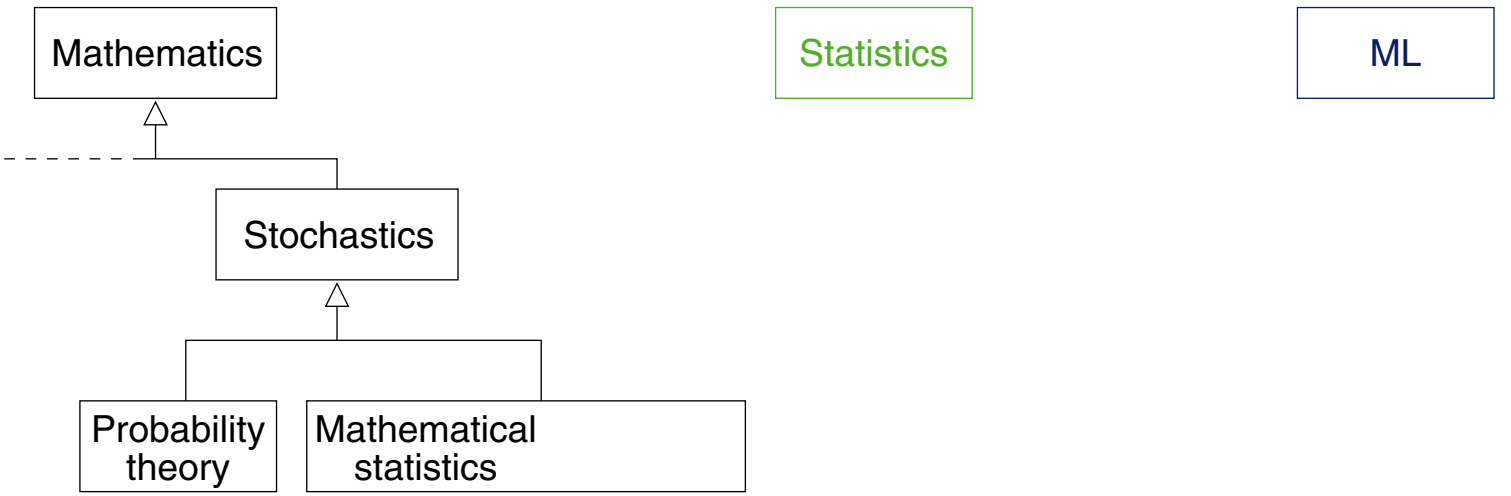
# Chapter ML:VII

## VII. Bayesian Learning

- Approaches to Probability
- Conditional Probability
- Bayes Classifier
- Exploitation of Data
- Frequentist versus Subjectivist

# Approaches to Probability

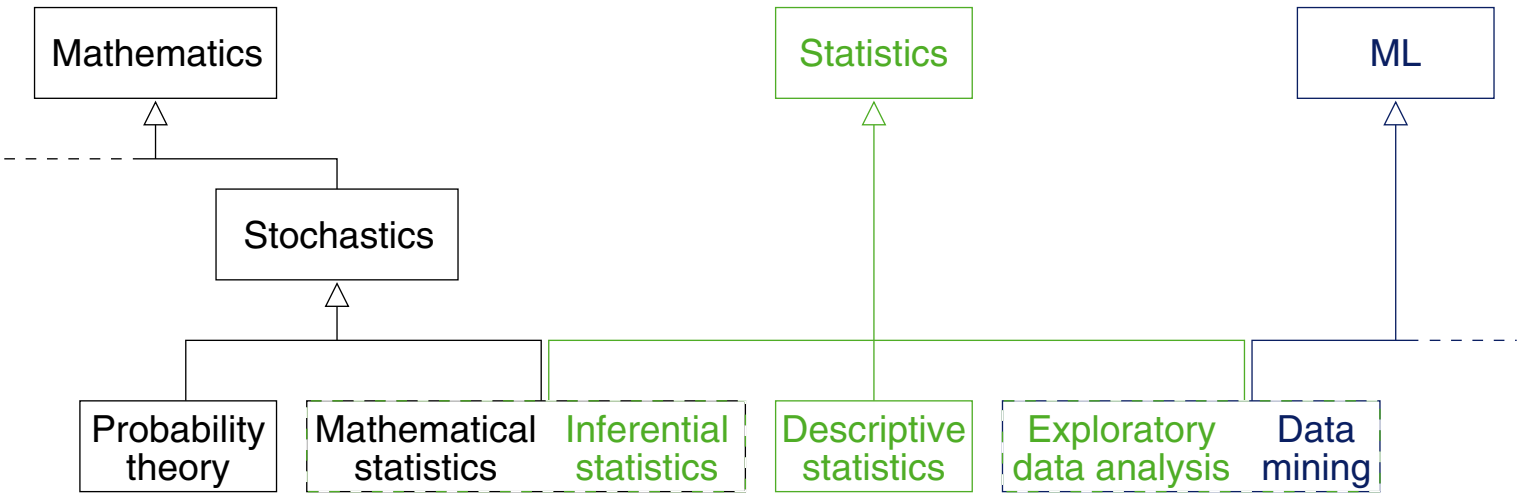
## Area Overview




- Probability theory: probability measures, Kolmogorov axioms
- Mathematical statistics: application of probability theory, Naive Bayes

# Approaches to Probability

## Area Overview




- Probability theory: probability measures, Kolmogorov axioms
- Mathematical statistics: application of probability theory, Naive Bayes
- Inferential statistics: parameter estimation, hypothesis (parameter) tests, confidence intervals
- Descriptive statistics: variances, contingencies
- Exploratory data analysis: histograms, principal component analysis
- Data mining: anomaly detection, cluster analysis

# Approaches to Probability

**Definition 1 (Random Experiment, Random Observation)**

A random experiment or random trial is a procedure that, at least theoretically, can be repeated infinite times. It is characterized as follows:

1. Configuration.
   A precisely specified system that can be reconstructed.
2. Procedure.
   An instruction of how to execute the experiment, based on the configuration.
3. Unpredictability of the outcome.

# Approaches to Probability

**Definition 1 (Random Experiment, Random Observation)**

A random experiment or random trial is a procedure that, at least theoretically, can be repeated infinite times. It is characterized as follows:

1. Configuration.
   A precisely specified system that can be reconstructed.
2. Procedure.
   An instruction of how to execute the experiment, based on the configuration.
3. Unpredictability of the outcome.

Random experiments whose configuration and procedure are not designed artificially are called *natural random experiments* or *natural random observations*.

Remarks:

- A procedure can be repeated several times using the same system, but also with different "copies" of the original system.

