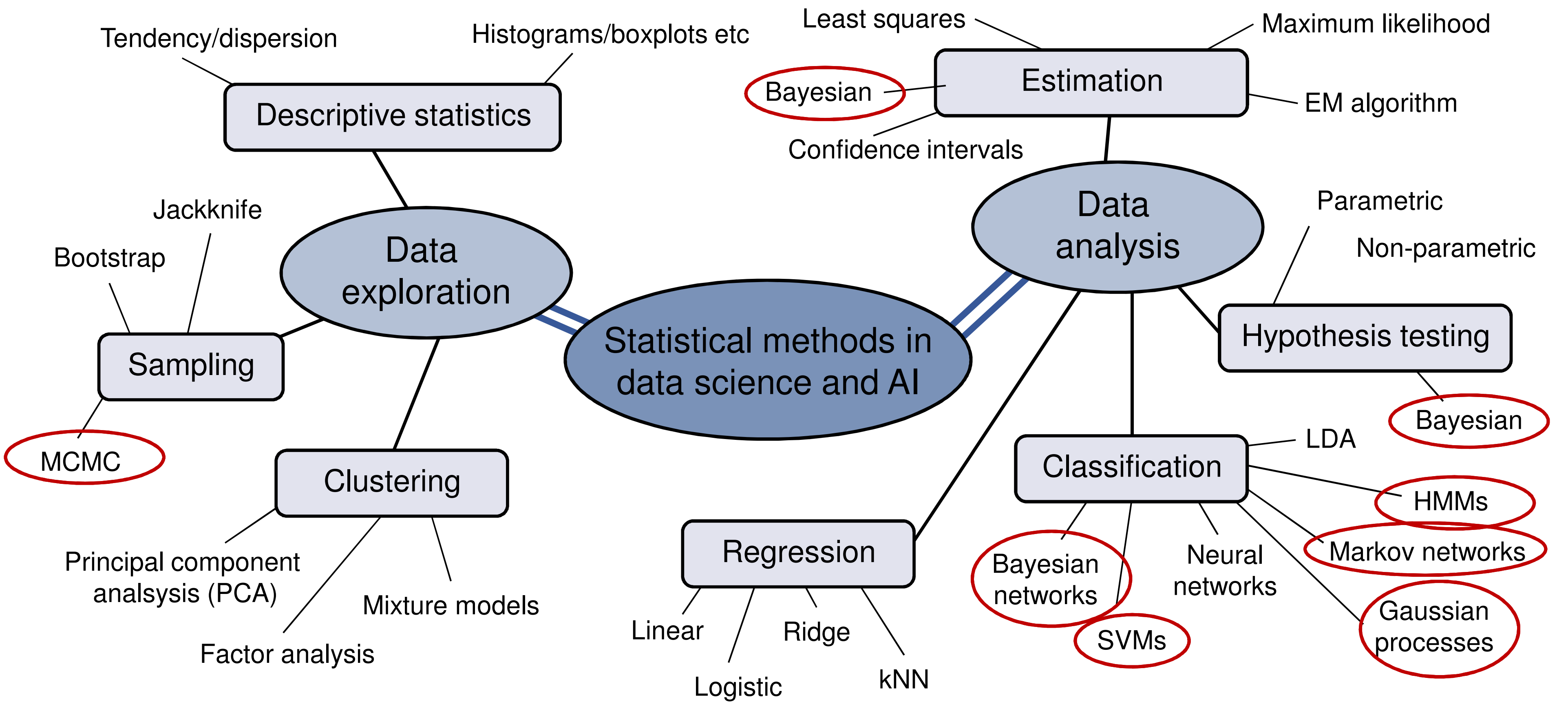
CHALMERS
UNIVERSITY OF TECHNOLOGY

Statistical methods in Data Science and AI

DAT405, 2019/2020, Module 3-4.

