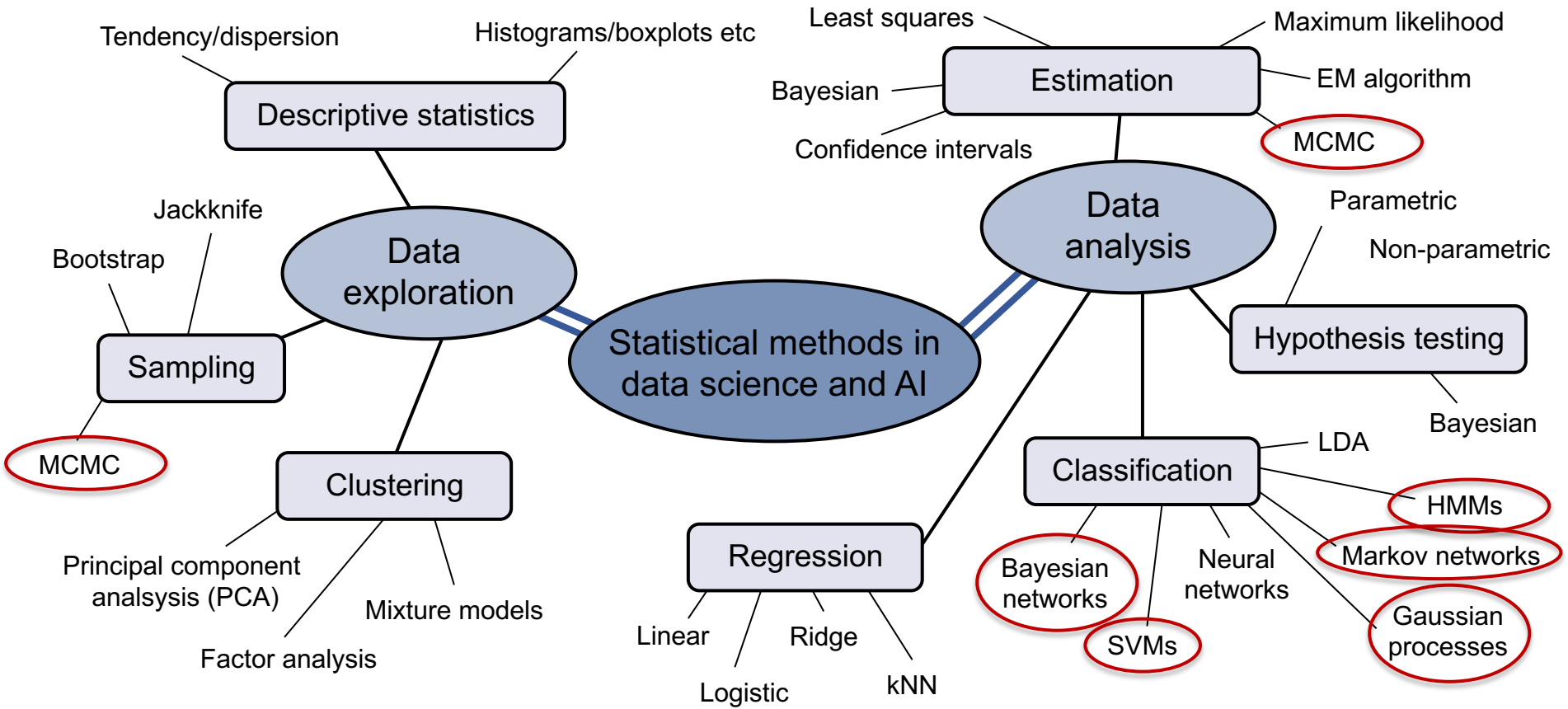
CHALMERS
UNIVERSITY OF TECHNOLOGY

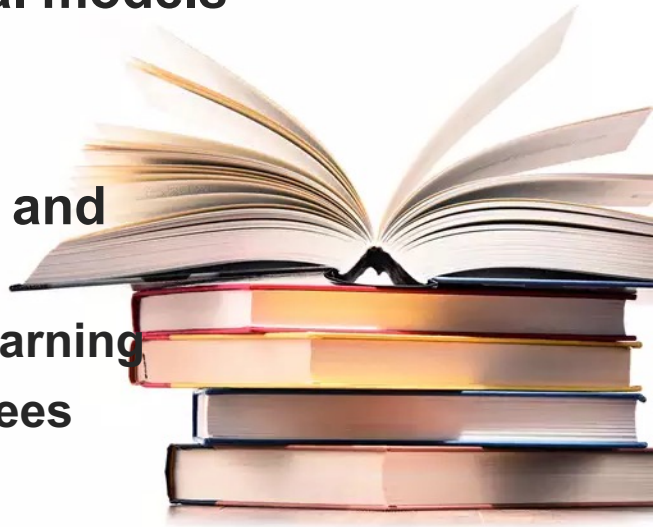
# DAT405 – Part 2:
# Statistical methods in Data Science and AI

DAT405/DIT407, LP4 2022-2023, Module 4


Statistical methods in data science and AI
Data exploration
Data analysis
Descriptive statistics
Tendency/dispersion
Histograms/boxplots etc
Sampling
Jackknife
Bootstrap
MCMC
Clustering
Principal component analsysis (PCA)
Factor analysis
Mixture models
Estimation
Least squares
Maximum likelihood
Bayesian
EM algorithm
Confidence intervals
MCMC
Parametric
Non-parametric
Hypothesis testing
Bayesian
Classification
LDA
HMMs
Markov networks
Gaussian processes
Bayesian networks
SVMs
Neural networks
Regression
Linear
Ridge
Logistic
kNN

# Syllabus – part 2

- Module 4: Bayesian statistics and graphical models
  - Lecture 7: Bayesian statistics
  - Lecture 8: Graphical models
- Module 5: Markov models, kernel methods and decision trees
  - Lecture 9: Markov models, reinforcement learning
  - Lecture 10: Kernel methods and decision trees



CHALMERS
UNIVERSITY OF TECHNOLOGY

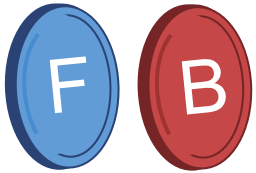
# Module 4: Bayesian statistics

# Probability theory and statistics

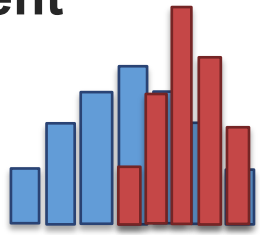
- a quick refresher

# Probability versus statistics

## Probability – predict the likelihood of a future event

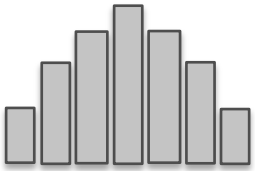



**Probability**
Given model, predict data



## Statistics – estimate the frequency of a past event


**Statistics**
Given data, predict model



# Sample space, events and random experiments

- A ***random experiment*** is a process that produces random ***outcomes***.
- The ***sample space*** is the set of all possible outcomes in an experiment.
- An ***event*** is the outcome, or a subset of possible outcomes, of an experiment.

