Lecture 6:

Probability Theory Basics

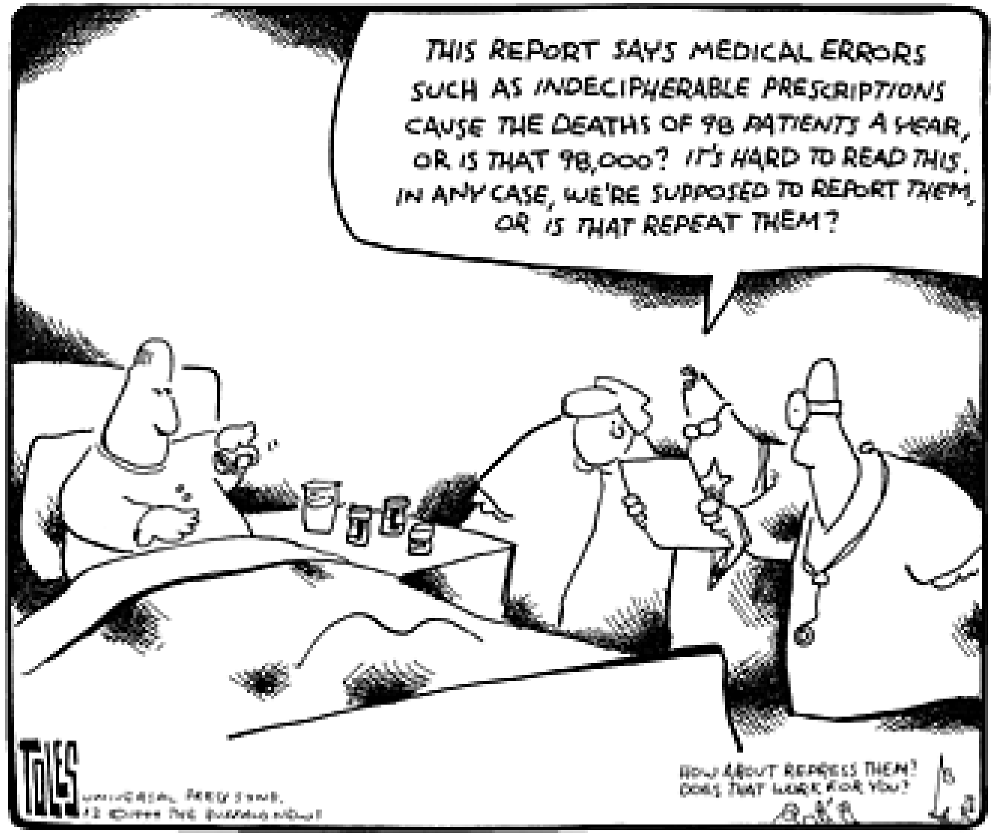
Markus Hohle

University California, Berkeley

Numerical Methods for Computational Science

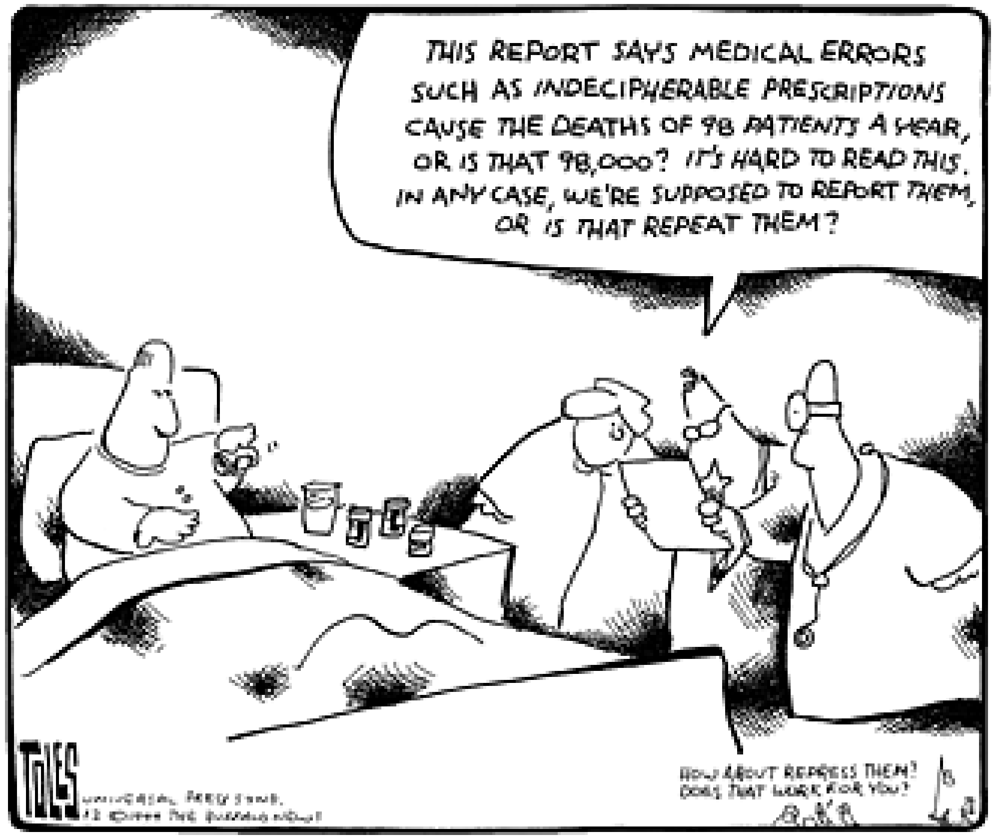
MSSE 273, 3 Units

Course Map



Outline

- Axioms of Probability
- Conditional Probabilities and Bayes Theorem
- Information and Entropy



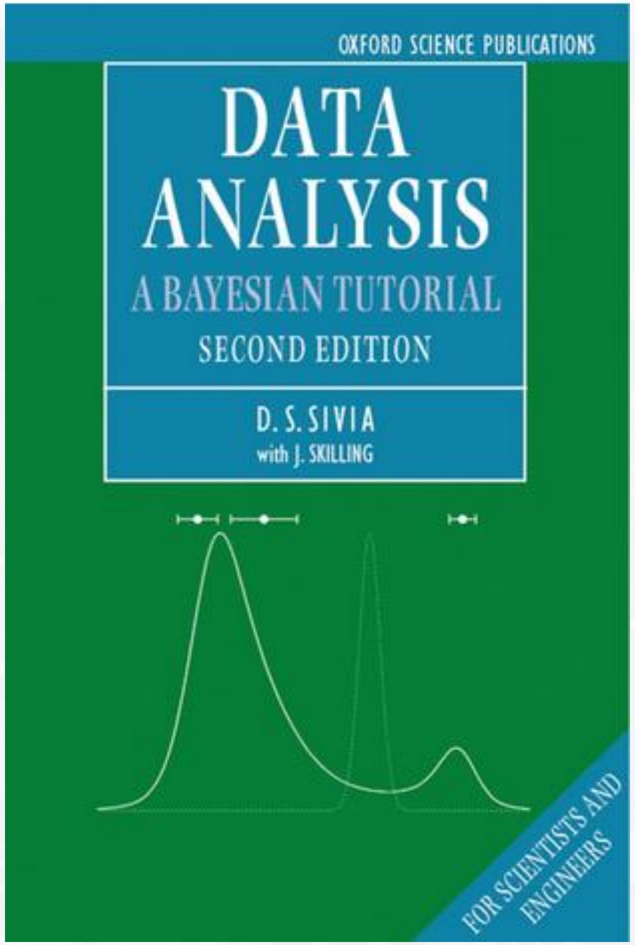
Outline

- **Axioms of Probability**

- Conditional Probabilities and Bayes Theorem

- Information and Entropy

here: **heuristic explanation** → more mathematical rigorous: see **Cox's theorem**



