# 통계분석

# Statistical Analysis

# Bayes Theorem

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

# Bayes Theorem

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$$

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

$$P(A|B) \leftrightarrow P(B|A)$$

Converting between conditional probabilities back and forth

# Bayes Theorem: Generalization

- Mutually exclusive events : ***Law of Total Probability***

$$B_1 \cup B_2 = S$$

$$B_1 \cap B_2 = \phi$$

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2)$$

# Bayes Theorem: Generalization

- Mutually exclusive events : ***Law of Total Probability***

$$B_1 \cup B_2 = S$$

$$B_1 \cap B_2 = \phi$$

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2)$$

- Bayes Theorem

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{2} P(A|B_i)P(B_i)}$$

# Bayes Theorem: Example

- **Screening Test for a Disease**

Suppose the frequency of the disease in the population (base rate) is 0.5%. The screening test is highly accurate with a 5% false positive rate and a 10% false negative rate.

You take the test and it comes back positive. What is the probability that you have the disease?

From MIT 18.05

# Bayes Theorem: Example

- **Screening Test for a Disease**

Suppose the frequency of the disease in the population (base rate) is 0.5%. The screening test is highly accurate with a 5% false positive rate and a 10% false negative rate.

You take the test and it comes back positive. What is the probability that you have the disease?

Let's rephrase by using language of probability theory.

# False Positive

- **Screening Test for a Disease**

Suppose the frequency of the disease in the population (base rate) is 0.5%. The screening test is highly accurate with a 5% false positive rate and a 10% false negative rate.

You take the test and it comes back positive. What is the probability that you have the disease?

D+ = You have the disease.
D- = You do not.
T+ = You tested positive.
T- = You tested negative.

False positive = test is positive when a person does not have the disease.

**The fact that test is positive** means that the test result indicates the tested person has the disease. But it is likely that the person does not have the disease in reality, even though the person is tested positive.

From MIT 18.05

# False Negative

- **Screening Test for a Disease**

Suppose the frequency of the disease in the population (base rate) is 0.5%. The screening test is highly accurate with a 5% false positive rate and a 10% false negative rate.

You take the test and it comes back positive. What is the probability that you have the disease?

D+ = You have the disease.
D- = You do not.
T+ = You tested positive.
T- = You tested negative.

False negative = test is negative when a person has the disease.

**The fact that test is negative** means that the test result indicates the tested person does not have the disease. But it is likely that the person has the disease in reality, even though the person is tested negative.

From MIT 18.05

# Probabilities

- **Screening Test for a Disease**

Suppose the frequency of the disease in the population (base rate) is **(1)0.5%**. The screening test is highly accurate with **(2)a 5% false positive rate** and **(3)a 10% false negative rate**.

You take the test and it comes back positive. What is the probability that you have the disease?

**(1)** $P(D+) = 0.005 \quad P(D-) = 0.995$

**(2)** False positive = test is positive ***when*** you do not has the disease

(Conditional Probability)

$$P(T+|D-) = 0.05$$

**(3)** False negative = test is negative ***when*** you do has the disease

(Conditional Probability)

$$P(T-|D+) = 0.10$$

# True Positive/Negative

- **Screening Test for a Disease**

Suppose the frequency of the disease in the population (base rate) is **(1)0.5%**. The screening test is highly accurate with **(2)a 5% false positive rate** and **(3)a 10% false negative rate**.