# Example: roll a die

- **Sample space: $S = \{1, 2, \ldots, 6\} = 6$ outcomes**
- **Events:**
  - "At least 3" $= \{3, 4, 5, 6\}$
  - "Six" $= \{6\}$
  - "Odd" $= \{1, 3, 5\}$
- **Probabilities**

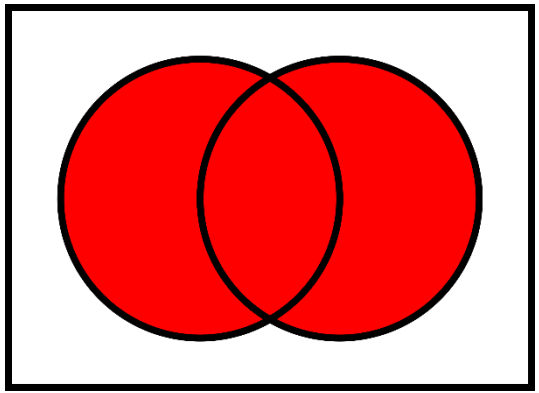
  $P(\text{at least } 3) = 4/6$

  $P(\text{six}) = 1/6$

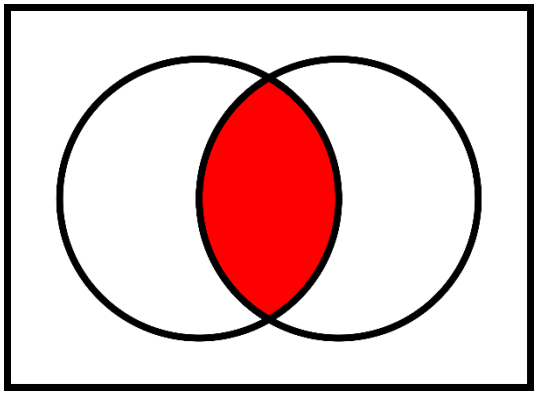
  $P(\text{odd}) = 3/6$

# Venn diagrams of set operations

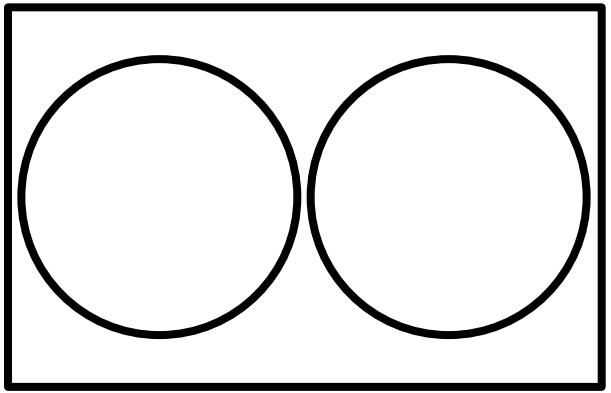
Union: $A \cup B$




Intersection: $A \cap B$




Mutually exclusive: $A \cap B = \phi$



# Combining events

- Now assume we want to combine two events
  - A= ”at least 3” B = ”odd number”
- **Union**

  A ∪ B = {3,4,5,6} ∪ {1,3,5} = {1,3,4,5,6}

  $\boldsymbol{P}$(A ∪ B) = 5/6
- **Intersection**

  A ∩ B = {3,4,5,6} ∩ {1,3,5} = {3,5}

  $\boldsymbol{P}$(A ∩ B) = 2/6

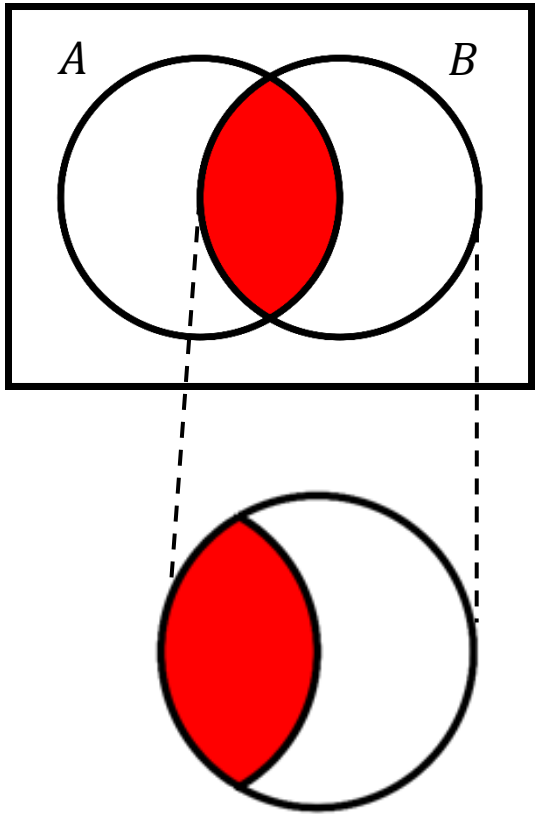
# Conditional probability

- The *conditional* probability of an event $A$ given the knowledge that event $B$ occurred

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

- Note also

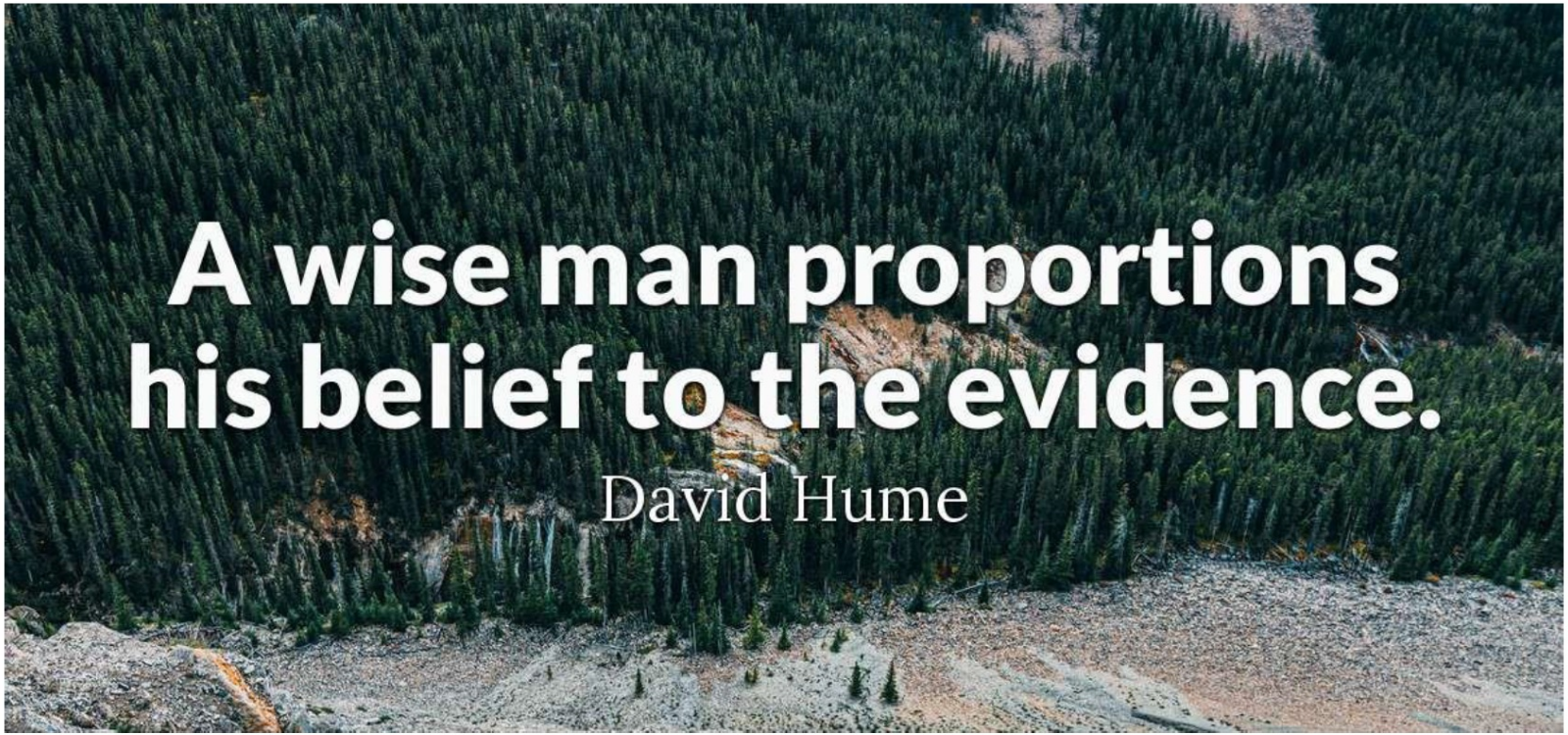
$$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$