Marina Axelson-Fisk


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

CHALMERS
UNIVERSITY OF TECHNOLOGY

Module 3.1: Bayesian statistics

Probability theory and statistics

- a quick refresher

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

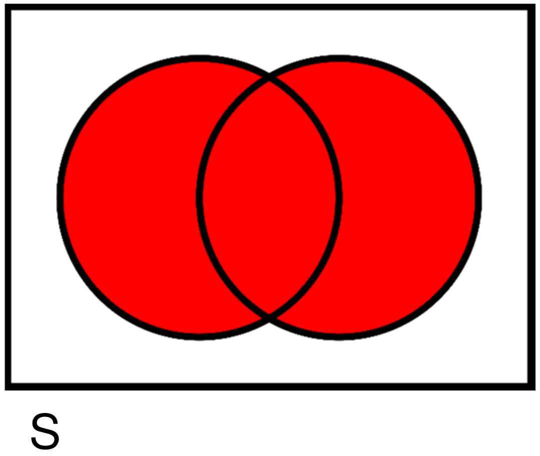
  $P(\text{at least 3}) = 4/6$

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

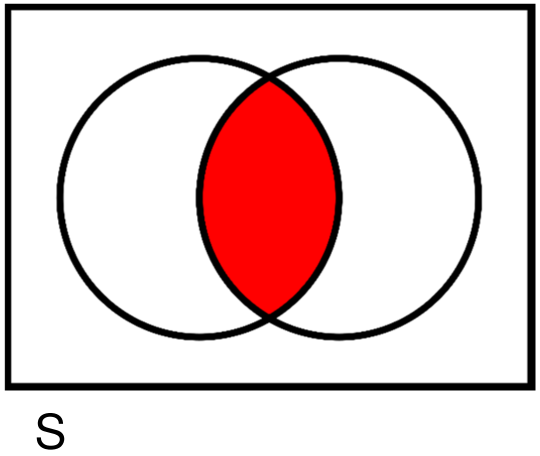
  $P(\text{odd}) = 3/6$

Venn diagrams of set operations

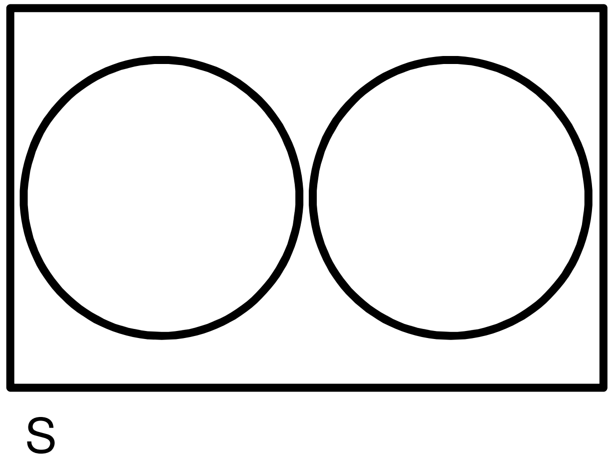
Union: $A \cup B$



Intersection: $A \cap B$



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



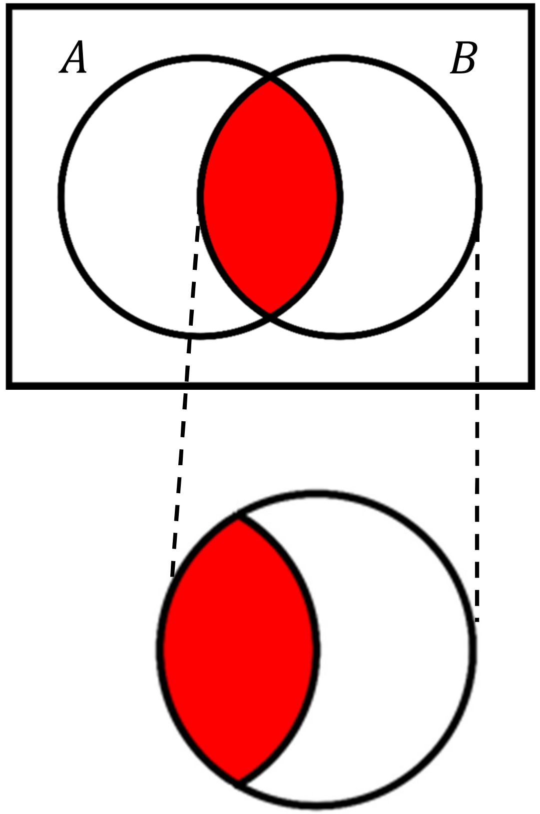
Conditional probability

- 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

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



Mutually exclusive and exhaustive events

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

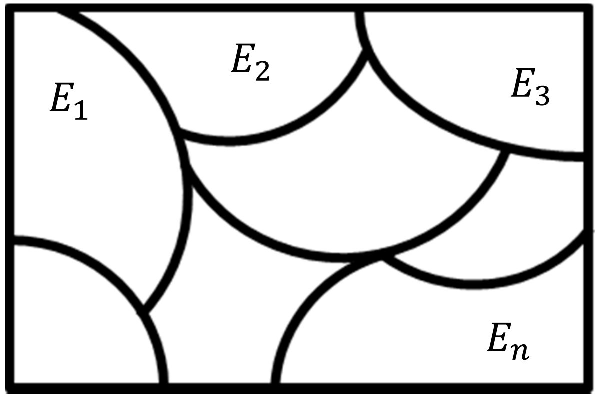
- ***mutually exclusive*** **if they cannot occur simultaneously**

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

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

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