D. S. Sivia: *"Data Analysis"*

Bayesian Statistics
Appendix B: Cox's Theorem

Probability Theory Basics

here: **heuristic explanation** → more mathematical rigorous: see **Cox's theorem**

set Ω of events

probability $P(e_i)$, that an event e_i occurs

$P(e_i)$

probability *P* that we **observe** an event e_i, **given**, it **occurred**

$P(observe|e_i)$

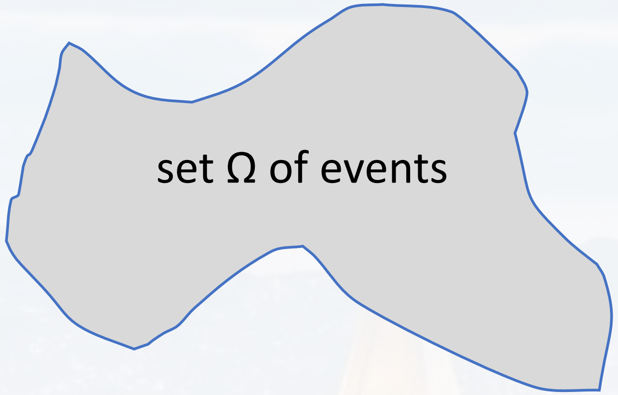
"given"

from data/observations → deriving Ω

from data/observations → understanding process
→ finding $P(e_i)$

Probability Theory Basics

here: **heuristic explanation** → more mathematical rigorous: see **Cox's theorem**

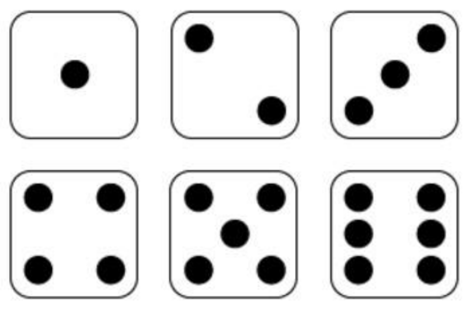


from data/observations → deriving Ω

1 2 2 1 2 2 1 2 1 1 2 1 1 1 2 observations

Is the observation 3 just rare **(and that's why we haven't observed it)**, or is 3 $\notin \Omega$?

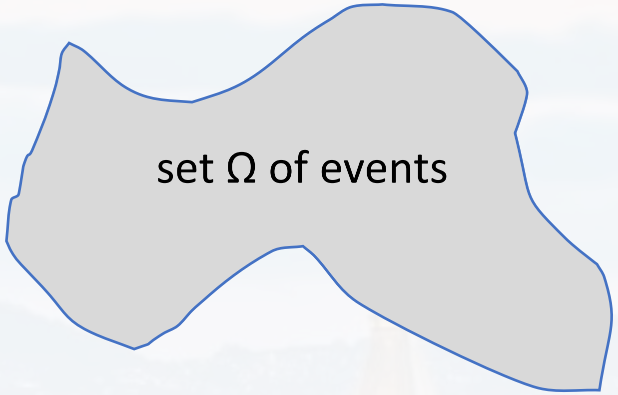
events can be ***discrete***



or ***continuous***

- car speed measured in a speed trap
- a person's weight, etc

$P(e_i)$, that an event e_i occurs



1st axiom: $P(e_i)$ is a **non-negative, real number**

2nd axiom: the probability that at **least one** of the events in the entire sample space will occur is **1** if events are **collectively exhaustive**

from 1st and 2nd : $P(e_i) = [0, 1]$ for any e_i

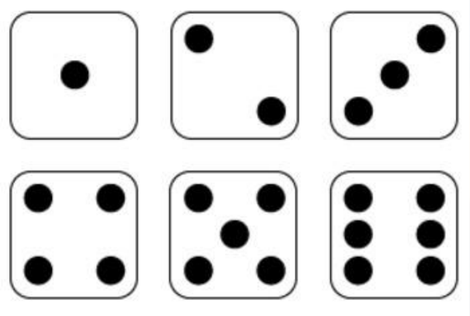
If events are **mutually exclusive**: **3rd axiom:** $P(\bigcup_{i=1}^{\infty} e_i) = \sum_{i=1}^{\infty} P(e_i)$

$$\bigcup_{i=1}^{\infty} e_i$$ means e_1 **or** e_2 **or** .... e_∞

$P(e_i)$, that an event e_i occurs

If events are **mutually exclusive**: **3rd axiom:** $P(\bigcup_{i=1}^{\infty} e_i) = \sum_{i=1}^{\infty} P(e_i)$

$\bigcup_{i=1}^{\infty} e_i$ means e_1 **or** e_2 **or** …. e_∞



The probability that we roll a **4 or a 6** equals…

$$P\left(e_4 \bigcup e_6\right) =$$

…the **probability** that we roll a **4 plus** the **probability** that we roll a **6**

$$P(e_4) + P(e_6)$$

"or" equals addition!

Probability Theory Basics

Axioms of Probability

$P(e_i)$, that an event e_i occurs

"or" equals addition!

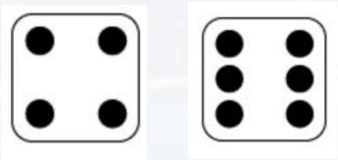
$$P(\cup_{i=1}^{\infty} e_i) = \sum_{i=1}^{\infty} P(e_i)$$

"and" equals multiplication!

$$P\left(\bigcap_{i=1}^{\infty} e_i\right) = \prod_{i=1}^{\infty} P(e_i)$$

two dice: The probability that we roll a **4 and a 6** equals…

$$P\left(e_4 \bigcap e_6\right) =$$



…the **probability** that we roll a **4 times** the **probability** that we roll a **6**

$$P(e_4)P(e_6)$$

$P(e_i)$, that an event e_i occurs

"or" equals addition!

$$P\left(\bigcup_{i=1}^{\infty} e_i\right) = \sum_{i=1}^{\infty} P(e_i)$$

"and" equals multiplication!

$$P\left(\bigcap_{i=1}^{\infty} e_i\right) = \prod_{i=1}^{\infty} P(e_i)$$

Be carful if events are not mutually exclusive (like a **set of events** or a **sequence of events**)