  In particular, a random experiment is called *ergodic* if its time average (= sequential analysis) is the same as its ensemble average (= parallel analysis). [Wikipedia]

- Note that random experiments are causal in the sense of cause and effect. The randomness of an experiment, i.e., the unpredictability of its outcome, is a consequence of the missing information about the causal chain. Hence a random experiment can turn into a deterministic process when new insights become known.

# Approaches to Probability

**Definition 2 (Sample Space, Event Space)**

A set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ is called sample space of a random experiment, if each experiment outcome is associated with at most one element $\omega \in \Omega$. The elements in $\Omega$ are called outcomes.

Let $\Omega$ be a finite sample space. Each subset $A \subseteq \Omega$ is called an event; an event $A$ occurs iff the experiment outcome $\omega$ is a member of $A$. The set of all events, $\mathcal{P}(\Omega)$, is called the event space or $\sigma$-algebra.

# Approaches to Probability

**Definition 2 (Sample Space, Event Space)**

A set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ is called sample space of a random experiment, if each experiment outcome is associated with at most one element $\omega \in \Omega$. The elements in $\Omega$ are called outcomes.

Let $\Omega$ be a finite sample space. Each subset $A \subseteq \Omega$ is called an event; an event $A$ occurs iff the experiment outcome $\omega$ is a member of $A$. The set of all events, $\mathcal{P}(\Omega)$, is called the event space or $\sigma$-algebra.

Examples:

| | |
|---|---|
| Experiment: | Rolling a dice. |
| Sample space: | $\Omega = \{1, 2, 3, 4, 5, 6\}$ |
| Some event: | $A = \{2, 4, 6\}$ |

# Approaches to Probability

**Definition 2 (Sample Space, Event Space)**

A set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ is called sample space of a random experiment, if each experiment outcome is associated with at most one element $\omega \in \Omega$. The elements in $\Omega$ are called outcomes.

Let $\Omega$ be a finite sample space. Each subset $A \subseteq \Omega$ is called an event; an event $A$ occurs iff the experiment outcome $\omega$ is a member of $A$. The set of all events, $\mathcal{P}(\Omega)$, is called the event space or $\sigma$-algebra.

Examples:

| | | |
|---|---|---|
| Experiment: | Rolling a dice. | Rolling two dice at the same time. |
| Sample space: | $\Omega = \{1, 2, 3, 4, 5, 6\}$ | $\Omega = \{\{1,1\}, \{1,2\}, \ldots, \{2,2\}, \ldots, \{6,6\}\}$ |
| Some event: | $A = \{2, 4, 6\}$ | $B = \{\{1, 2\}\}$ |

# Approaches to Probability

**Definition 2 (Sample Space, Event Space)**

A set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ is called sample space of a random experiment, if each experiment outcome is associated with at most one element $\omega \in \Omega$. The elements in $\Omega$ are called outcomes.

Let $\Omega$ be a finite sample space. Each subset $A \subseteq \Omega$ is called an event; an event $A$ occurs iff the experiment outcome $\omega$ is a member of $A$. The set of all events, $\mathcal{P}(\Omega)$, is called the event space or $\sigma$-algebra.

Examples:

Experiment: Rolling a dice.

Sample space: $\Omega = \{1, 2, 3, 4, 5, 6\}$

Some event: $A = \{2, 4, 6\}$

Rolling two dice at the same time.

$\Omega = \{\{1, 1\}, \{1, 2\}, \ldots, \{2, 2\}, \ldots, \{6, 6\}\}$

$B = \{\{1, 2\}\}$

Rolling two dice in succession.

$\Omega = \{(1, 1), (1, 2), \ldots, (2, 1), \ldots, (6, 6)\}$

$B = \{(1, 2), (2, 1)\}$

# Approaches to Probability

## Definition 3 (Important Event Types)

Let $\Omega$ be a finite sample space, and let $A \subseteq \Omega$ and $B \subseteq \Omega$ be two events. Then we agree on the following notation:

1. $\emptyset$ — The impossible event.
2. $\Omega$ — The certain event.
3. $\overline{A} := \Omega \setminus A$ — The complementary event of $A$.
4. $|A| = 1$ — An elementary event.

5. $A \subseteq B \quad \Leftrightarrow$ $A$ is a subevent of $B$, "$A$ entails $B$", $A \Rightarrow B$
6. $A = B \quad \Leftrightarrow$ $A \subseteq B$ and $B \subseteq A$
7. $A \cap B = \emptyset \quad \Leftrightarrow$ $A$ and $B$ are incompatible (otherwise, they are compatible).

# Approaches to Probability

## Definition 3 (Important Event Types)

Let $\Omega$ be a finite sample space, and let $A \subseteq \Omega$ and $B \subseteq \Omega$ be two events. Then we agree on the following notation:

1. $\emptyset$ — The impossible event.
2. $\Omega$ — The certain event.
3. $\overline{A} := \Omega \setminus A$ — The complementary event of $A$.
4. $|A| = 1$ — An elementary event.
5. $A \subseteq B$ $\Leftrightarrow$ $A$ is a subevent of $B$, "$A$ entails $B$", $A \Rightarrow B$
6. $A = B$ $\Leftrightarrow$ $A \subseteq B$ and $B \subseteq A$
7. $A \cap B = \emptyset$ $\Leftrightarrow$ $A$ and $B$ are incompatible (otherwise, they are compatible).

Example (Point 5) :

| | | |
|---|---|---|
| $\{2\}$ | $\subset$ | $\{2, 4, 6\}$ |
| "Roll a two." | $\succ$ | "Even number roll." |
| "2" | entails | "Even number roll." |
| "2" | $\Rightarrow$ | "2 or 4 or 6" |

Remarks:

- Alternative and semantically equivalent notations for the probability of the joint event "$A$ and $B$" :
  1. $P(A, B)$
  2. $P(A \wedge B)$
  3. $P(A \cap B)$

# Approaches to Probability

## How to Capture the Nature of Probability

1. Classic, symmetry-based
2. Frequentist
3. Axiomatic
4. Subjectivist, Bayesian, prognostic

# Approaches to Probability

## How to Capture the Nature of Probability (continued)

1. Classic, symmetry-based
2. Frequentist
3. Axiomatic
4. Subjectivist, Bayesian, prognostic

**Definition 4 (Classical / Laplace Probability [1749-1827])**

If each elementary event $\{\omega\}$, $\omega \in \Omega$, gets assigned the same probability (equiprobable events), then the probability $P(A)$ of an event $A$ is defined as follows:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{number of cases favorable for } A}{\text{number of total outcomes possible}}$$

Remarks:

- A random experiment whose configuration and procedure imply an equiprobable sample space, be it by definition or by construction, is called Laplace experiment. The probabilities of the outcomes are called Laplace probabilities.