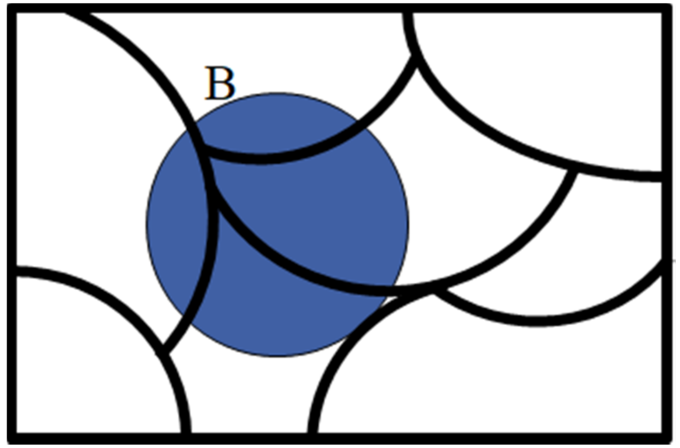
Sample space S



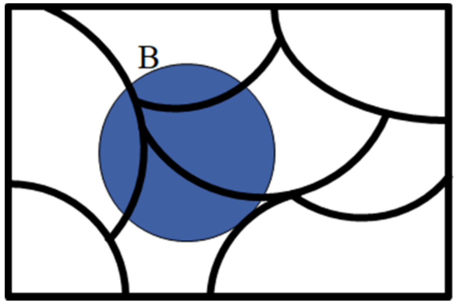
Total law of probability

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

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



Bayes' rule



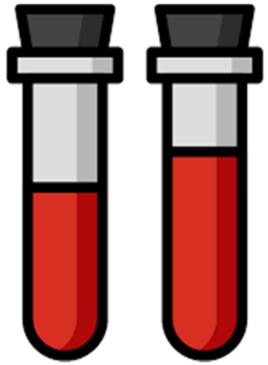
- **Bayes' rule**

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

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

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

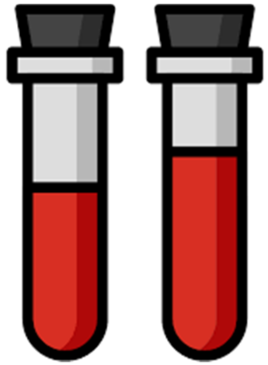
Example



- Assume that 0.0015 individuals in our population has a certain disease D.
- 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: $$\boldsymbol{P}(\text{ill}\,|+) = \frac{\boldsymbol{P}(+|\text{ill})\boldsymbol{P}(\text{ill})}{\boldsymbol{P}(+)}$$



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

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

Random variables and probability distributions

- A ***random variable*** is a function of the outcomes in a random experiment.

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

- Assumes values according to a ***probability distribution***.

