\ que\ Hannah\ demora\ en\ el\ correo\ en\ un\ día\ aleatorio\text{"} \Rightarrow T \sim N(6, 1.3^2)$$

Two friends Sarah and Hannah often go to the post office together. They travel on Sarah's scooter. Sarah always drives Hannah to the post office and drops her off there. Sarah then drives around until she is ready to pick Hannah up some time later. Their experience has been that the time Hannah takes in the post office can be approximated by a normal distribution with mean 6 minutes and standard deviation 1.3 minutes. **How many minutes after having dropped Hannah off should Sarah return if she wants to be at least 95% certain that Hannah will not keep her waiting?**

$$T = \text{"}tiempo\ que\ Hannah\ demora\ en\ el\ correo\ en\ un\ día\ aleatorio\text{"} \Rightarrow T \sim N(6, 1.3^2)$$

$$P(T \leq t) = 95\%$$

$$P(T \leq t) = 0.95$$

```
mean, std = 6, 1.3
```

```
# P(T≤t)=0.95
norm.ppf(0.95, loc=mean, scale=std)
```

```
8.138309715036915
```

```
norm.cdf(8.2, loc=mean, scale=std)
```

```
0.9547063390392431
```

```
norm.cdf(8.1, loc=mean, scale=std)
```

```
0.946886285109998
```

Two friends Sarah and Hannah often go to the post office together. They travel on Sarah's scooter. Sarah always drives Hannah to the post office and drops her off there. Sarah then drives around until she is ready to pick Hannah up some time later. Their experience has been that the time Hannah takes in the post office can be approximated by a normal distribution with mean 6 minutes and standard deviation 1.3 minutes. **How many minutes after having dropped Hannah off should Sarah return if she wants to be at least 95% certain that Hannah will not keep her waiting?**

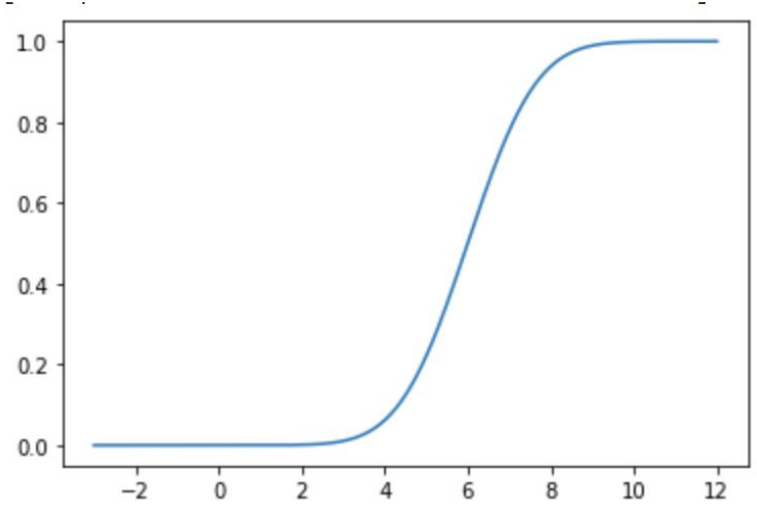
$$T = "tiempo\ que\ Hannah\ demora\ en\ el\ correo\ en\ un\ día\ aleatorio" \Rightarrow T \sim N(6, 1.3^2)$$

$$P(T \leq t) = 95\%$$
$$P(T \leq t) = 0.95$$

```
mean, std = 6, 1.3
```

```
# P(T≤t)=0.95
norm.ppf(0.95, loc=mean, scale=std)
```

```
8.138309715036915
```



SUMMARY

Today we've learnt:

- Introduction to Normal distribution
  - Formula
  - Definition
  - Graph
- Calculating probabilities of a random variable that follows a normal distribution using Python
- Modelling a real life problem using normal distribution

Assembler
Institute of Technology