# Thomas Bayes

(1701 – 1761)

- Developed the idea of using probability to represent **uncertainty about beliefs**
- Most importantly: gave a method on **updating beliefs** given new evidence


A wise man proportions
his belief to the evidence.
David Hume

# Bayes' rule

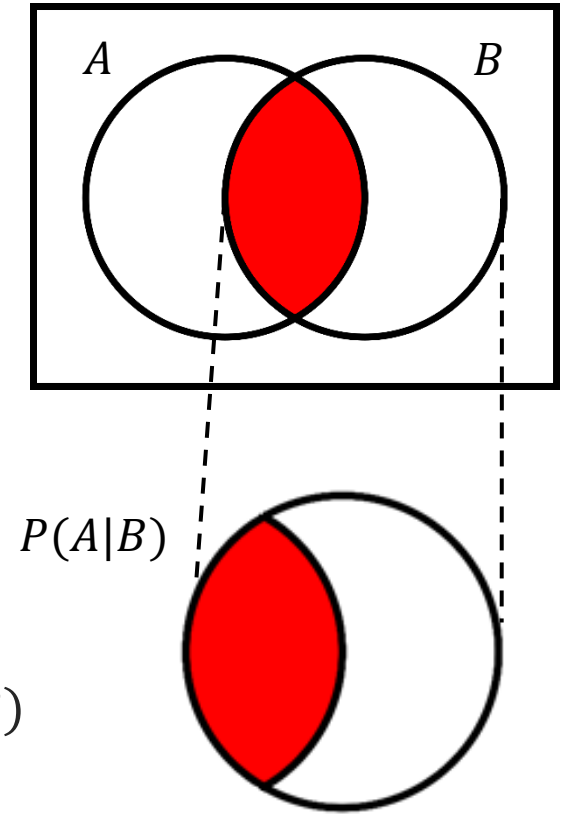
- **Bayes' rule**

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

**Proof:**

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

$\Rightarrow P(A|B)P(B) = P(B|A)P(A)$, **then divide by** $P(B)$



# Bayes' rule interpretation

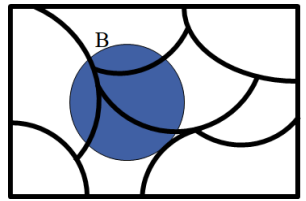
posterior

likelihood

prior

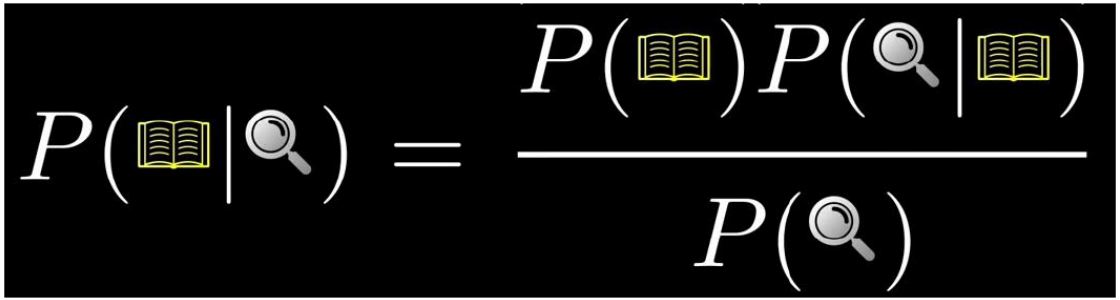
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

normalizer




We have prior information $P(A)$ of event $A$, and then update the posterior probability $P(A|B)$ as more information/data $B$ is achieved.

# Bayes' rule interpretation




**Prior:**
*Before* making observation 🔍 you think the probability of your hypothesis 📖 is $P(📖)$

**Posterior:**
*After* making observation 🔍 you think the probability of your hypothesis 📖 is $P(📖|🔍)$

# Example: spam or ham?

- Suppose I get an email with the word ”invest”. Is it more likely to spam or ham?
  - Hard to estimate P(spam|”invest”). Bayes’ rule will help!
- Estimated proportions of emails sent to me that are spam or ham:
  - P(spam) = 0.4
  - P(ham) = 0.6
- Proportions of emails containing the word ”invest”
  - P(”invest”|spam) = 0.05
  - P(”invest”|ham) = 0.01

Easier to estimate!

# Example: (cont)

$$P(\text{spam}|\text{"invest"}) = \frac{P(\text{"invest"}|\text{spam})P(\text{spam})}{P(\text{"invest"})} = \frac{0.05 \cdot 0.4}{P(\text{"invest"})} = \frac{0.02}{P(\text{"invest"})}$$

$$P(\text{ham}|\text{"invest"}) = \frac{P(\text{"invest"}|\text{ham})P(\text{ham})}{P(\text{"invest"})} = \frac{0.01 \cdot 0.6}{P(\text{"invest"})} = \frac{0.006}{P(\text{"invest"})}$$

$$\Rightarrow P(\text{spam}|\text{"invest"}) > P(\text{ham}|\text{"invest"})$$

Didn't have to estimate P("invest")!

# Mutually exclusive and exhaustive events

Events $E_1, E_2, \ldots, E_n$ are

Also called *pairwise disjoint*.