You take the test and it comes back positive. What is the probability that you have the disease?

$$P(D+) = 0.005 \quad P(D-) = 0.995$$

$$P(T+|D-) = 0.05$$

$$\longrightarrow P(T-|D-) = 1 - P(T+|D-) = 0.95$$

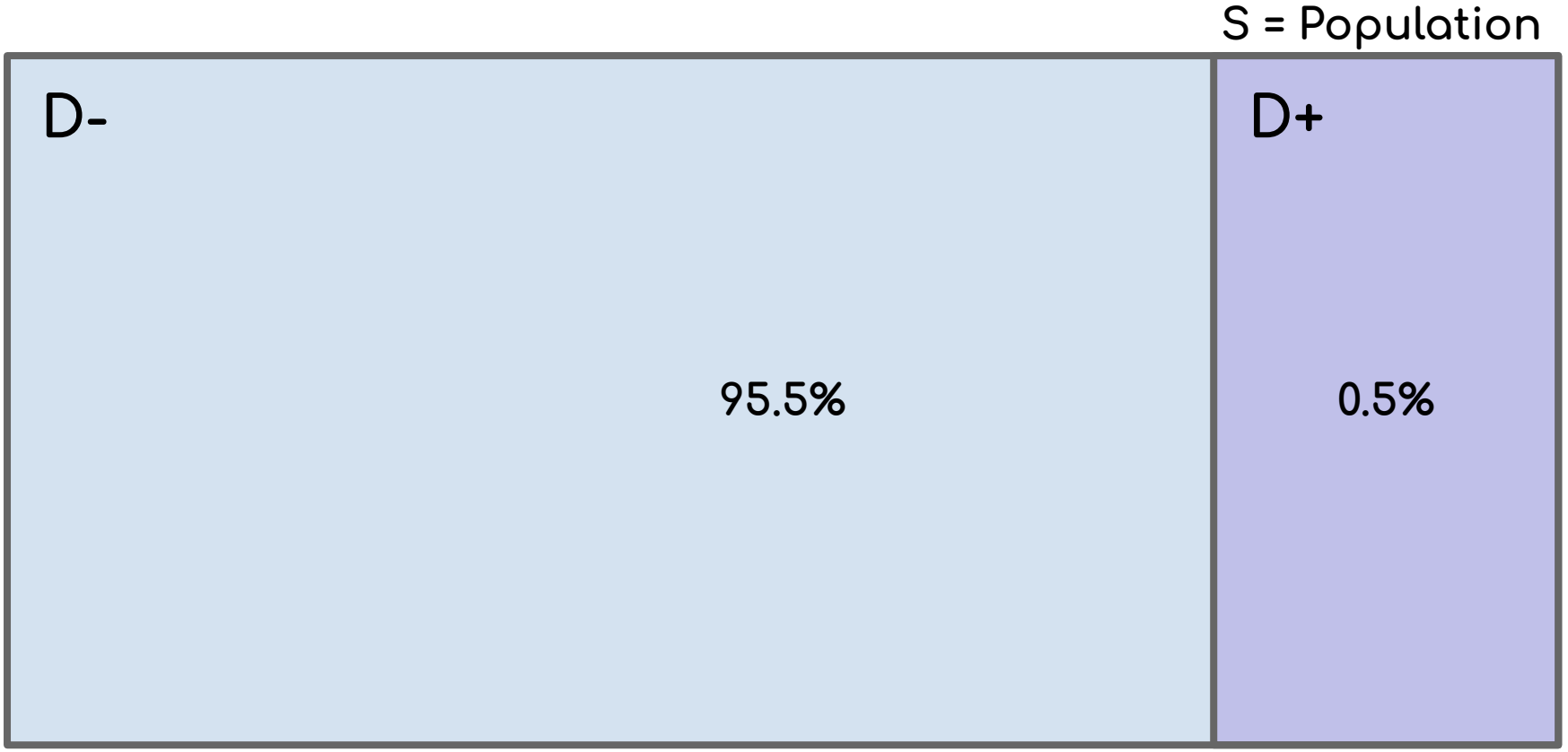
**True negative rate**

$$P(T-|D+) = 0.10$$

$$\longrightarrow P(T+|D+) = 1 - P(T-|D+) = 0.90$$

**True positive rate**

# Bayes Theorem: Example



# Bayes Theorem: Example

# Bayes Theorem: Example

# Bayes Theorem: Example

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

- You take the test and it comes back positive. **[Condition!]**

$$\text{You} \in T+$$

- What is the **probability that you have the disease**?