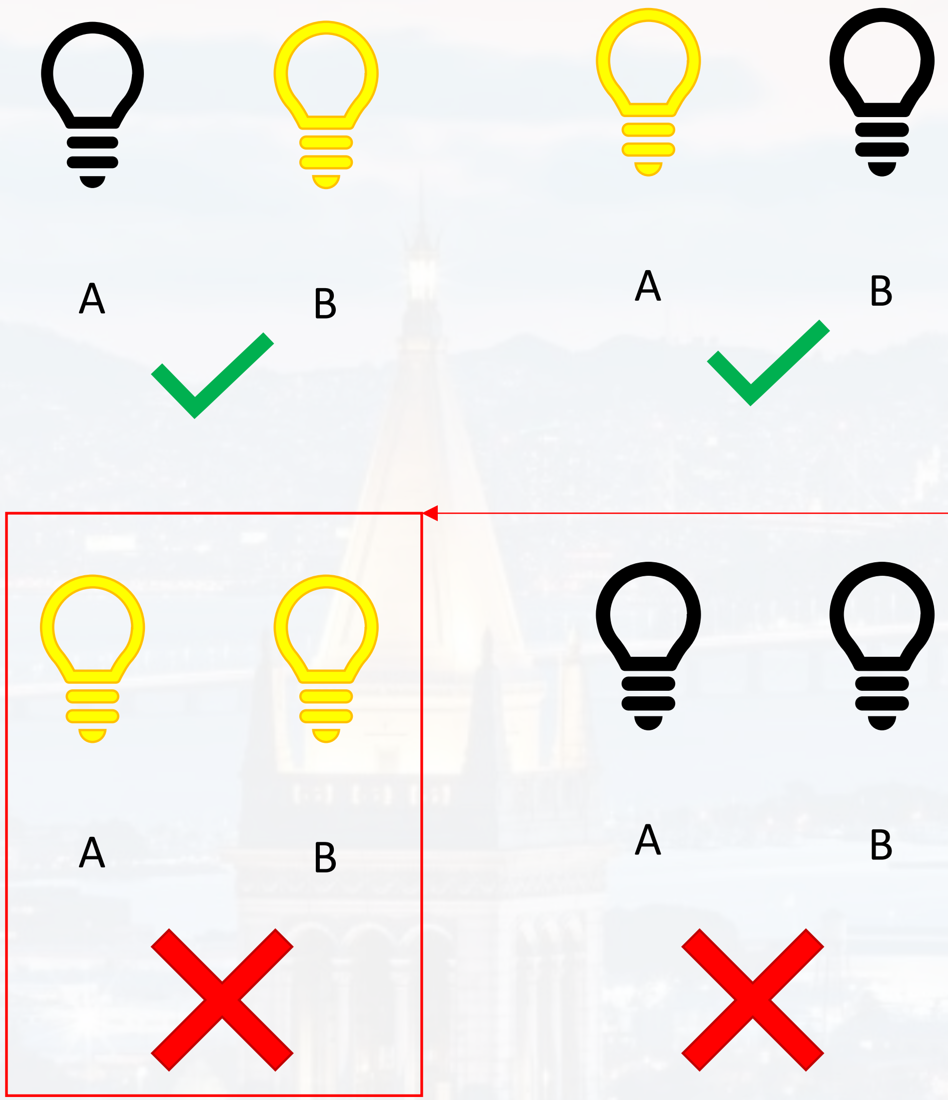
two light bulbs A and B:

What is the probability that A or B is turned on?

A B

Probability Theory Basics

two light bulbs A and B:



What is the probability that A **or** B is turned on?

$$P(A) + P(B)$$

$$\boxed{P(A)P(B)}$$

$$P\left(A \bigcup B\right) = P(A) + P(B) - P\left(A \bigcap B\right)$$

$$\boxed{= P(A) + P(B) - P(A)P(B)}$$

$P(e_i)$, that an event e_i occurs

"or" equals addition!

$$P(\cup_{i=1}^{\infty} e_i) = \sum_{i=1}^{\infty} P(e_i)$$

"and" equals multiplication!

$$P\left(\bigcap_{i=1}^{\infty} e_i\right) = \prod_{i=1}^{\infty} P(e_i)$$

$$P\left(e_4 \bigcup e_6\right) = P(e_4) + P(e_6) - P(e_4)P(e_6)$$

inclusion - exclusion principle

$$P\left(A \bigcup B\right) = P(A) + P(B) - P\left(A \bigcap B\right)$$

$$P\left(A \bigcup B \bigcup C\right) = P(A) + P(B) + P(B) + P(A)P(B)P(C) - P(A)P(B) - P(A)P(C) - P(C)P(B)$$

Axioms of Probability

$P(e_i)$, that an event e_i occurs

"or" equals addition!

$$P(\cup_{i=1}^{\infty} e_i) = \sum_{i=1}^{\infty} P(e_i)$$

"and" equals multiplication!

$$P\left(\bigcap_{i=1}^{\infty} e_i\right) = \prod_{i=1}^{\infty} P(e_i)$$

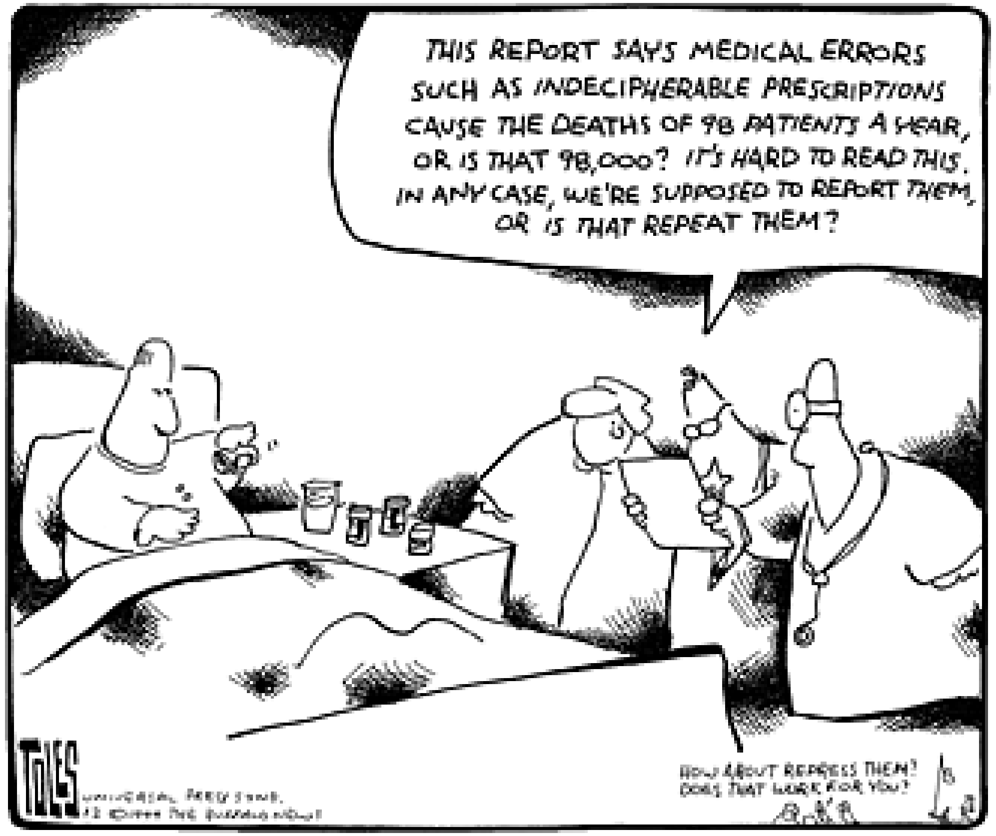
$$P\left(A \bigcup B\right) = P(A) + P(B) - P\left(A \bigcap B\right)$$