- ***mutually exclusive*** if they cannot occur simultaneously

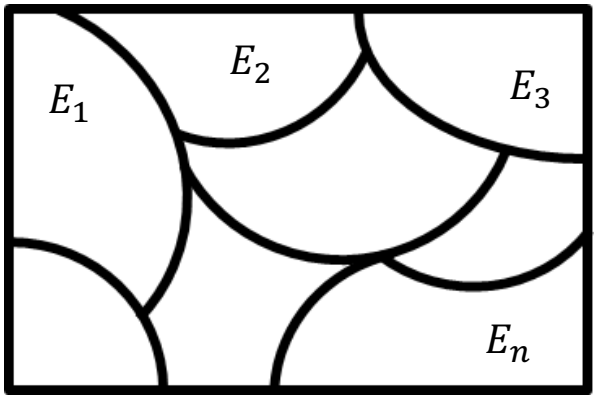
$$E_i \cap E_j = \phi, \; i \neq j$$

- ***exhaustive*** if they cover the sample space

$$E_1 \cup E_2 \cup \cdots \cup E_n = \bigcup_{i=1}^{n} E_i = S$$

Also called a *partition*.

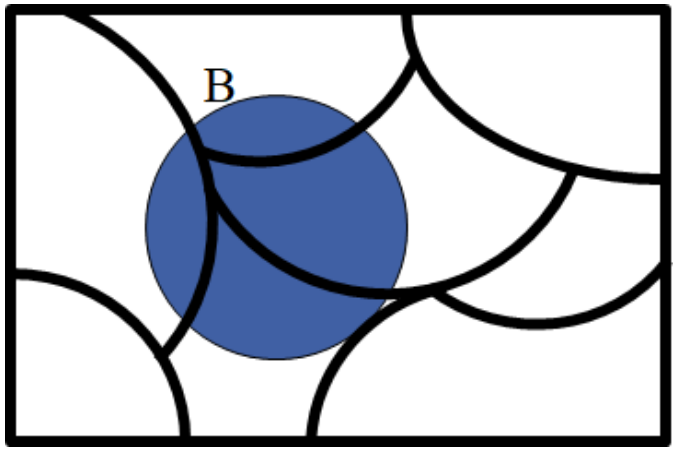
Sample space $S$



# Total law of probability

- **For mutually exclusive and exhaustive events $E_1, E_2, \ldots, E_n$ we get for any other event $B$**

$$P(B) = \sum_{i=1}^{n} P(B|E_i)$$




Example: P(apple) =
P(green apple) + P(red apple) + P(other apples)

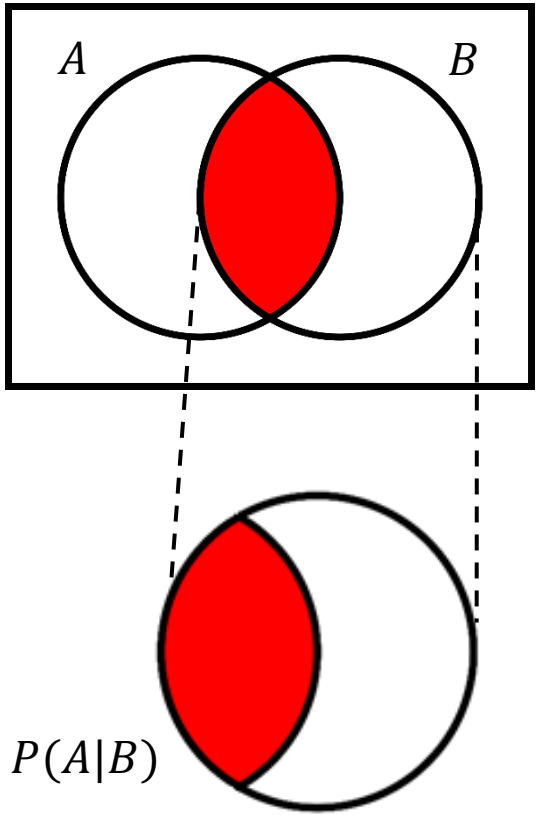
# Bayes' rule – extended

- **Bayes' rule**

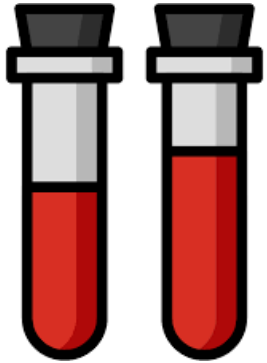
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- **For mutually exclusive and exhaustive events $E_1, E_2, \ldots, E_n$ we get**

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{\sum_{i=1}^{n} P(B|E_i)}$$



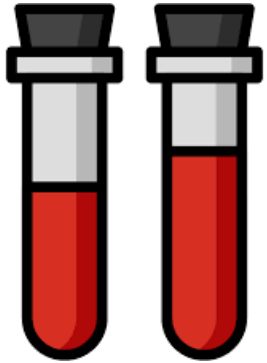
# Example: applying Bayes' rule



- Assume that 15 out of 10,000 individuals in a population have a certain disease $D$.
- The test is not perfect: when testing for the disease
  - an ill person always tests positive
  - a healthy person tests positive with probability 0.0002
- **Given that you tested positive, what is the probability that you have the disease?**

# Example (cont.)

**Bayes' rule:** $\boldsymbol{P}(\text{ill}\,|\text{positive}) = \dfrac{\boldsymbol{P}(\text{positive}|\text{ill})\boldsymbol{P}(\text{ill})}{\boldsymbol{P}(\text{positive})}$



- **We have**
  - $\boldsymbol{P}(\text{ill}) = \mathbf{0.0015}$ **and** $\boldsymbol{P}(\text{healthy}) = \mathbf{1 - 0.0015 = 0.9985}$
  - $\boldsymbol{P}(\text{positive}|\text{ill}) = \mathbf{1},\ \boldsymbol{P}(\text{positive}|\text{healthy}) = \mathbf{0.002}$
  - $\boldsymbol{P}(\text{positive}) = \boldsymbol{P}(\text{positive}|\text{ill})\boldsymbol{P}(\text{ill}) + \boldsymbol{P}(\text{positive}|\text{healthy})\boldsymbol{P}(\text{healthy})$

Would you call this test good?

**Hence**

$$\boldsymbol{P}(\text{ill}\,|\text{positive}) = \frac{\boldsymbol{P}(\text{positive}|\text{ill})\boldsymbol{P}(\text{ill})}{\boldsymbol{P}(\text{positive})} = \frac{1 \cdot 0.0015}{1 \cdot 0.0015 + 0.002 \cdot 0.9985} = \mathbf{0.43}$$

# Is Steve a librarian or a farmer?

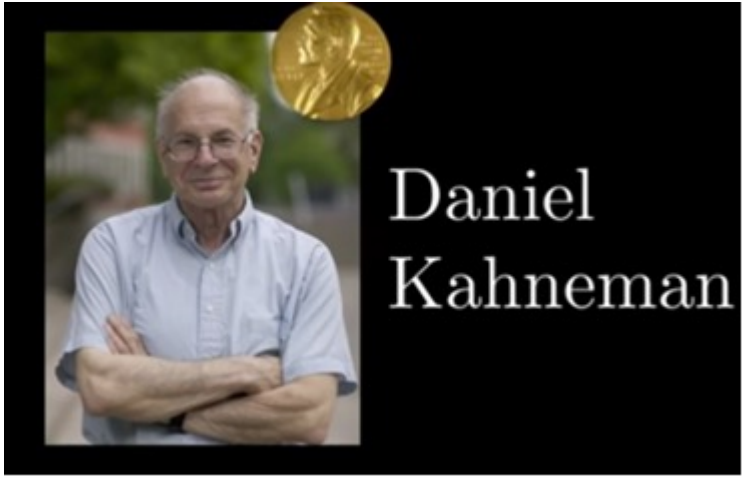
Steve is very shy and withdrawn, invariably helpful but with very little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.