  Since Laplace probabilities are defined by the experiment configuration along with the experiment procedure, they need not to be estimated.

- The assumption that a given experiment is a Laplace experiment is called Laplace assumption. If the Laplace assumption cannot be presumed, the probabilities can only be obtained from a (possibly large) number of trials.

- Strictly speaking, the Laplace probability as introduced above is not a definition but a circular definition: the probability concept is defined by means of the concept of equiprobability, i.e., another kind of probability.

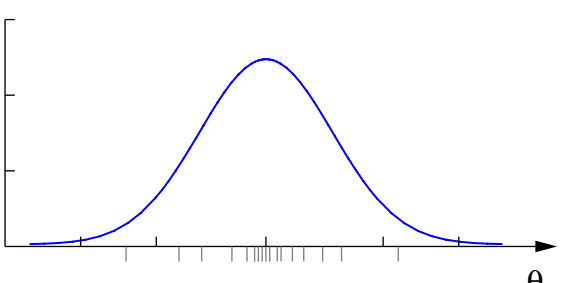
# Approaches to Probability

## How to Capture the Nature of Probability (continued)

1. Classic, symmetry-based
2. Frequentist
3. Axiomatic
4. Subjectivist, Bayesian, prognostic

Basis is the empirical law of large numbers:

In a random experiment, the average of the outcomes obtained from a large number of trials is close to the expected value, and it will become closer as more trials are performed.



Remarks:

- Basic assumption of the frequentist approach is that—given enough data—the data will explain the hypothesis (and no additional assumptions or prior knowledge are necessary).
- Inspired by the empirical law of large numbers, scientists have tried to develop a frequentist probability concept, which is completely based on the (fictitious) limit of the relative frequencies. [von Mises, 1951]

  These attempts failed since such a limit formation is possible only within mathematical settings (infinitesimal calculus), where accurate repetitions unto infinity can be made.

# Approaches to Probability

## How to Capture the Nature of Probability (continued)

1. Classic, symmetry-based
2. Frequentist
3. Axiomatic
4. Subjectivist, Bayesian, prognostic

Axiomatic approach to phenomena modeling:

(a) Postulate a function $P()$ that assigns a "probability" to each event in $\mathcal{P}(\Omega)$.

(b) Specify the required properties of $P()$ in the form of axioms.

# Approaches to Probability

## How to Capture the Nature of Probability (continued)

1. Classic, symmetry-based
2. Frequentist
3. Axiomatic
4. Subjectivist, Bayesian, prognostic

Axiomatic approach to phenomena modeling:

(a) Postulate a function $P()$ that assigns a "probability" to each event in $\mathcal{P}(\Omega)$.

(b) Specify the required properties of $P()$ in the form of axioms.

# Approaches to Probability

## How to Capture the Nature of Probability (continued)

1. Classic, symmetry-based
2. Frequentist
3. Axiomatic
4. Subjectivist, Bayesian, prognostic

Consider (prior) knowledge about the hypotheses:

$$p(h \mid D) = \frac{p(D \mid h) \cdot p(h)}{p(D)}$$

# Approaches to Probability

## How to Capture the Nature of Probability (continued)

1. Classic, symmetry-based
2. Frequentist
3. Axiomatic
4. Subjectivist, Bayesian, prognostic

Consider (prior) knowledge about the hypotheses:

$$p(h \mid D) = \frac{p(D \mid h) \cdot p(h)}{p(D)}$$

- Likelihood: How well does $h$ explain (= entail, induce, evoke) the data $D$?
- Prior: How probable is the hypothesis $h$ a priori (= in principle)?

# Approaches to Probability

## How to Capture the Nature of Probability (continued)

1. Classic, symmetry-based
2. Frequentist
3. Axiomatic
4. Subjectivist, Bayesian, prognostic

Consider (prior) knowledge about the hypotheses:

$$p(h \mid D) = \frac{p(D \mid h) \cdot p(h)}{p(D)}$$

- Likelihood: How well does $h$ explain (= entail, induce, evoke) the data $D$?
- Prior: How probable is the hypothesis $h$ a priori (= in principle)?

$$p(h \mid D) \propto p(D \mid h) \cdot p(h)$$

Remarks:

- Likelihood is the hypothetical probability that an event that has already occurred (here: an experiment parameterized by $h$) would yield a specific outcome (here: $D$, which typically is a sequence of feature-value pairs $(\mathbf{x}, c)$).

  The concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes. I.e., $p(D \mid h)$ is called likelihood since we reason about a past experiment. [Mathworld]

- Probability pattern versus likelihood pattern. Let $B$ be known.
  - With probabilities $P(\cdot \mid B)$ we reason about the future, i.e., possible "consequence" events "caused" by $B$.
  - With likelihoods $P(B \mid \cdot)$ we reason about the past, i.e., possible and already occurred "precursor" or "condition" events for $B$. More specifically, $B$ may denote a "data event", **D**$=D$, and we reason about the occurred "parameter event", *H*$=h$.

- The symbol »$\propto$« means "is proportional to".

Remarks (frequentist versus subjectivist) :

- If applicable and if properly applied the frequentist approach and the Bayesian approach will lead to the same result in most cases.
- The frequentist approach cannot handle singleton or rare events. Example:

  “What are the chances that the first human mission to Mars will become a success?”
- “It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel. Doubtless, much of the disagreement is merely terminological and would disappear under sufficiently sharp analysis.” [Savage, 1954]

# Approaches to Probability

## Axiomatic Approach to Probability

**Definition 5 (Probability Measure** [Kolmogorov 1933]**)**

Let $\Omega$ be a set, called sample space, and let $\mathcal{P}(\Omega)$ be the set of all events, called event space. A function $P$, $P : \mathcal{P}(\Omega) \to \mathbf{R}$, which maps each event $A \in \mathcal{P}(\Omega)$ onto a real number $P(A)$, is called probability measure if it has the following properties:

1. $P(A) \geq 0$ (Axiom I)
2. $P(\Omega) = 1$ (Axiom II)
3. $A \cap B = \emptyset$ implies $P(A \cup B) = P(A) + P(B)$ (Axiom III)

# Approaches to Probability

## Axiomatic Approach to Probability

**Definition 5 (Probability Measure** [Kolmogorov 1933]**)**

Let $\Omega$ be a set, called sample space, and let $\mathcal{P}(\Omega)$ be the set of all events, called event space. A function $P$, $P : \mathcal{P}(\Omega) \to \mathbf{R}$, which maps each event $A \in \mathcal{P}(\Omega)$ onto a real number $P(A)$, is called probability measure if it has the following properties:

1. $P(A) \geq 0$ (Axiom I)
2. $P(\Omega) = 1$ (Axiom II)
3. $A \cap B = \emptyset \quad \rightarrow \quad P(A \cup B) = P(A) + P(B)$ (Axiom III)