inclusion - exclusion principle

complement probability for not A, $\bar{A}$:

$$\boxed{P(\bar{A}) = 1 - P(A)}$$

because: $P(e_i) = [0, 1]$ for any e_i



Outline

- Axioms of Probability

- Conditional Probabilities and Bayes Theorem

- Information and Entropy

Conditional Probabilities and Bayes Theorem

$P(A \cap B)$ probability ***P*** that the events ***A* and *B*** occur

so far: A and B were independent $P(A \cap B) = P(A)P(B) = P(B)P(A)$

now: **conditional probabilities** | "given" or "under the condition"

Thomas Bayes (1701 - 1761)

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

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

Bayes Theorem

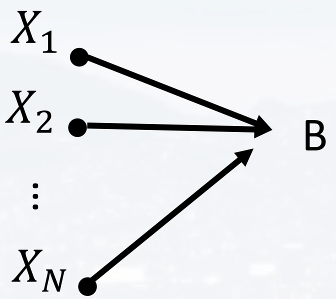
posterior $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$ **prior**

$$\boxed{P(A|B)P(B) = P(B|A)\mathrm{P(A)}}$$

Bayes Theorem

Thomas Bayes
(1701 - 1761)

posterior $\quad P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$ **prior**



$$P(B) = \sum_{n=1}^{N} P(B|X_n)P(X_n)$$

$$P(B) = \int P(B|X)P(X)\,dX$$

marginalization

Probability $P(B)$ that I am going to be too late for a meeting:

$$\begin{aligned} P(B) = {} & P(B|I\ forgot\ that\ I\ have\ a\ meeting)\ P(I\ forgot\ that\ I\ have\ a\ meeting) + \\ & P(B|I\ got\ sick)\ P(I\ got\ sick) + \\ & P(B|BART\ was\ too\ late)\ P(BART\ was\ too\ late) + \ \ldots \end{aligned}$$

example: cancer diagnosis from blood test

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

Marginalization $P(B) = \sum_{n=1}^{N} P(B|X_n)P(X_n)$

+ : positive test result
D : diseased
H : health

statement 1: If a person is **diseased**, there is a **95% probability** that the test is **positive.**

statement 2: The **prevalence** for the disease in the average **population** is **0.001%** .

statement 3: **5% of healthy** patients have **a positive result** (aka p-value).

A person takes the test and gets a positive test result. **What is the probability that the person is sick**?

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{0.95\,P(D)}{P(+)} = \frac{0.95 \cdot 0.00001}{P(+)} = \frac{0.95 \cdot 0.00001}{P(+|D)P(D) + P(+|H)P(H)}$$

statement 1 statement 2 marginalization

example: cancer diagnosis from blood test

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

+ : positive test result
D : diseased
H : health

Marginalization $P(B) = \sum_{n=1}^{N} P(B|X_n)P(X_n)$

statement 1: If a person is **diseased**, there is a **95% probability** that the test is **positive.**
statement 2: The **prevalence** for the disease in the average **population** is **0.001%** .
statement 3: **5% of healthy** patients have **a positive result** (aka p-value).

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{\mathbf{0.95}\, P(D)}{P(+)} = \frac{\mathbf{0.95} \cdot \mathbf{0.00001}}{P(+)} = \frac{\mathbf{0.95} \cdot \mathbf{0.00001}}{P(+|D)P(D) + P(+|H)P(H)}$$

statement 1 statement 2 marginalization

$$= \frac{\mathbf{0.95} \cdot \mathbf{0.00001}}{P(+|D)P(D) + P(+|H)[\mathbf{1 - P(D)}]}$$

complement probability

example: cancer diagnosis from blood test

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

+ : **positive test result**
D : **diseased**
H : **health**

Marginalization $P(B) = \sum_{n=1}^{N} P(B|X_n)P(X_n)$

statement 1: If a person is **diseased**, there is a **95% probability** that the test is **positive.**
statement 2: The **prevalence** for the disease in the average **population** is **0.001%** .
statement 3: **5% of healthy** patients have **a positive result** (aka p-value).

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|H)[1 - P(D)]}$$

$$\boxed{= \frac{1}{1 + \frac{P(+|H)[1 - P(D)]}{P(+|D)P(D)}}} = \frac{1}{1 + \frac{0.05\,[1 - 0.00001]}{0.95 \cdot 0.00001}} = 1/5000$$

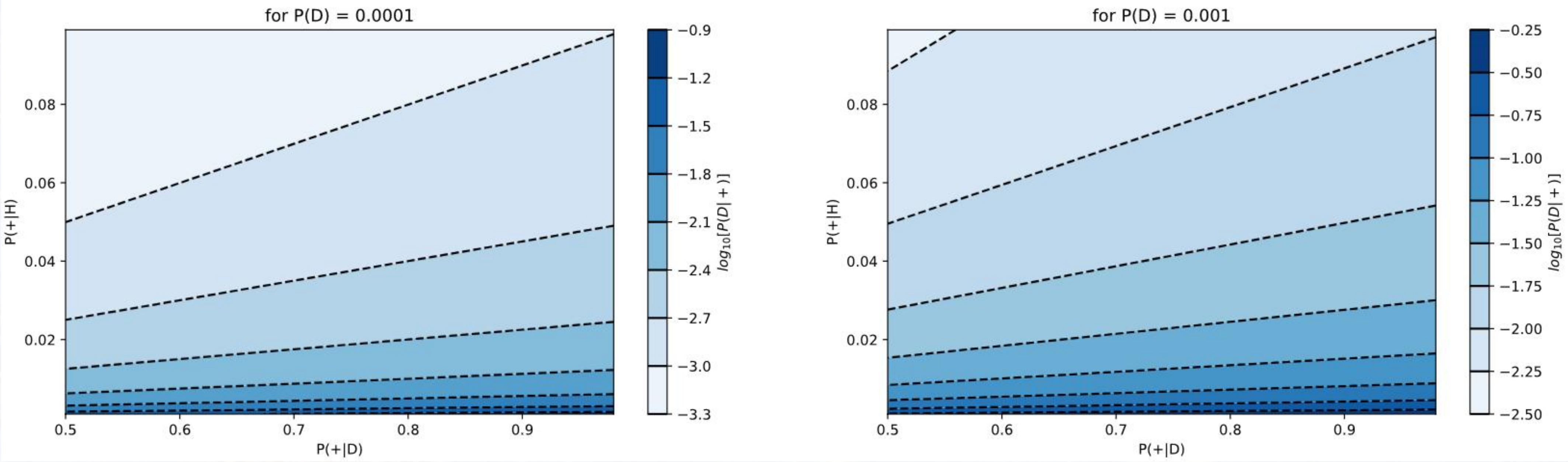
example: cancer diagnosis from blood test