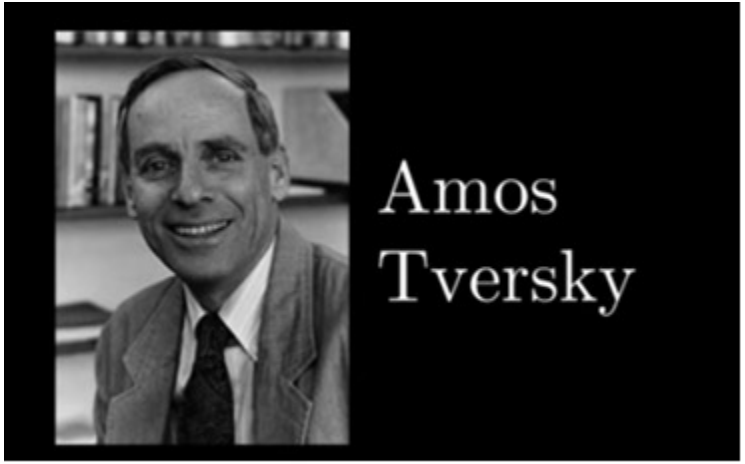
Two options:
- Steve is a librarian, or
- Steve is a farmer

What do you think?

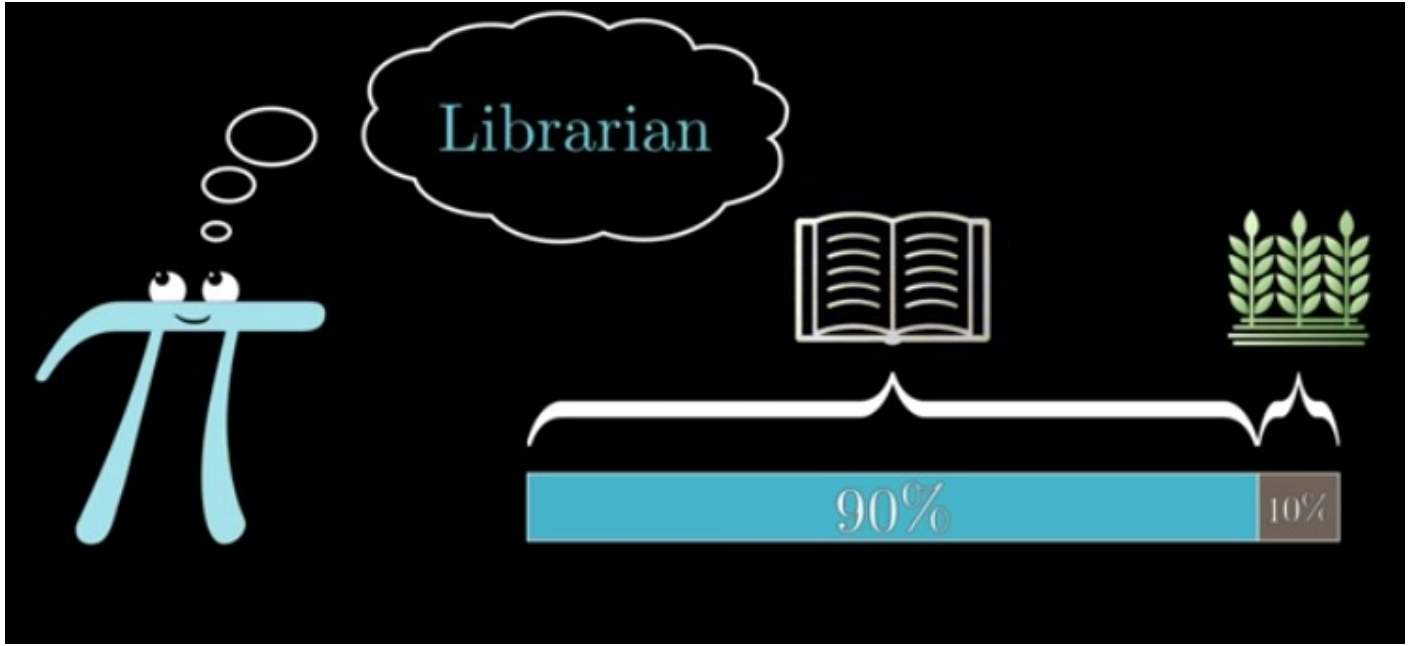
Description of Steve

# Study by Kahneman and Tversky

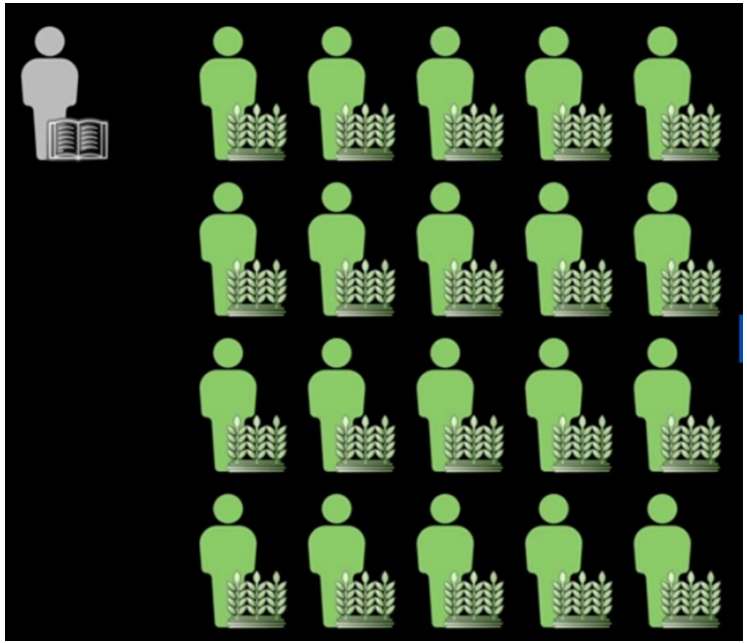


# What people in the study answered



# Background fact



There were about 20 times as many farmers than librarians in the US at that time.

Did you consider the librarian vs farmer ratio?

Most people in the study didn't.

# Background fact

| | 2017 |
|---|---|
| 441 Biblioteks- och arkivassistenter m.fl. | |
| Män | 921 |
| 611 Växtodlare inom jordbruk och trädgård | |
| Män | 13 366 |
| 612 Djuruppfödare och djurskötare | |
| Män | 3 505 |
| 613 Växtodlare och djuruppfödare, blandad drift | |
| Män | 2 934 |

There were about 20 times as many farmers than librarians in Sweden in 2017.

Librarians

Crop farmers

Animal farmers

Mixed farmers

Data source

# Is Steve a librarian or a farmer?

- Let
  - D = description (of Steve)
  - L = librarian
  - F = farmer
- We would like to know

  P(L|D) = ?