# Approaches to Probability

## Axiomatic Approach to Probability (continued)

**Definition 5 (Probability Measure** [Kolmogorov 1933]**)**

Let $\Omega$ be a set, called sample space, and let $\mathcal{P}(\Omega)$ be the set of all events, called event space. A function $P$, $P : \mathcal{P}(\Omega) \to \mathbf{R}$, which maps each event $A \in \mathcal{P}(\Omega)$ onto a real number $P(A)$, is called probability measure if it has the following properties:

1. $P(A) \geq 0$ (Axiom I)
2. $P(\Omega) = 1$ (Axiom II)
3. $A \cap B = \emptyset \quad \rightarrow \quad P(A \cup B) = P(A) + P(B)$ (Axiom III)

**Definition 6 (Probability Space)**

Let $\Omega$ be a sample space, let $\mathcal{P}(\Omega)$ be an event space, and let $P : \mathcal{P}(\Omega) \to \mathbf{R}$ be a probability measure. Then the tuple $(\Omega, P)$, as well as the triple $(\Omega, \mathcal{P}(\Omega), P)$, is called probability space.

# Approaches to Probability

## Axiomatic Approach to Probability (continued)

**Definition 5 (Probability Measure** [Kolmogorov 1933]**)**

Let $\Omega$ be a set, called sample space, and let $\mathcal{P}(\Omega)$ be the set of all events, called event space. A function $P$, $P : \mathcal{P}(\Omega) \to \mathbf{R}$, which maps each event $A \in \mathcal{P}(\Omega)$ onto a real number $P(A)$, is called probability measure if it has the following properties:

1. $P(A) \geq 0$ (Axiom I)
2. $P(\Omega) = 1$ (Axiom II)
3. $A \cap B = \emptyset \quad \rightarrow \quad P(A \cup B) = P(A) + P(B)$ (Axiom III)

**Definition 6 (Probability Space)**

Let $\Omega$ be a sample space, let $\mathcal{P}(\Omega)$ be an event space, and let $P : \mathcal{P}(\Omega) \to \mathbf{R}$ be a probability measure. Then the tuple $(\Omega, P)$, as well as the triple $(\Omega, \mathcal{P}(\Omega), P)$, is called probability space.

We can work with probabilities without interpreting them.

# Approaches to Probability

## Axiomatic Approach to Probability (continued)

**Theorem 7 (Implications of Kolmogorov Axioms)**

1. $P(A) + P(\overline{A}) = 1$ (from Axioms II, III)

2. $P(\emptyset) = 0$ (from 1. with $A = \Omega$)

3. Monotonicity law of the probability measure:
   $A \subseteq B \Rightarrow P(A) \leq P(B)$ (from Axioms I, II)

4. "Sum rule" or "addition rule" :
   $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (from Axiom III)

5. Let $A_1, A_2, \ldots, A_k$ be mutually exclusive (incompatible), then holds:
   $P(A_1 \cup A_2 \cup \ldots \cup A_k) = P(A_1) + P(A_2) + \ldots + P(A_k)$

Remarks:

- The three axioms are also called the Axiom System of Kolmogorov.
- $P(A)$ is called “probability of the occurrence of $A$.”
- Observe that nothing is said about how to interpret the probabilities $P()$. An axiomatic approach does not explain but specifies properties “only”.
- Also observe that nothing is said about the distribution of the probabilities $P()$.
- A function that provides the three properties of a probability measure is called a non-negative, normalized, and additive measure.

# Chapter ML:VII (continued)

## VII. Bayesian Learning

- Approaches to Probability
- Conditional Probability
- Bayes Classifier
- Exploitation of Data
- Frequentist versus Subjectivist

# Conditional Probability

## Basic Definition

**Definition 8 (Conditional Probability)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then the probability of the occurrence of event $A$ given that event $B$ is known to have occurred is defined as follows:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0$$

$P(A \mid B)$ is called "probability of $A$ under condition $B$."

# Conditional Probability

## Basic Definition (continued)

**Definition 8 (Conditional Probability)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then the probability of the occurrence of event $A$ given that event $B$ is known to have occurred is defined as follows:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0$$

$P(A \mid B)$ is called "probability of $A$ under condition $B$."

# Conditional Probability

## Basic Definition (continued)

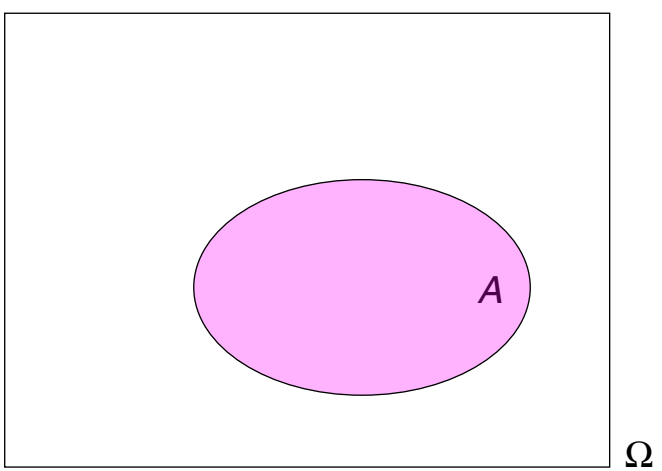
**Definition 8 (Conditional Probability)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then the probability of the occurrence of event $A$ given that event $B$ is known to have occurred is defined as follows:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0$$

$P(A \mid B)$ is called "probability of $A$ under condition $B$."

$A$ : The road is wet.




$A \equiv A \mid \Omega$

# Conditional Probability

## Basic Definition (continued)

**Definition 8 (Conditional Probability)**

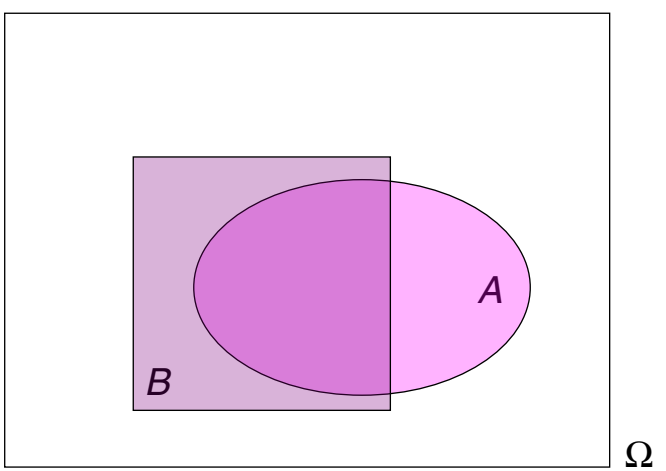
Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then the probability of the occurrence of event $A$ given that event $B$ is known to have occurred is defined as follows:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0$$

$P(A \mid B)$ is called "probability of $A$ under condition $B$."