$$\text{You} \in D+$$

$$\longrightarrow P(D+|T+)?$$

# Bayes Theorem: Example

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

- We want to calculate $P(D+|T+)$?
- We now know $P(D+),\ P(T+|D-),\ P(T-|D+)$
- We can calculate $P(T+|D+) = 1 - P(T-|D+)$

→ We can use Bayes theorem to calculate **P(D+|T+)** with A = T+ and B = D+

# Bayes Theorem: Example

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

- **What we need to know additionally: P(T+)**

***Law of Total Probability!***

$$P(T+) = P(T+|D+)P(D+) + P(T+|D-)P(D-)$$
$$= 0.9 \times 0.005 + 0.995 \times 0.05 = 0.05425$$

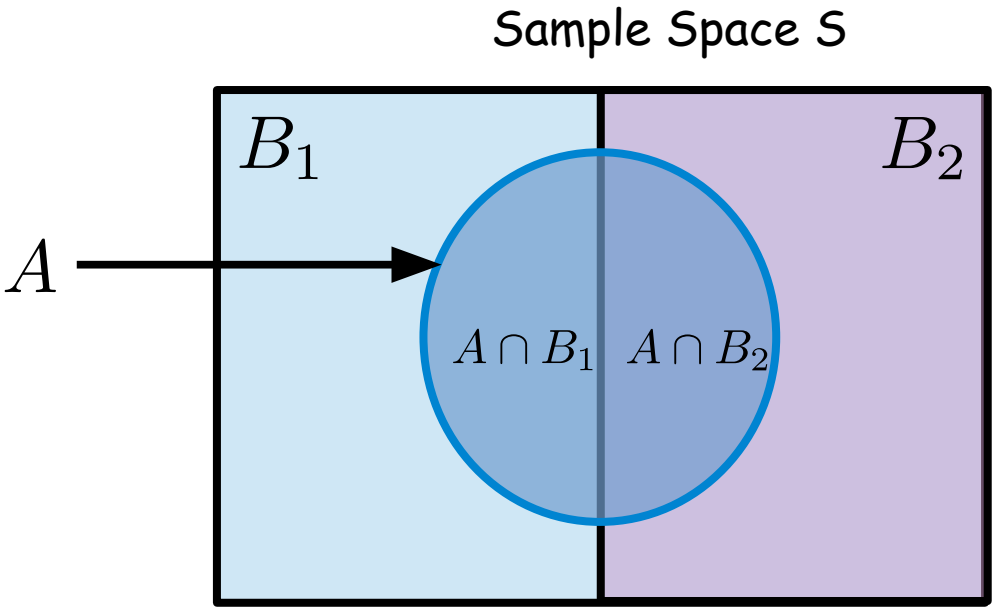
# Recall: Law of Total Probability

- **For mutually exclusive (or disjoint) events**

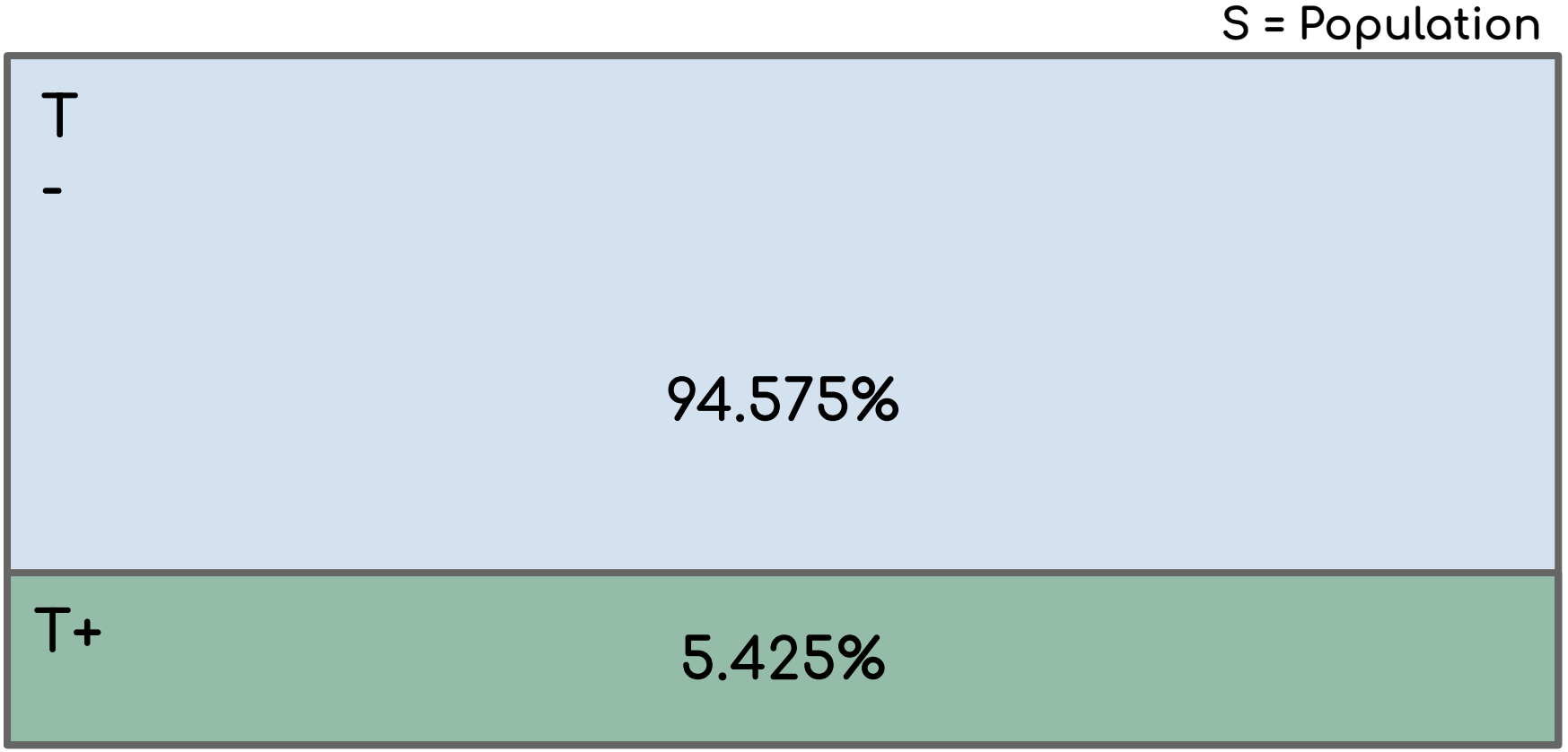
$$B_1 \cup B_2 = S \quad B_1 \cap B_2 = \phi$$

$$\rightarrow P(A) = P(A \cap B_1) + P(A \cap B_2)$$

$$= P(A|B_1)P(B_1) + P(A|B_2)P(B_2)$$



# Bayes Theorem: Example



# Bayes Theorem: Example

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+)}$$

- The probability that you really do have the disease when you are tested positive.

$$P(D+|T+) = \frac{0.9 \times 0.005}{0.05425} = 0.082949 \approx 8.3\%$$

# Bayes Theorem: Example

$$P(D+|T+) = \frac{P(T+|D+)P(D+)}{P(T+)}$$

- The probability that you really do have the disease when you are tested positive.

$$P(D+|T+) = \frac{0.9 \times 0.005}{0.05425} = 0.082949 \approx 8.3\%$$

- The probability that you do not have the disease even though you are tested positive.

$$P(D-|T+) \approx 91.7\%$$

# This Test is Accurate?

- The probability that you do not have the disease even though you are tested positive.

$$P(D-|T+) \approx 91.7\%$$

**<u>Can you say that this screening test is accurate?</u>**

- Let us say that you definitely know that you do not have the disease. In this case, the screening test has a 95% accuracy.

$$P(T-|D-) = 0.95$$

- But, when you know only that the test result is positive, the probability that you really have the disease is 8.3%, which is too low.