# Is Steve a librarian or a farmer?

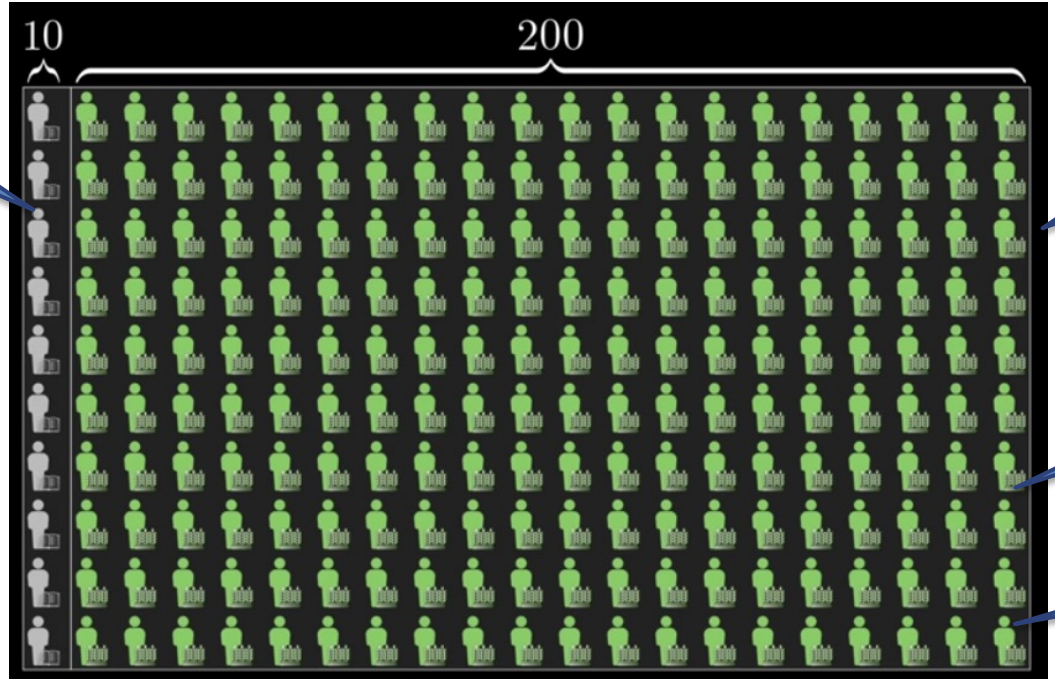
- **Bayes' rule**

$$P(L|D) = \frac{P(L)P(D|L)}{P(D)}$$

# Estimate the prior P(L)

Without using information/data D

Librarians




Visualize the entire population

Farmers

$P(L) \approx 1/21 \approx 5\%$

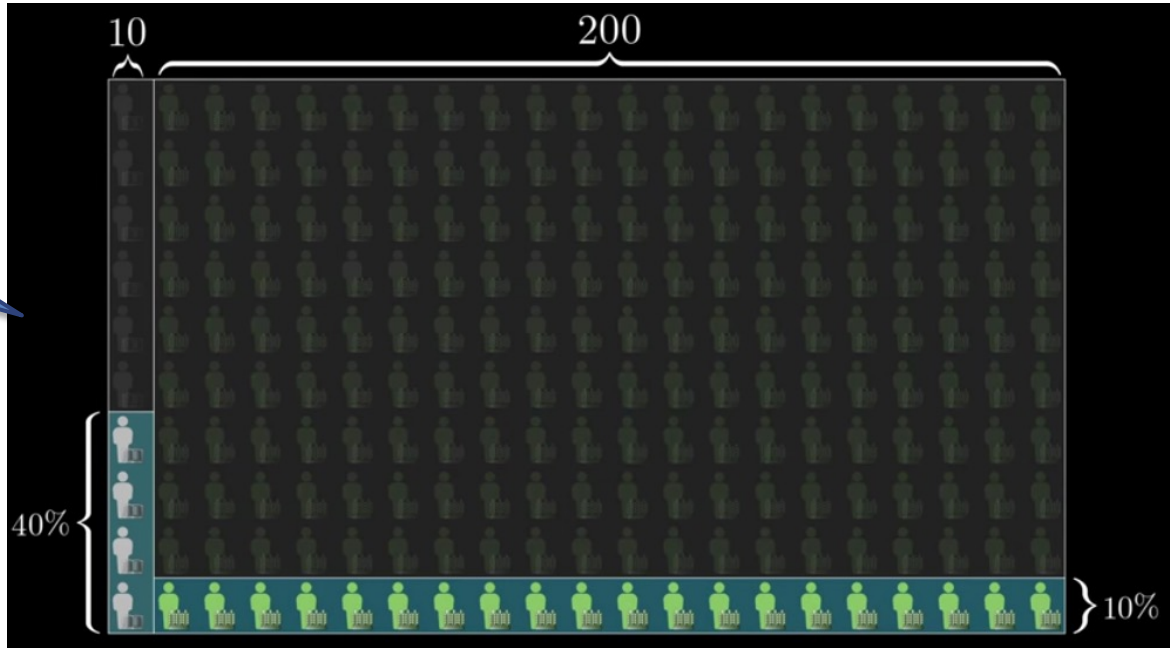
# Add the info D

Our data

Mark the individuals that match the description D

# Estimate the likelihoods P(D|L) and P(D|F)

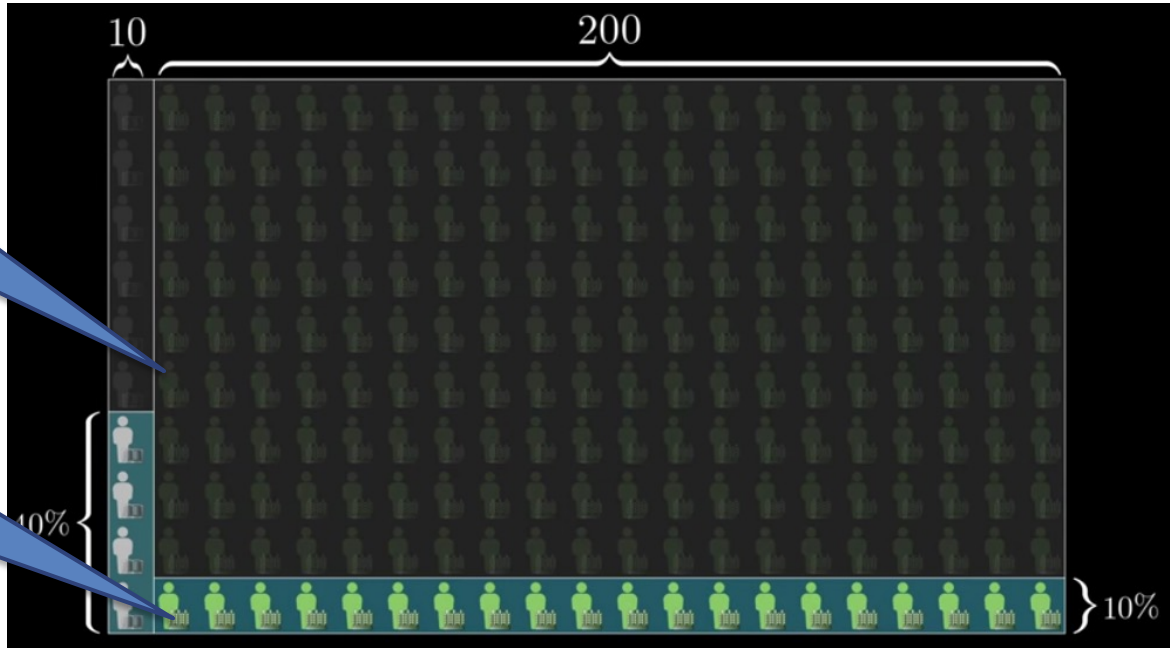
Estimates:
$P(D|L) = 40\%$
$P(D|F) = 10\%$