$A$ : The road is wet.

$B$ : It's raining.




$B \equiv B \mid \Omega$

# Conditional Probability

## Basic Definition (continued)

**Definition 8 (Conditional Probability)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then the probability of the occurrence of event $A$ given that event $B$ is known to have occurred is defined as follows:

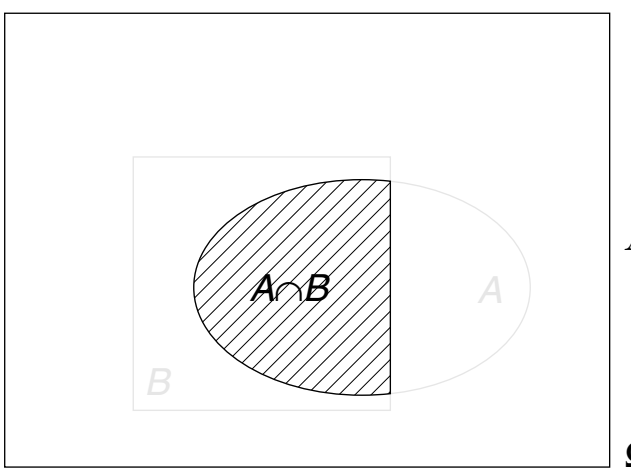
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0$$

$P(A \mid B)$ is called "probability of $A$ under condition $B$."

$A$ : The road is wet.

$B$ : It's raining.

$A \cap B$ : The road is wet and it's raining.




$A \cap B \equiv A \cap B \mid \Omega$

# Conditional Probability

## Basic Definition (continued)

**Definition 8 (Conditional Probability)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then the probability of the occurrence of event $A$ given that event $B$ is known to have occurred is defined as follows:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0$$

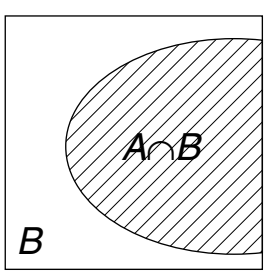
$P(A \mid B)$ is called "probability of $A$ under condition $B$."

$A$ : The road is wet.

$B$ : It's raining.

$A \cap B$ : The road is wet and it's raining.

$A \mid B$ : The road is wet when it's raining.




$$A \mid B \equiv A \cap B \mid B$$

Here:

$$P(A \mid B) > P(A)$$

Remarks (conditional probability) :

- Considered as a function in a parameter (event), $A$, and a constant (event), $B$, the conditional probability $P(A \mid B)$ fulfills the Kolmogorov Axioms and in turn defines a probability measure. This probability measure is denoted as $P_B()$.

- Important consequences (deductions) from the conditional probability definition:

  1. $P(A \cap B) = P(B) \cdot P(A \mid B)$ (see multiplication rule given statistical independence)
  2. $P(A \cap B) = P(B \cap A) = P(A) \cdot P(B \mid A)$
  3. $P(B) \cdot P(A \mid B) = P(A) \cdot P(B \mid A) \Leftrightarrow P(A \mid B) = \frac{P(A \cap B)}{P(B)} \overset{(\star)}{=} \frac{P(A) \cdot P(B \mid A)}{P(B)}$
  4. $P(\overline{A} \mid B) = 1 - P(A \mid B)$ or $P_B(\overline{A}) = 1 - P_B(A)$ (see Point 1 of Implications)

$(\star)$ The identity shows the (simple) Bayes rule.

- While Deduction 4 is obvious since $P(A \mid B) \equiv P_B()$ is a probability measure, the interpretation of complementary events in conditions may be confusing.

  In particular, the following inequality must be assumed: $P(A \mid \overline{B}) \neq 1 - P(A \mid B)$

  For illustrating purposes, consider the probability $P(A \mid B) = 0.9$ for event $A$ "The road is wet" given event $B$ "It's raining". Observe that this probability cannot give us any clue regarding the wetness of the road under the complementary event $\overline{B}$ "It's not raining".

Remarks (conditional event algebra) :

- As a probability measure, the argument of $P()$ is an event, and in this sense $A \mid B$ denotes the conditional event "$A$ when given $B$". However, except for rare cases the event $A \mid B$ is not an element in the event space $2^\Omega$ (similarly: not a subset of the sample space $\Omega$) and does not satisfy Kolmogorov's axioms:

  In standard probability theory the probability measure—precisely: its domain, a $\sigma$-algebra such as $2^\Omega$—is *not closed under conditioning* (see Lewis's triviality result).

- A conditional event cannot be combined with other events but only be formed at "top-level". E.g., we cannot interpret the "event" $B \wedge (A \mid B)$, and thus cannot compute $P(B, (A \mid B))$.

  This restriction becomes clear when recalling the definition of $P(A \mid B)$ : The conditional event $A \mid B$ is not treated as an element in the domain $2^\Omega$ of $P()$, but, $P(A \mid B)$ is defined as the quotient of $P(A \cap B)$ and $P(B)$, say, $P(A \cap B)$ is normalized wrt. $P(B)$.

  In this sense, $P(B, (A \mid B)) \equiv P((B \mid \Omega), (A \mid B))$, requires a normalization wrt. to $P(\Omega)$ and $P(B)$ at the same time, which cannot be afforded with the algebraic structure of a $\sigma$-algebra.

  Note that $A \cap B$ is an element in $2^\Omega$, as guaranteed by the $\sigma$-algebra requirement.

- The syntax $P_B(A)$ (versus $P(A \mid B)$) reminds the fact that "event conditioning" is not an operation of basic set theory but a normalization of a probability value.

- With the invention of so-called conditional event algebras, CEA, the above mentioned restriction can be overcome.