+ : **positive test result**
D : **diseased**
H : **health**

statement 1:	sensitivity	$P(D\|+) = 95\%$
statement 2:	prior	$P(D) = 0.001\%$
statement 3:	p-value or false positive rate	$P(+\|H) = 5\%$

$$P(D|+) = \frac{1}{1 + \frac{P(+|H)[1 - P(D)]}{P(+|D)P(D)}}$$

check: `PlotPD_Plus.py`



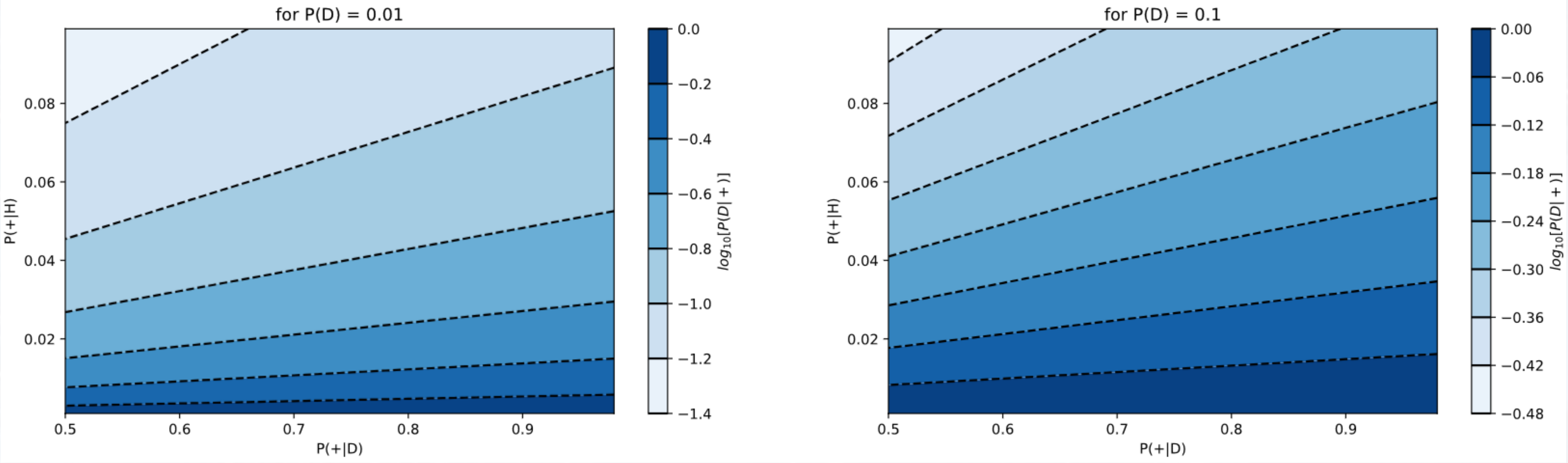
example: cancer diagnosis from blood test

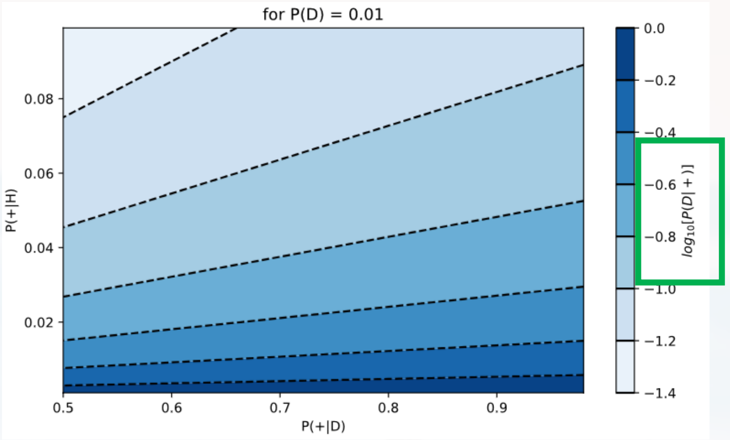
+ : positive test result
D : diseased
H : health

statement 1:	sensitivity	P(D|+) = 95%
statement 2:	prior	P(D) = 0.001%
statement 3:	p-value or false positive rate	P(+|H) = 5%

$$P(D|+) = \frac{1}{1 + \frac{P(+|H)[1 - P(D)]}{P(+|D)P(D)}}$$

check: `PlotPD_Plus.py`





statement 1:	sensitivity	P(D|+) = 95%
statement 2:	prior	P(D) = 0.001%
statement 3:	p-value or false positive rate	P(+|H) = 5%

odds ratios:

$$\rho_1 = \frac{P(+|H)}{P(+|D)}$$

$$\rho_2 = \frac{1-P(D)}{P(D)}$$

$$P(D|+) = \frac{1}{1+\dfrac{P(+|H)[1-P(D)]}{P(+|D)P(D)}}$$

log odds ratios:

$$r_1 = \log\left[\frac{P(+|H)}{P(+|D)}\right]$$

$$r_2 = \log\left[\frac{1-P(D)}{P(D)}\right]$$

$$P(D|+) = \frac{1}{1+e^{r_1}e^{r_2}}$$

Probability Theory Basics

log odds ratios: $r_1 = \log\left[\frac{P(+|H)}{P(+|D)}\right]$

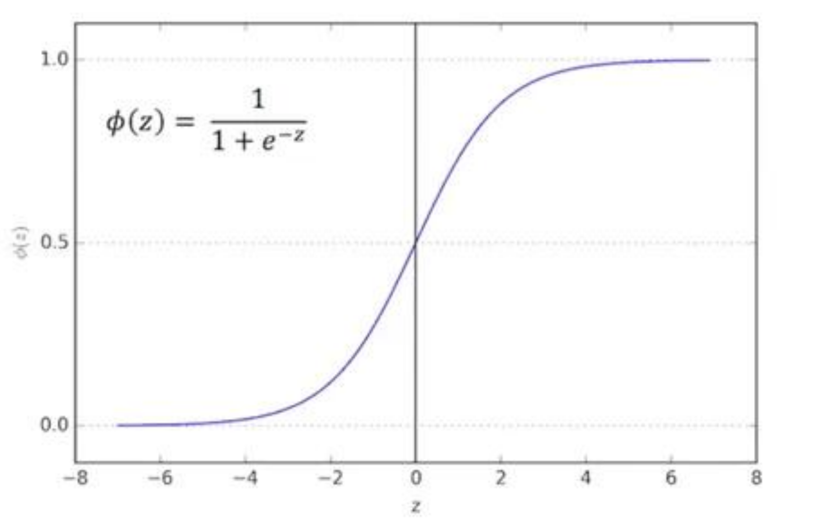
$r_2 = \log\left[\frac{1-P(D)}{P(D)}\right]$