# Use Bayes' rule to compute the posteriors

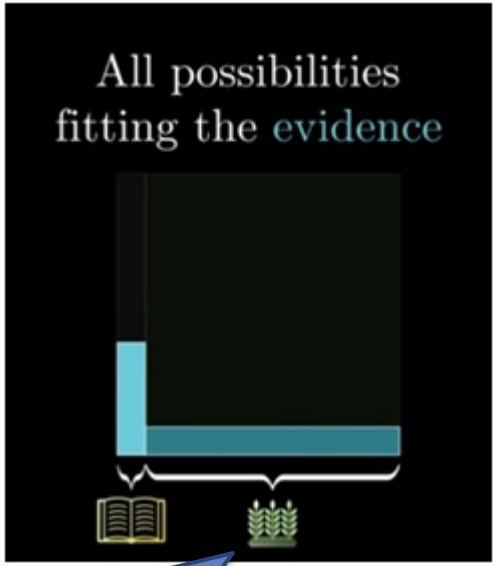
$P(L|D) = 4/(4 + 20) \approx 17\%$
$P(F|D) = 20/(4 + 20) \approx 83\%$

So it is 4 times more likely that Steve is a farmer.

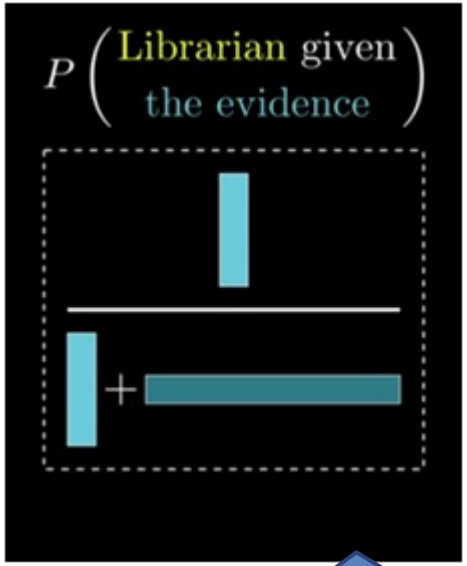


# What we did


Proportion of L and F

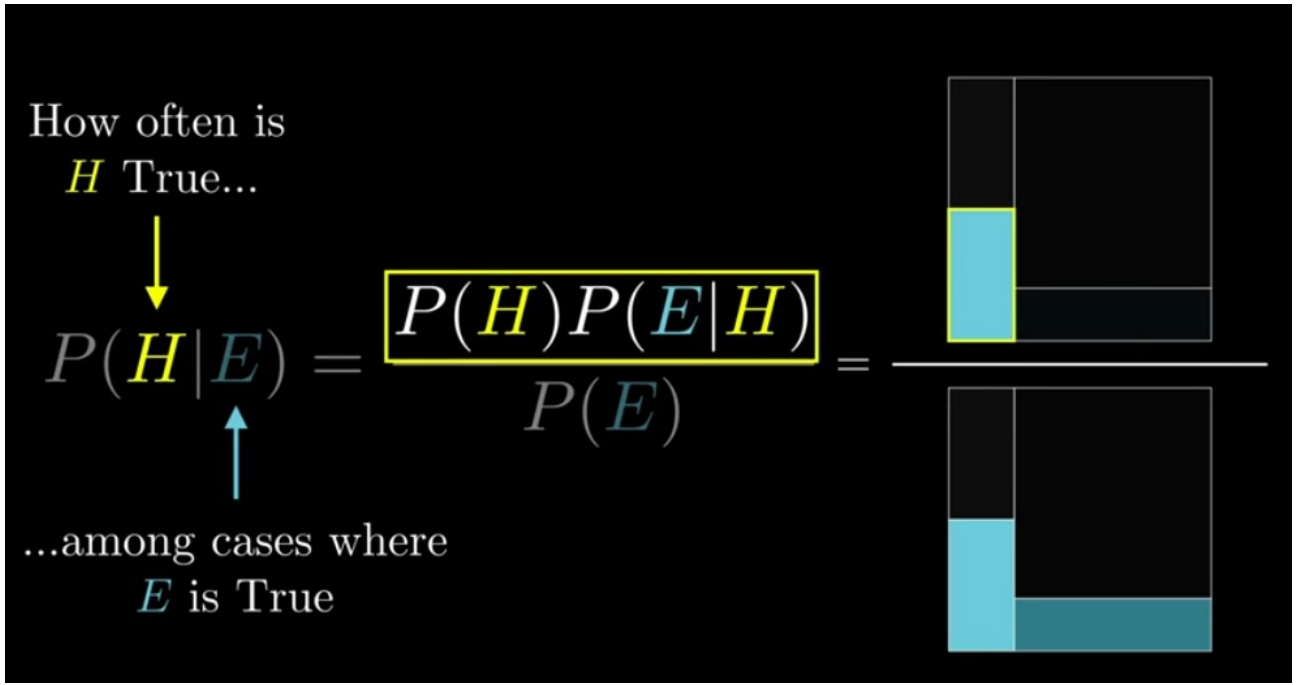



Restricting to those satisfying D

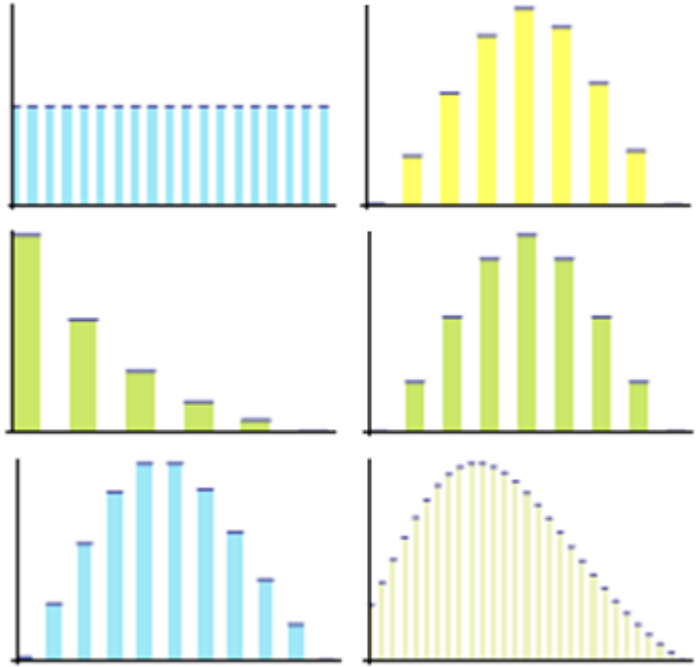



Proportion of L satisfying D

# General case in one slide



# Random variables and probability distributions



# Random variables and probability distributions

- A *random variable* is a function of the outcomes in a random experiment.
- Assumes values according to a *probability distribution*.
  - *Discrete r.v.*: finite or countable number of values,
  - *Continuous r.v*: takes all real values in given intervals