# Conditional Probability

## Total Probability

**Theorem 9 (Total Probability)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A_1, \ldots, A_k$ be mutually exclusive events with $\Omega = A_1 \cup \ldots \cup A_k$, $P(A_i) > 0$, $i = 1, \ldots, k$. Then for each $B \in \mathcal{P}(\Omega)$ holds:

$$P(B) = \sum_{i=1}^{k} P(A_i) \cdot P(B \mid A_i)$$

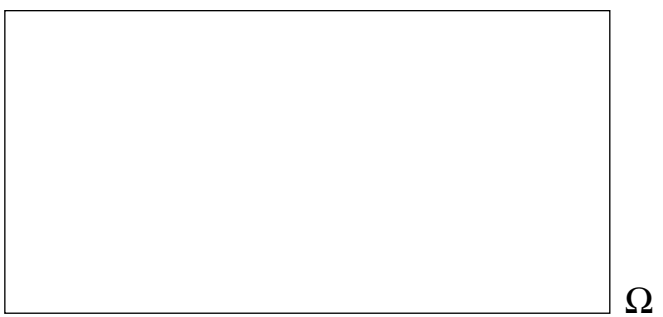
# Conditional Probability

## Total Probability (continued)

**Theorem 9 (Total Probability)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A_1, \ldots, A_k$ be mutually exclusive events with $\Omega = A_1 \cup \ldots \cup A_k$, $P(A_i) > 0$, $i = 1, \ldots, k$. Then for each $B \in \mathcal{P}(\Omega)$ holds:

$$P(B) = \sum_{i=1}^{k} P(A_i) \cdot P(B \mid A_i)$$



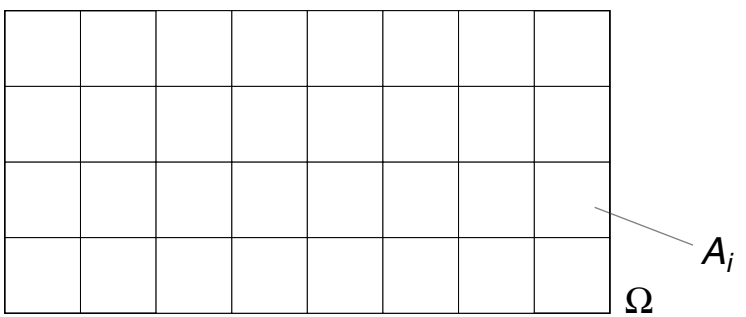
# Conditional Probability

## Total Probability (continued)

**Theorem 9 (Total Probability)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A_1, \ldots, A_k$ be mutually exclusive events with $\Omega = A_1 \cup \ldots \cup A_k$, $P(A_i) > 0$, $i = 1, \ldots, k$. Then for each $B \in \mathcal{P}(\Omega)$ holds:

$$P(B) = \sum_{i=1}^{k} P(A_i) \cdot P(B \mid A_i)$$



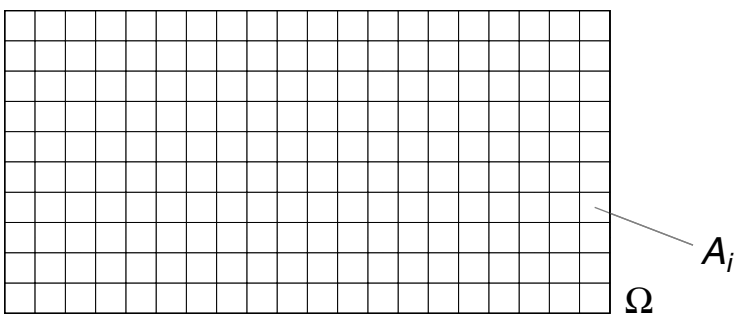
# Conditional Probability

## Total Probability (continued)

**Theorem 9 (Total Probability)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A_1, \ldots, A_k$ be mutually exclusive events with $\Omega = A_1 \cup \ldots \cup A_k$, $P(A_i) > 0$, $i = 1, \ldots, k$. Then for each $B \in \mathcal{P}(\Omega)$ holds:

$$P(B) = \sum_{i=1}^{k} P(A_i) \cdot P(B \mid A_i)$$



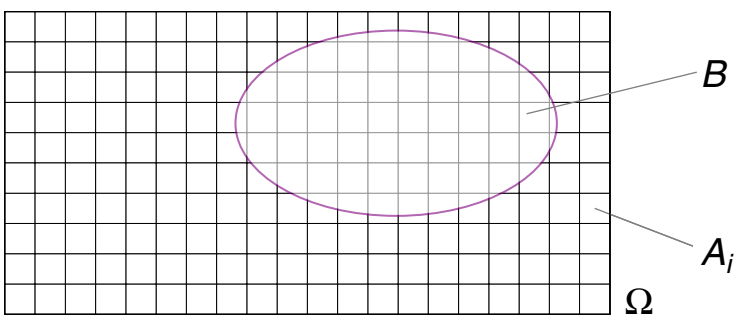
# Conditional Probability

## Total Probability (continued)

**Theorem 9 (Total Probability)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A_1, \ldots, A_k$ be mutually exclusive events with $\Omega = A_1 \cup \ldots \cup A_k$, $P(A_i) > 0$, $i = 1, \ldots, k$. Then for each $B \in \mathcal{P}(\Omega)$ holds:

$$P(B) = \sum_{i=1}^{k} P(A_i) \cdot P(B \mid A_i)$$



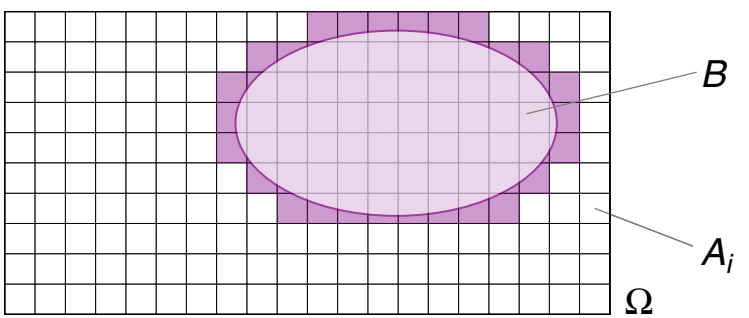
# Conditional Probability

## Total Probability (continued)

**Theorem 9 (Total Probability)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A_1, \ldots, A_k$ be mutually exclusive events with $\Omega = A_1 \cup \ldots \cup A_k$, $P(A_i) > 0$, $i = 1, \ldots, k$. Then for each $B \in \mathcal{P}(\Omega)$ holds:

$$P(B) = \sum_{i=1}^{k} P(A_i) \cdot P(B \mid A_i)$$



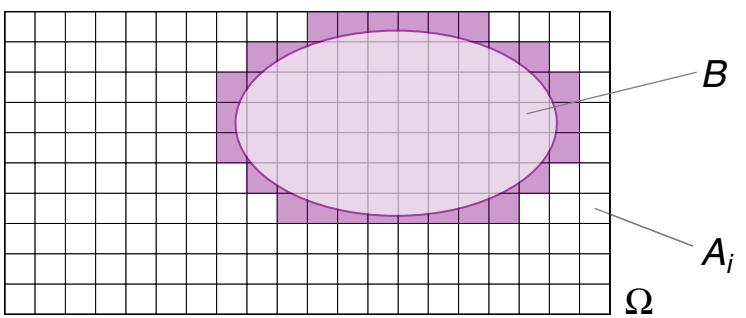
# Conditional Probability

## Total Probability (continued)

**Theorem 9 (Total Probability)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A_1, \ldots, A_k$ be mutually exclusive events with $\Omega = A_1 \cup \ldots \cup A_k$, $P(A_i) > 0$, $i = 1, \ldots, k$. Then for each $B \in \mathcal{P}(\Omega)$ holds:

$$P(B) = \sum_{i=1}^{k} P(A_i) \cdot P(B \mid A_i)$$




**Proof**

$$\begin{aligned}
P(B) &= P(\Omega \cap B) \\
&= P((A_1 \cup \ldots \cup A_k) \cap B) && \text{(exploitation of completeness of the } A_i\text{)} \\
&= P((A_1 \cap B) \cup \ldots \cup (A_k \cap B)) && \text{(exploitation of exclusiveness of the } A_i\text{)} \\
&= \sum_{i=1}^{k} P(A_i \cap B) = \sum_{i=1}^{k} P(B \cap A_i) = \sum_{i=1}^{k} P(A_i) \cdot P(B \mid A_i)
\end{aligned}$$