# Base Rate Fallacy

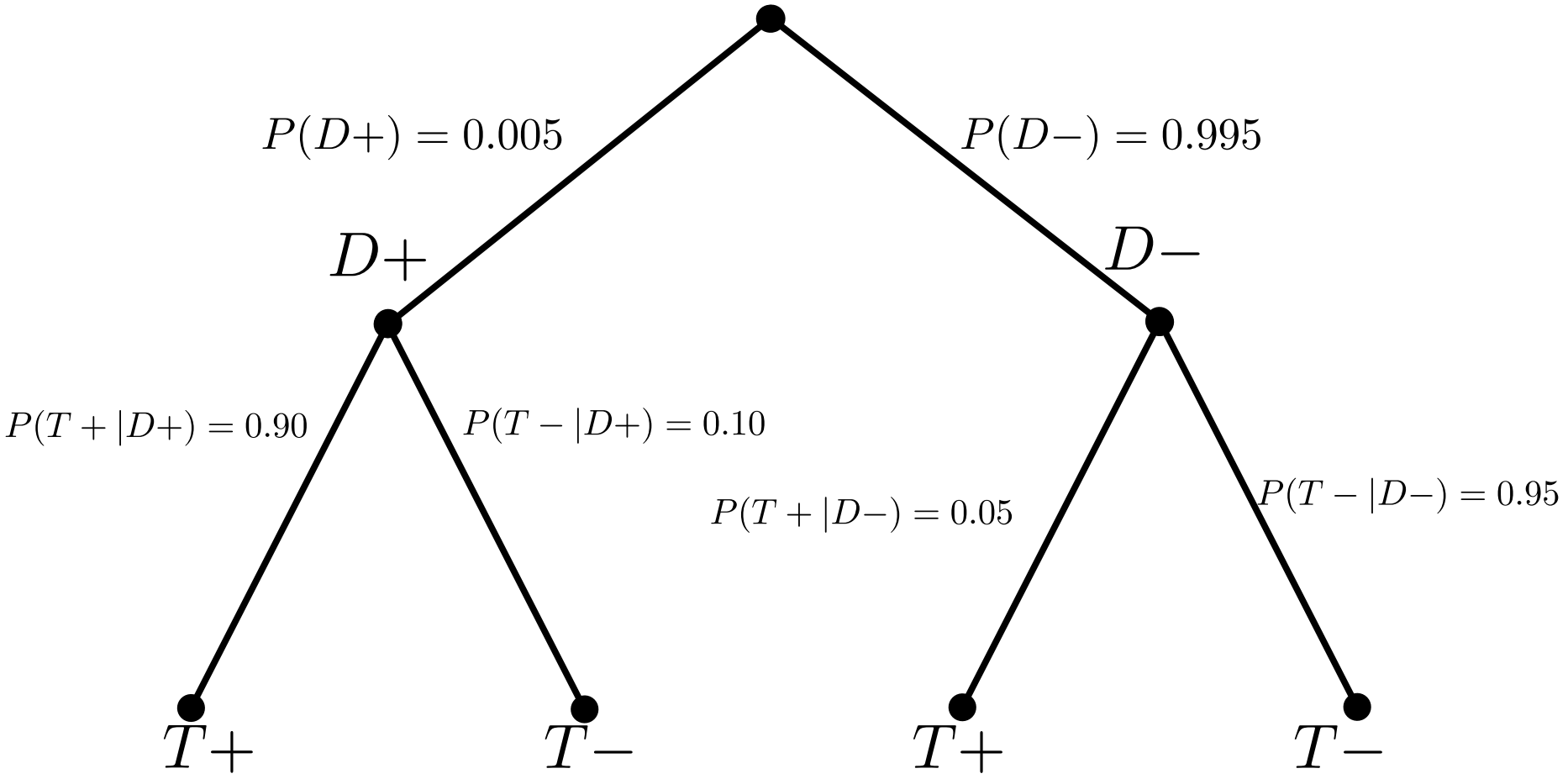
"95% of all tests are accurate does not imply 95% of positive tests are accurate."