$$X: S \to \mathbb{R}$$

$$P(a \leq X \leq b) = ?$$

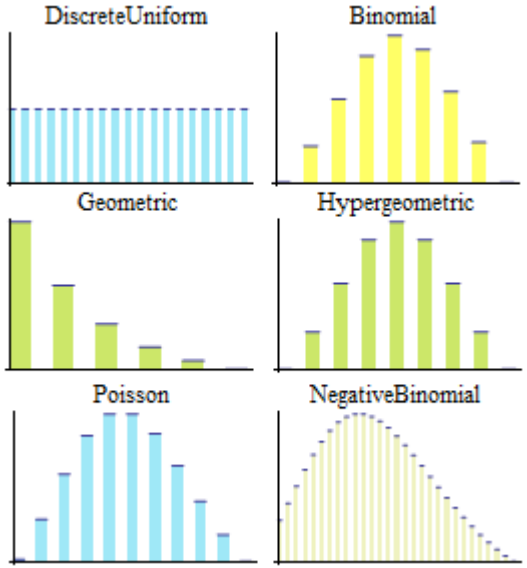
$$P(X = a) > 0$$

$$P(X = a) = 0$$

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

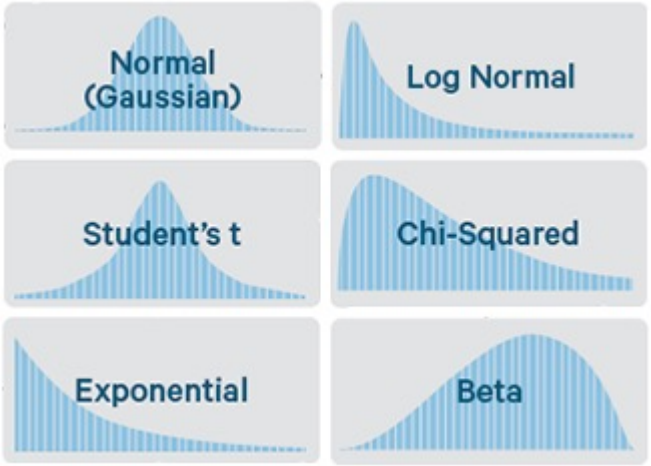
# Probability distributions

- Typically depend on one or more ***parameters***
- Common **discrete** distributions
  - ***Uniform***: $U(a, b)$
  - ***Binomial***: $Bin(n, p)$
  - ***Geometric***: $Geo(p)$
  - ***Hypergeometric***: $HGeo(N, K, n)$
  - ***Poisson***: $Poi(\lambda)$
  - ***Negative binomial***: $NB(r, p)$



# Probability distributions

- Common continuous distributions
  - *Uniform:* $U[a, b]$
  - *Normal (Gaussian)*: $N(\mu, \sigma^2)$
  - *Student's t:* $t_{n-1}$
  - *Exponential*: $Exp(\lambda)$
  - *Chi-square:* $\chi^2_{n-1}$
  - *Beta*: $Beta(\alpha, \beta)$



# Expected Value and Variance

Two key characteristics of a random variable

- **Expected value:**
  - **Mean value of random variable**
- **Variance:**
  - **Measure of how far, on average, the random variable is from its mean.**

# Expected Value and Variance

- **For a discrete random variable $X$ the expected value is the weighted average of the possible outcomes**

$$\mu = \mathbb{E}[X] = \sum_i x_i * P(X = x_i)$$

- **Intuitively measures the value you can expect to get on average in some random experiment**
  - *E.g. Rolling a fair die once: 1/6 + 2/6 + 3/6 + 4/6 + 5/6 + 6/6 = 2 ½ = 3.5.*

# Expected Value and Variance

- For a discrete random variable $\boldsymbol{X}$ the variance is the weighted average of the square distance to the mean

$$Var[\boldsymbol{X}] = \sum_i (x_i - \mu)^2 * P(X = x_i)$$

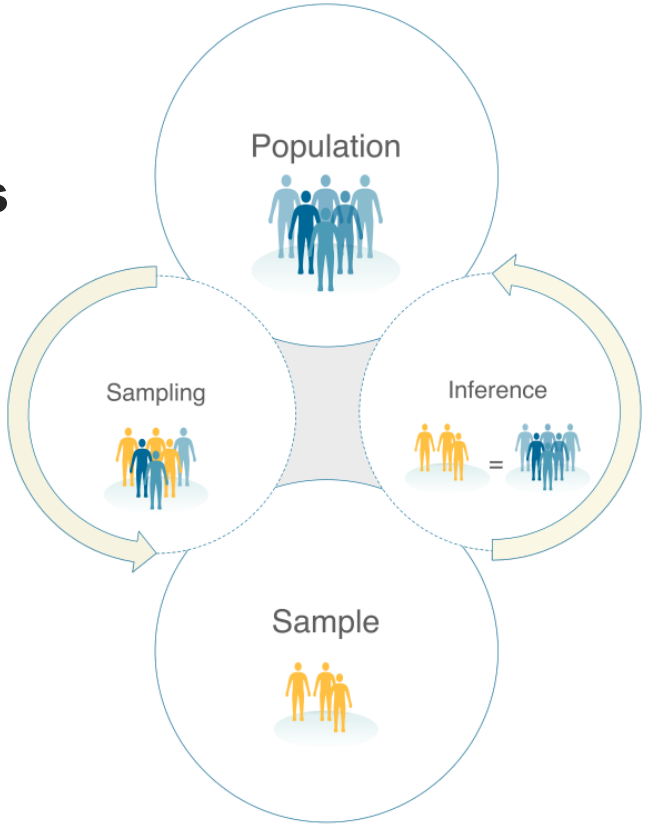
- Intuitively measures how spread out the values of the random variable are.

# Statistical inference

**Estimation and analysis of these parameters in random samples to draw conclusions of the underlying population.**

**Two main paradigms:**

- ***Frequentism***
- ***Bayesianism***



# *Classical* or *frequentist* probability theory:

- Probabilities are *relative frequencies* of the event in a large number of trials.

# *Bayesian* probability theory:

- Probabilities are *reasonable expectation* of an event, quantifying personal beliefs and prior information, and including the degree of certainty in those beliefs.

# Frequentism versus Bayesianism

| Frequentism | Bayesianism |
|---|---|
| + Objective | + More natural |
| + Trade of between errors | + Logically rigourous |
| + Design controls bias | + Can explore different priors |
| + Long prosperous history | + Data can be added |
| - p-value depends on design | - Prior is subjective |
| - Ad-hoc notions of ”data more extreme” | - Assigning probabilities to hypotheses |
| - Fully specified designs ahead | |