Remarks:

- The theorem of total probability states that the probability of an arbitrary event equals the sum of the probabilities of the subevents into which the event has been partitioned.

# Conditional Probability

## Independence of Events

**Definition 10 (Statistical Independence of two Events)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then $A$ and $B$ are called statistically independent iff the following equation holds:

$$P(A \cap B) = P(A) \cdot P(B) \qquad \text{“multiplication rule”}$$

# Conditional Probability

## Independence of Events (continued)

**Definition 10 (Statistical Independence of two Events)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then $A$ and $B$ are called statistically independent iff the following equation holds:

$$P(A \cap B) = P(A) \cdot P(B) \qquad \text{“multiplication rule”}$$

If statistical independence is given for the events $A$ and $B$, and if $0 < P(B) < 1$, then the following identities hold:

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B) \\ P(A \mid B) &= P(A \mid \overline{B}) \\ P(A \mid B) &= P(A) \end{aligned}$$

# Conditional Probability

## Independence of Events (continued)

**Definition 10 (Statistical Independence of two Events)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then $A$ and $B$ are called statistically independent iff the following equation holds:

$$P(A \cap B) = P(A) \cdot P(B) \qquad \text{“multiplication rule”}$$

If statistical independence is given for the events $A$ and $B$, and if $0 < P(B) < 1$, then the following identities hold:

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

$$P(A \mid B) = P(A \mid \overline{B})$$

$$P(A \mid B) = P(A)$$

# Conditional Probability

## Independence of Events (continued)

**Definition 10 (Statistical Independence of two Events)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then $A$ and $B$ are called statistically independent iff the following equation holds:

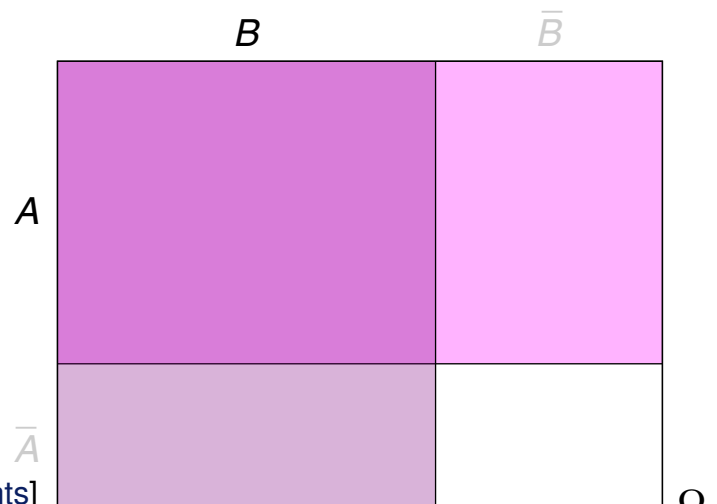
$$P(A \cap B) = P(A) \cdot P(B) \qquad \text{"multiplication rule"}$$

If statistical independence is given for the events $A$ and $B$, and if $0 < P(B) < 1$, then the following identities hold:

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

$$P(A \mid B) = P(A \mid \overline{B})$$

$$P(A \mid B) = P(A)$$




[dependent events]

# Conditional Probability

## Independence of Events (continued)

**Definition 10 (Statistical Independence of two Events)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then $A$ and $B$ are called statistically independent iff the following equation holds:

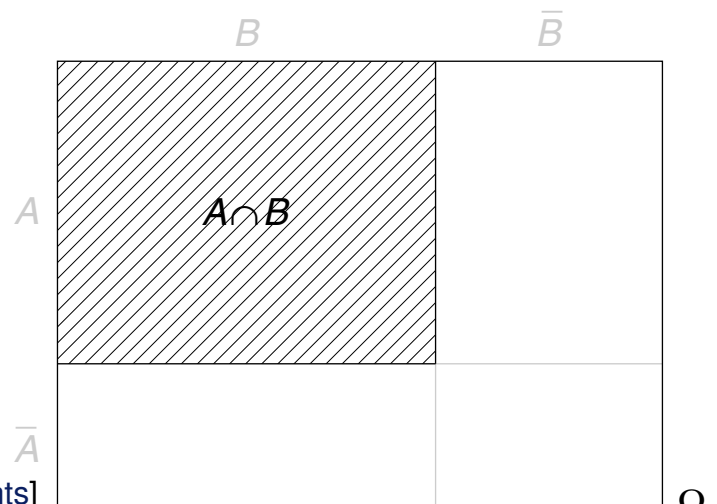
$$P(A \cap B) = P(A) \cdot P(B) \qquad \text{"multiplication rule"}$$

If statistical independence is given for the events $A$ and $B$, and if $0 < P(B) < 1$, then the following identities hold:

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

$$P(A \mid B) = P(A \mid \overline{B})$$

$$P(A \mid B) = P(A)$$




[dependent events]

# Conditional Probability

## Independence of Events (continued)

**Definition 10 (Statistical Independence of two Events)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then $A$ and $B$ are called statistically independent iff the following equation holds:

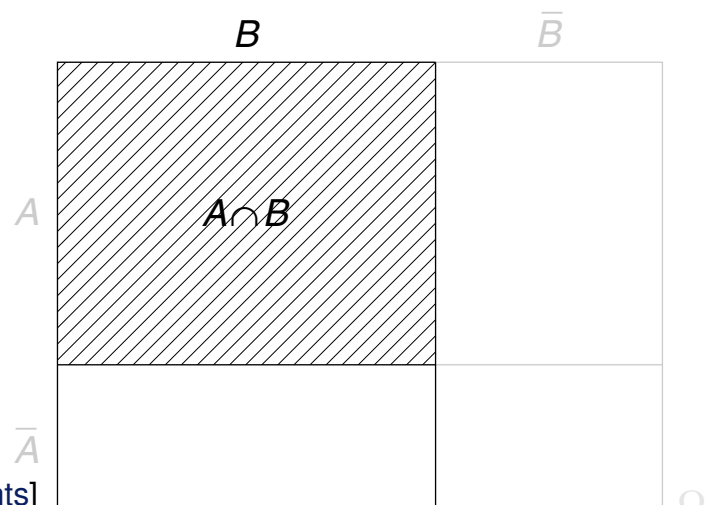
$$P(A \cap B) = P(A) \cdot P(B) \qquad \text{"multiplication rule"}$$