- Base rate fallacy

  It is easy to confuse the meaning of P(A|B) and P(B|A).

https://en.wikipedia.org/wiki/Base_rate_fallacy

# Bayes Theorem: Using Trees



# Bayes Theorem: Using Table

10,000 people

| | $D+$ | $D-$ | total |
|---|---|---|---|
| $T+$ | | | |
| $T-$ | | | |
| total | | | $10,000$ |

# Bayes Theorem: Using Table

10,000 people

| | $D+$ | $D-$ | total |
|---|---|---|---|
| $T+$ | | | |
| $T-$ | | | |
| total | 50 | $9,950$ | $10,000$ |

$P(D+) = 0.005 \quad P(D-) = 0.995$

# Bayes Theorem: Using Table

10,000 people

| | $D+$ | $D-$ | total |
|---|---|---|---|
| $T+$ | 45 | | |
| $T-$ | 5 | | |
| total | 50 | $9,950$ | $10,000$ |

$$P(T+|D+) = 0.90$$
$$P(T-|D+) = 0.10$$

# Bayes Theorem: Using Table

10,000 people

| | $D+$ | $D-$ | total |
|---|---|---|---|
| $T+$ | 45 | 498 | 543 |
| $T-$ | 5 | 9452 | 9457 |
| total | 50 | $9,950$ | $10,000$ |

$$P(T+|D-) = 0.05$$
$$P(T-|D-) = 0.95$$

# Bayes Theorem: Using Table

10,000 people

| | $D+$ | $D-$ | total |
|---|---|---|---|
| $T+$ | 45 | 498 | 543 |
| $T-$ | 5 | 9452 | 9457 |
| total | 50 | $9,950$ | $10,000$ |

When the test is positive, what is the probability that the tested person really has the disease?

# Bayes Theorem: Using Table

10,000 people

| | $D+$ | $D-$ | total |
|---|---|---|---|
| $T+$ | 45 | 498 | 543 |
| $T-$ | 5 | 9452 | 9457 |
| total | 50 | $9,950$ | $10,000$ |

**When the test is positive**, what is the probability that the tested person really has the disease?

$$45/543 \approx 0.083$$

# Bayes Theorem: Using Table

| | $D+$ | $D-$ | total |
|---|---|---|---|
| $T+$ | 45 | 498 | 543 |
| $T-$ | 5 | 9452 | 9457 |
| total | 50 | $9,950$ | $10,000$ |

$P(D+|T+) = 0.083$

$P(T+|D+) = 0.95$ ⟶

***What makes this difference?***

# Bayes Theorem: Using Table

| | $D+$ | $D-$ | total |
|---|---|---|---|
| $T+$ | 45 | 498 | 543 |
| $T-$ | 5 | 9452 | 9457 |
| total | 50 | $9,950$ | $10,000$ |

$P(D+|T+) = 0.083$

$P(T+|D+) = 0.95$

$N(D+|T+) = 45$ vs. $N(D-|T+) = 498$

That is because the base rate is too small, i.e., this disease is too rare.

# Bayes Theorem: Using Table

What if the base rate changes?

$$P(D+) = 0.05, \quad P(D-) = 0.95$$

| | $D+$ | $D-$ | total |
|---|---|---|---|
| $T+$ | | | |
| $T-$ | | | |
| total | | | $10,000$ |

# Bayes Theorem: Using Table

What if the base rate changes?

$$P(D+) = 0.05, \quad P(D-) = 0.95$$

| | $D+$ | $D-$ | total |
|---|---|---|---|
| $T+$ | 450 | 475 | 925 |
| $T-$ | 50 | $9,025$ | $9,075$ |
| total | 500 | $9,500$ | $10,000$ |

$$P(D+|T+) = 0.486$$