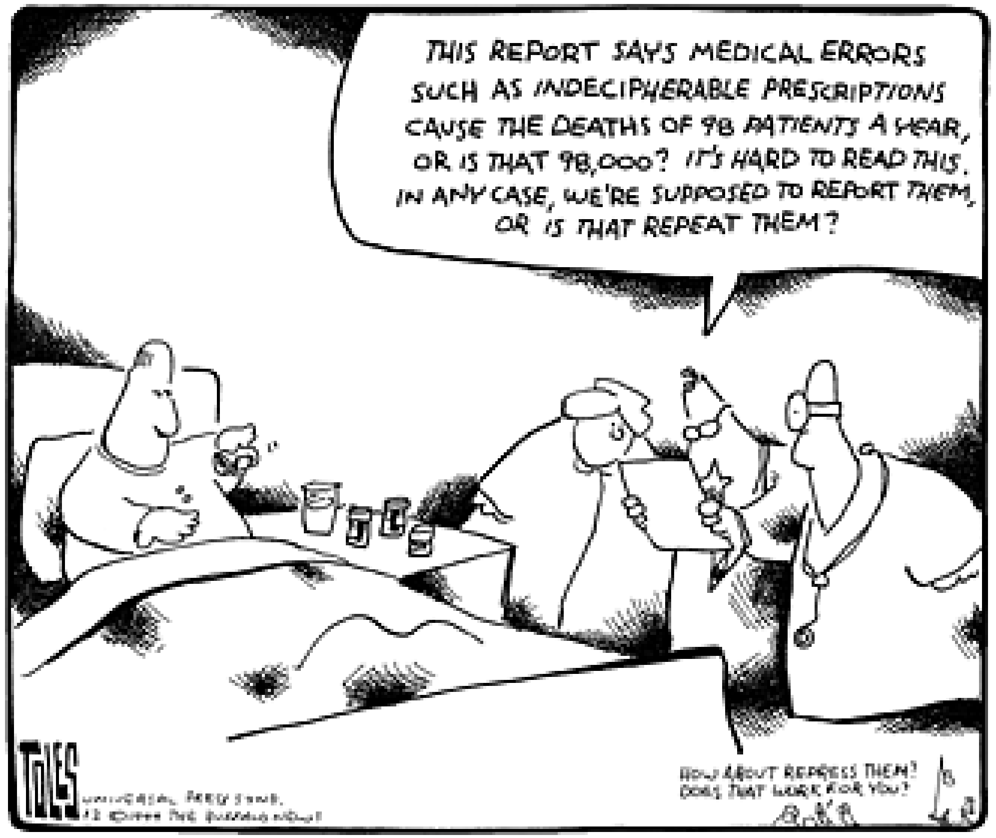
$$P(D|+) = \frac{1}{1 + \frac{P(+|H)[1-P(D)]}{P(+|D)P(D)}}$$

$$\boxed{P(D|+) = \frac{1}{1 + e^{r_1} e^{r_2}}}$$

logistic (or logit or sigmoid) function

- logistic regression
- transfer function ANN
- bound growth (Verhulst equation)
- binding affinity ligand/receptor





Outline

- Axioms of Probability
- Conditional Probabilities and Bayes Theorem
- **Information and Entropy**

<u>in this section:</u>

entropy S is a **measure of information** we have **about a system**

without mathematical proof: $\boldsymbol{S = -\sum_{i=1}^{I} p_i \, ln(p_i)}$

entropy is important in statistics, physics, informatics etc

important: **entropy has nothing to do with order/disorder**

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

problem:

I want to rewatch a series on Netflix.
Where have I left off, i. e. which episodes have I watched already ?

Probability Theory Basics

entropy S is a **measure of information** we have **about a system**

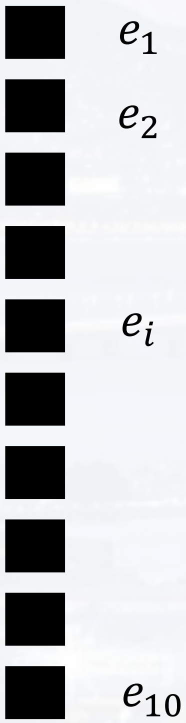
problem:
I want to rewatch a series on Netflix. Where have I left off, i. e. which episodes have I watched already?

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

p_i: probability that I have watched episode e_i

case 1): no idea, no information → all $p_i = 0.5$

list of episodes



Probability Theory Basics

entropy S is a **measure of information** we have **about a system**

problem:
I want to rewatch a series on Netflix. Where have I left off, i. e. which episodes have I watched already?

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

p_i: probability that I have watched episode e_i

case 1): no idea, no information → all $p_i = 0.5$

$$S = -\sum_{i=1}^{10} \frac{1}{2} \, ln\left(\frac{1}{2}\right) = \sum_{i=1}^{10} \frac{1}{2} \, ln(2) = \frac{5}{2} \, ln(2) \approx \mathbf{1.73}$$

list of episodes

e_1

e_2

e_i

e_{10}

Probability Theory Basics

entropy S is a **measure of information** we have **about a system**

problem:
I want to rewatch a series on Netflix. Where have I left off, i. e. which episodes have I watched already?

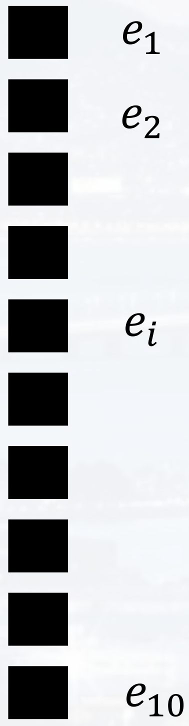
$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

p_i: probability that I have watched episode e_i

case 1): no idea, no information → all $p_i = 0.5$ $\quad S \approx \mathbf{1.73}$

case 2): usually, I remember that I have **watched some of the first episodes**, I am **not sure about 2 or 3 episodes** and I know that I **haven't watched the last episodes**

list of episodes



Probability Theory Basics

entropy S is a **measure of information** we have **about a system**

problem:
I want to rewatch a series on Netflix. Where have I left off, i. e. which episodes have I watched already?

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

p_i: probability that I have watched episode e_i

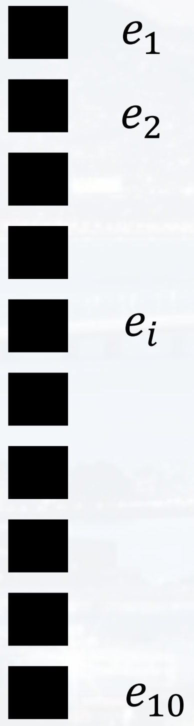
case 1): no idea, no information → all $p_i = 0.5$ $\quad S \approx \mathbf{1.73}$

case 2): usually, I remember that I have **watched some of the first episodes**, I am **not sure about 2 or 3 episodes** and I know that I **haven't watched the last episodes**

$$S = -\sum_{i=1}^{4} 1 \, ln(1) \; - 0.75 \, ln(0.75) - 0.5 \, ln(0.5)$$

$$- 0.25 \, ln(0.25) - \sum_{i=1}^{3} 0 \, ln(0) \approx \mathbf{0.91}$$

list of episodes



entropy S is a **measure of information** we have **about a system**

problem:
I want to rewatch a series on Netflix. Where have I left off, i. e. which episodes have I watched already?