If statistical independence is given for the events $A$ and $B$, and if $0 < P(B) < 1$, then the following identities hold:

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

$$P(A \mid B) = P(A \mid \overline{B})$$

$$P(A \mid B) = P(A)$$




[dependent events]

# Conditional Probability

## Independence of Events (continued)

**Definition 10 (Statistical Independence of two Events)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then $A$ and $B$ are called statistically independent iff the following equation holds:

$$P(A \cap B) = P(A) \cdot P(B) \qquad \text{“multiplication rule”}$$

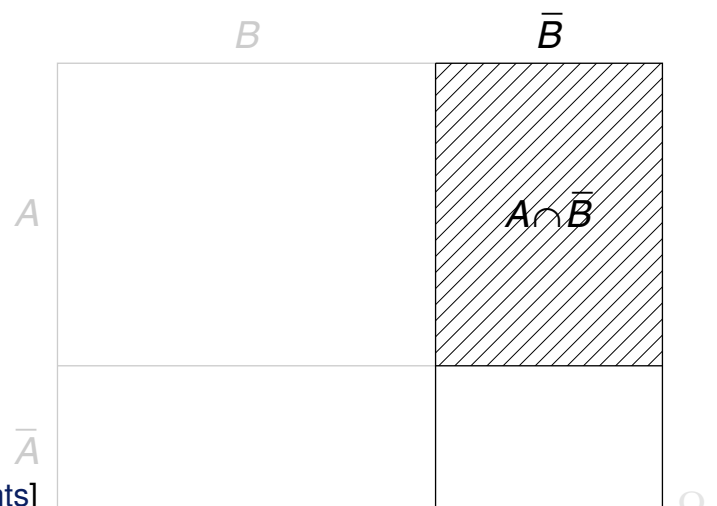
If statistical independence is given for the events $A$ and $B$, and if $0 < P(B) < 1$, then the following identities hold:

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

$$P(A \mid B) = P(A \mid \overline{B})$$

$$P(A \mid B) = P(A)$$

[dependent events]



# Conditional Probability

## Independence of Events (continued)

**Definition 10 (Statistical Independence of two Events)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then $A$ and $B$ are called statistically independent iff the following equation holds:

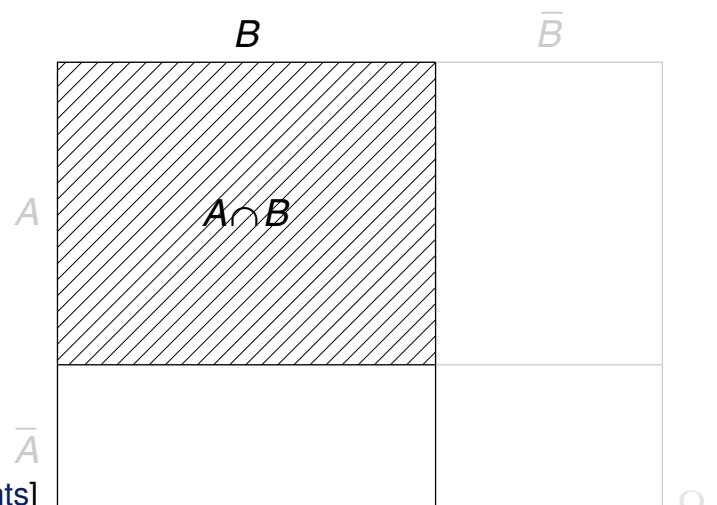
$$P(A \cap B) = P(A) \cdot P(B) \qquad \text{“multiplication rule”}$$

If statistical independence is given for the events $A$ and $B$, and if $0 < P(B) < 1$, then the following identities hold:

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

$$P(A \mid B) = P(A \mid \overline{B})$$

$$P(A \mid B) = P(A)$$




[dependent events]

# Conditional Probability

## Independence of Events (continued)

**Definition 10 (Statistical Independence of two Events)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then $A$ and $B$ are called statistically independent iff the following equation holds:

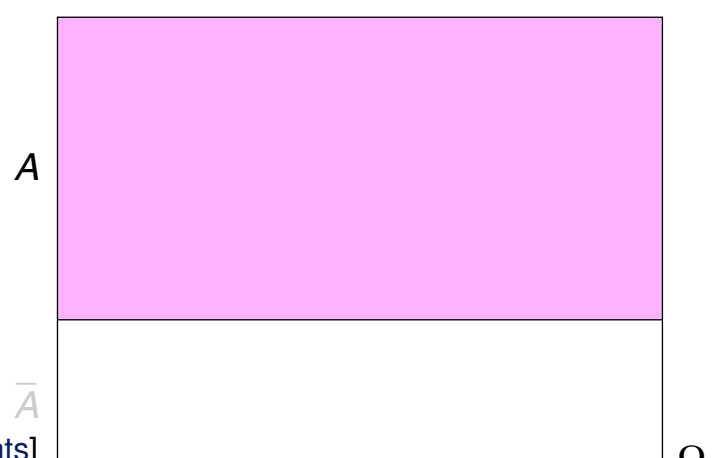
$$P(A \cap B) = P(A) \cdot P(B) \qquad \text{“multiplication rule”}$$

If statistical independence is given for the events $A$ and $B$, and if $0 < P(B) < 1$, then the following identities hold:

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

$$P(A \mid B) = P(A \mid \overline{B})$$

$$P(A \mid B) = P(A)$$




[dependent events]

# Conditional Probability

## Independence of Events (continued)

**Definition 11 (Statistical Independence of $k$ Events)**

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A_1, \ldots, A_k \in \mathcal{P}(\Omega)$ be $k$ events. Then the $A_1, \ldots, A_k$ are called jointly statistically independent under $P$ iff for all subsets $\{A_{i_1}, \ldots, A_{i_l}\} \subseteq \{A_1, \ldots, A_k\}$ the multiplication rule holds:

$$P(A_{i_1} \cap \ldots \cap A_{i_l}) = P(A_{i_1}) \cdot \ \ldots \ \cdot P(A_{i_l}),$$

where $i_1 < i_2 < \ldots < i_l$ and $2 \leq l \leq k$.