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

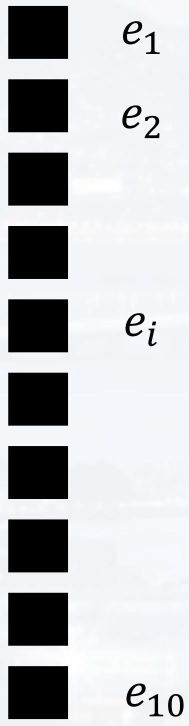
p_i: probability that I have watched episode e_i

case 1): no idea, no information → all $p_i = 0.5$ $S \approx \mathbf{1.73}$

case 2): some information $S \approx \mathbf{0.91}$

case 3): **all** information

list of episodes



$$S = -\sum_{i=1}^{10} 1 \, ln(1) = 0$$

Probability Theory Basics

entropy S is a **measure of information** we have **about a system**

problem:
I want to rewatch a series on Netflix. Where have I left off, i. e. which episodes have I watched already?

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

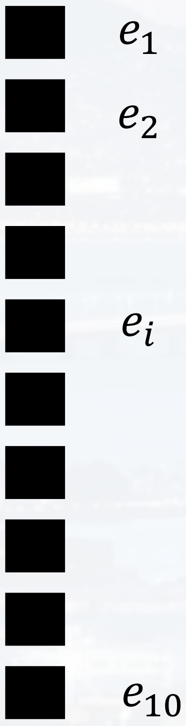
p_i: probability that I have watched episode e_i

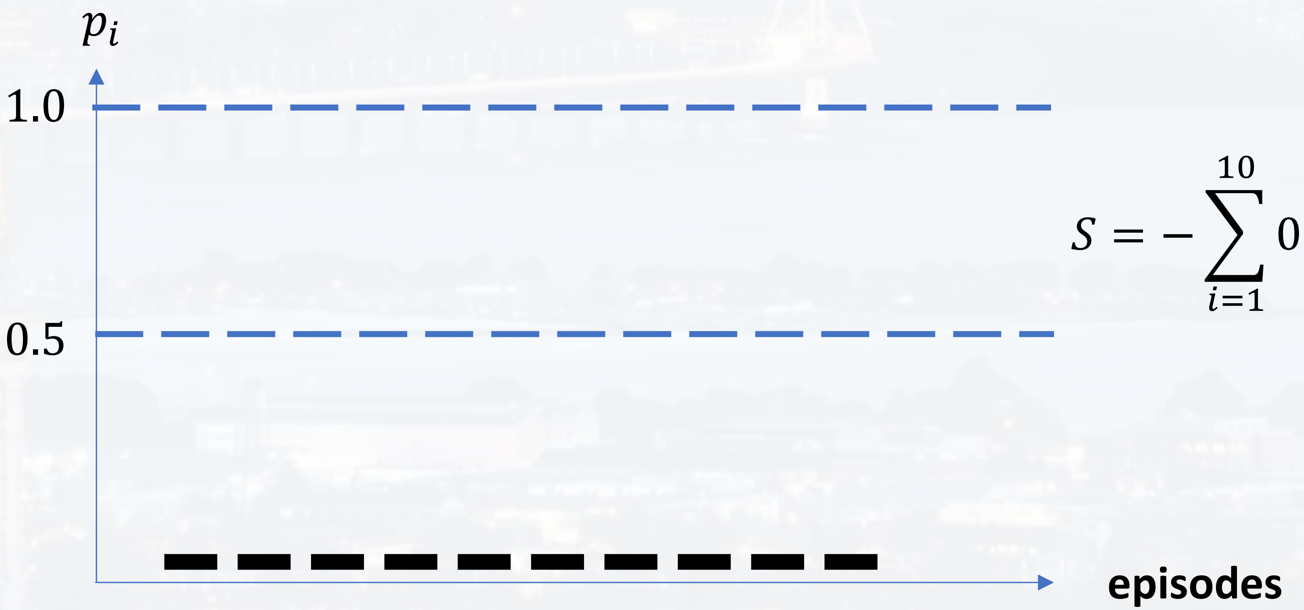
case 1): no idea, no information → all $p_i = 0.5$ $S \approx \mathbf{1.73}$

case 2): some information $S \approx \mathbf{0.91}$

case 3): **all** information

list of episodes





$$S = -\sum_{i=1}^{10} 0 \, ln(0) = 0$$

Probability Theory Basics

entropy S is a **measure of information** we have **about a system**

problem:
I want to rewatch a series on Netflix. Where have I left off, i. e. which episodes have I watched already?

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

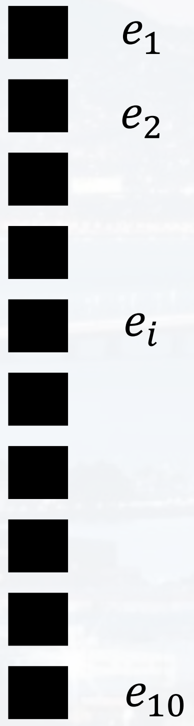
p_i: probability that I have watched episode e_i

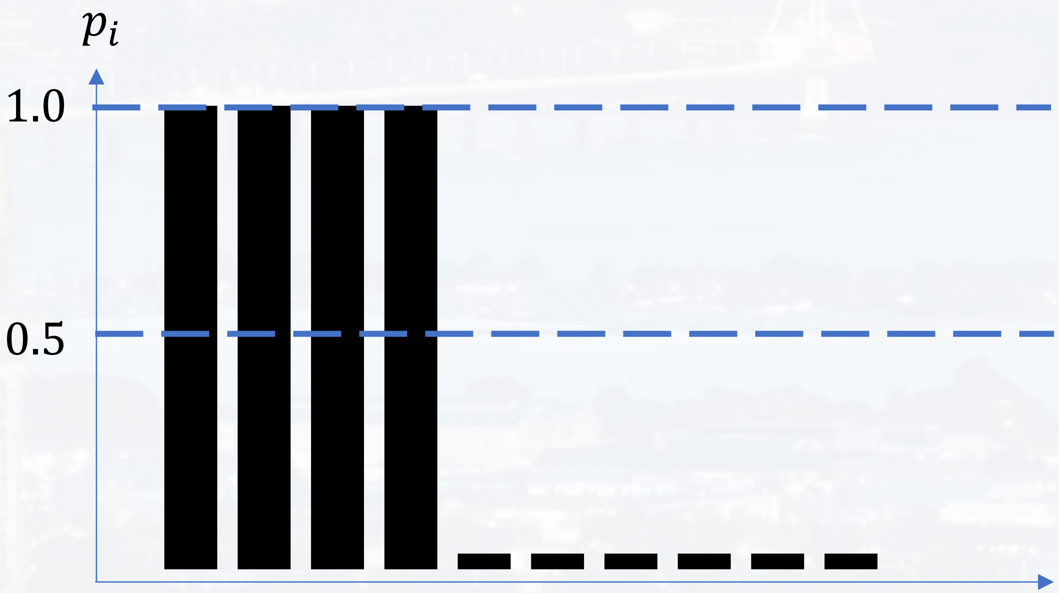
case 1): no idea, no information → all $p_i = 0.5$ $S \approx \mathbf{1.73}$

case 2): some information $S \approx \mathbf{0.91}$

case 3): **all** information

list of episodes





$$S = -\sum_{i=1}^{6} 0 \, ln(0)$$

$$+\sum_{i=1}^{4} 1 \, ln(1) = 0$$

entropy S is a **measure of information** we have **about a system**

problem:
I want to rewatch a series on Netflix. Where have I left off, i. e. which episodes have I watched already?

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

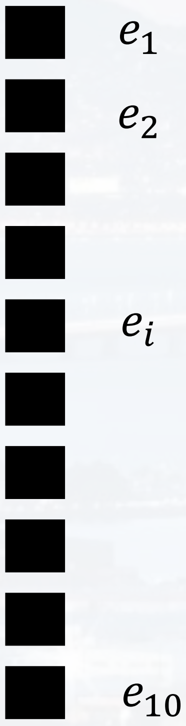
p_i: probability that I have watched episode e_i

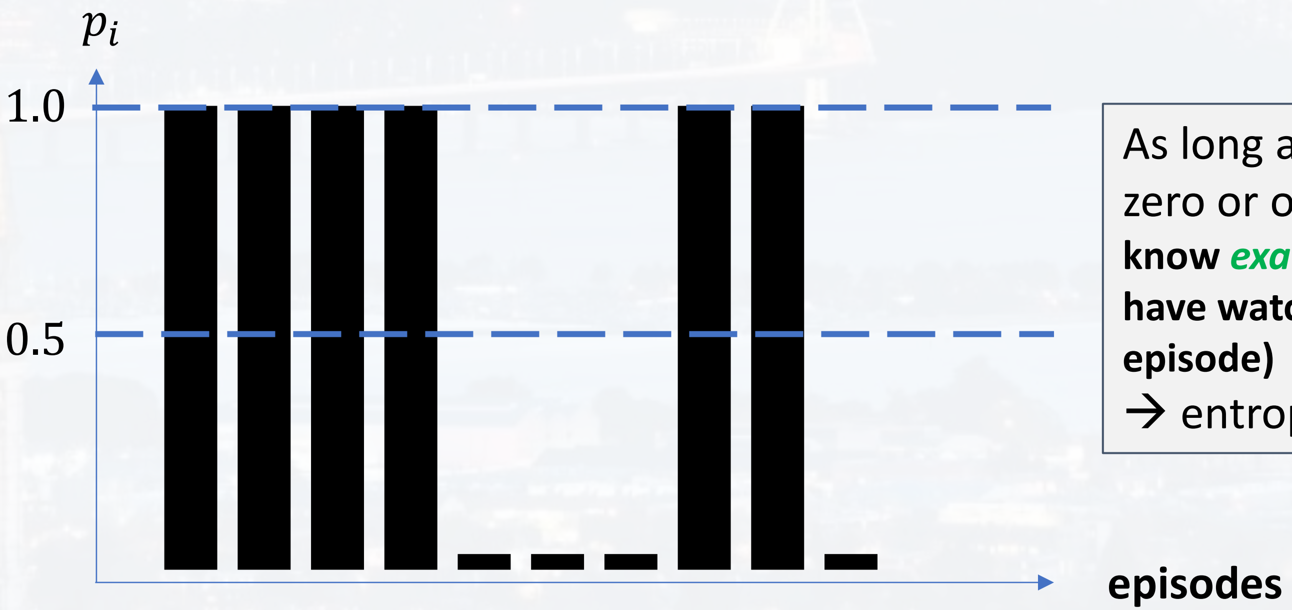
case 1): no idea, no information → all $p_i = 0.5$ $S \approx \mathbf{1.73}$

case 2): some information $S \approx \mathbf{0.91}$

case 3): **all** information

list of episodes





As long as the p_i are zero or one **(i. e. I know *exactly* if I have watched the episode)**
→ entropy = 0

Probability Theory Basics

entropy S is a **measure of information** we have **about a system**

problem:
I want to rewatch a series on Netflix. Where have I left off, i. e. which episodes have I watched already?

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

p_i: probability that I have watched episode e_i

case 1):	no idea, no information → all $p_i = 0.5$	$S \approx \mathbf{1.73}$
case 2):	some information	$S \approx \mathbf{0.91}$
case 3):	**all** information	$S \approx \mathbf{0.00}$

list of episodes

e_1

e_2

e_i

e_{10}

The lower the entropy, the more information!

If $p_i = 0.5 = \bar{p}_i$, maximum entropy (uniform distribution)!

Probability Theory Basics

entropy S is a **measure of information** we have **about a system**

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

two states ↑ or ↓

and three entities → system

possible states of the system

all three up	two up	one up	all three down
↑↑↑	↑↑↓	↑↓↓	↓↓↓
	↑↓↑	↓↑↓	
	↓↑↑	↓↓↑	

eight possible states

no idea (before the experiment)

$$S = -\sum_{i=1}^{8} \frac{1}{8} \, ln\left(\frac{1}{8}\right) = \sum_{i=1}^{8} \frac{1}{8} \, ln(8) = ln(8) \approx \mathbf{2.08}$$

Probability Theory Basics

entropy S is a **measure of information** we have **about a system**

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

two states ↑ or ↓

and three entities → system

possible states of the system

all three up	two up	one up	all three down
↑↑↑	↑↑↓	↑↓↓	~~↓↓↓~~
	↑↓↑	↓↑↓	
	↓↑↑	↓↓↑	

one measurement
→ at least one arrow up

seven possible states

$$S = \sum_{i=1}^{7} \frac{1}{7} \, ln(7) + 0 \, \ln(0) \; = ln(7) \approx \mathbf{1.95}$$

entropy S is a **measure of information** we have **about a system**

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

two states ↑ or ↓

and three entities → system

possible states of the system

all three up	two up	one up	all three down
↑↑↑	↑↑↓	~~↑↓↓~~	~~↓↓↓~~
	↑↓↑	~~↓↑↓~~	
	↓↑↑	~~↓↓↑~~	

second measurement
→ another arrow up

four possible states

$$S = \sum_{i=1}^{4} \frac{1}{4} \, ln(4) \; = ln(4) \approx \mathbf{1.39}$$

entropy S is a **measure of information** we have **about a system**

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

two states ↑ or ↓

and three entities → system

possible states of the system

all three up	two up	one up	all three down
~~↑↑↑~~	↑↑↓	~~↑↓↓~~	~~↓↓↓~~
	↑↓↑	~~↓↑↓~~	
	↓↑↑	~~↓↓↑~~	

third measurement
→ one arrow down

three possible states

$S = ln(3) \approx \mathbf{1.10}$

The lower the entropy, the more information!

Thank you very much for you attention!