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

  - *Discrete r.v.*: finite or countable number of values,

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

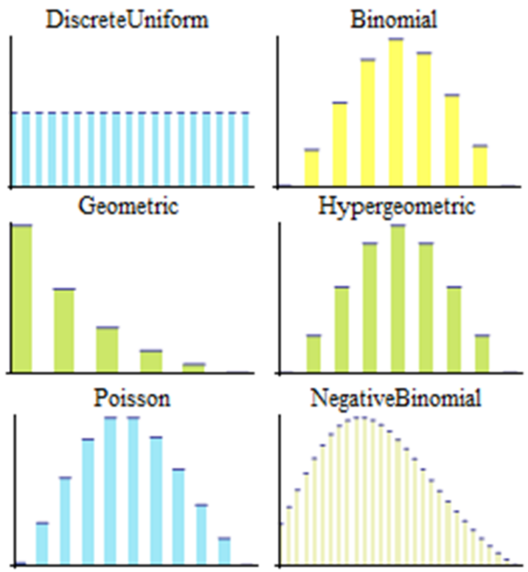
  - *Continuous r.v*: takes all real values in given intervals

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

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

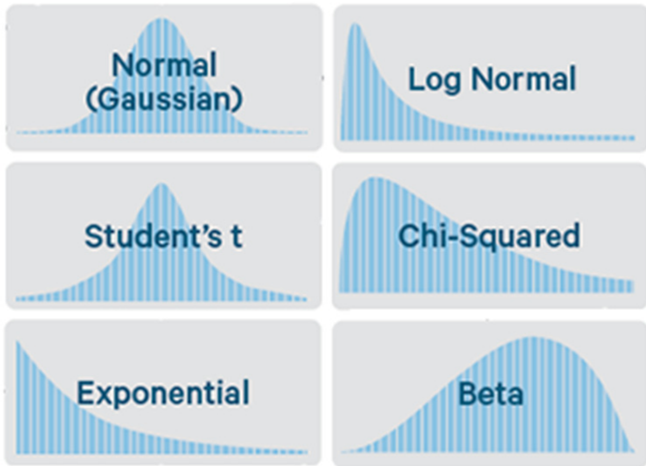
Probability distributions

- 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:* χ^2_{n-1}
  - *Beta*: $Beta(\alpha, \beta)$

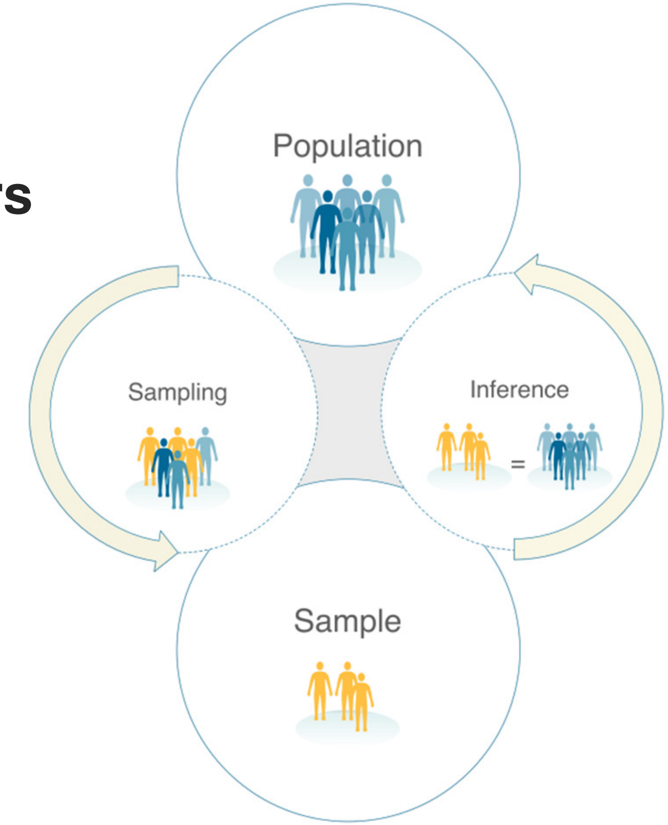


Statistical inference

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

Two main paradigms:

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

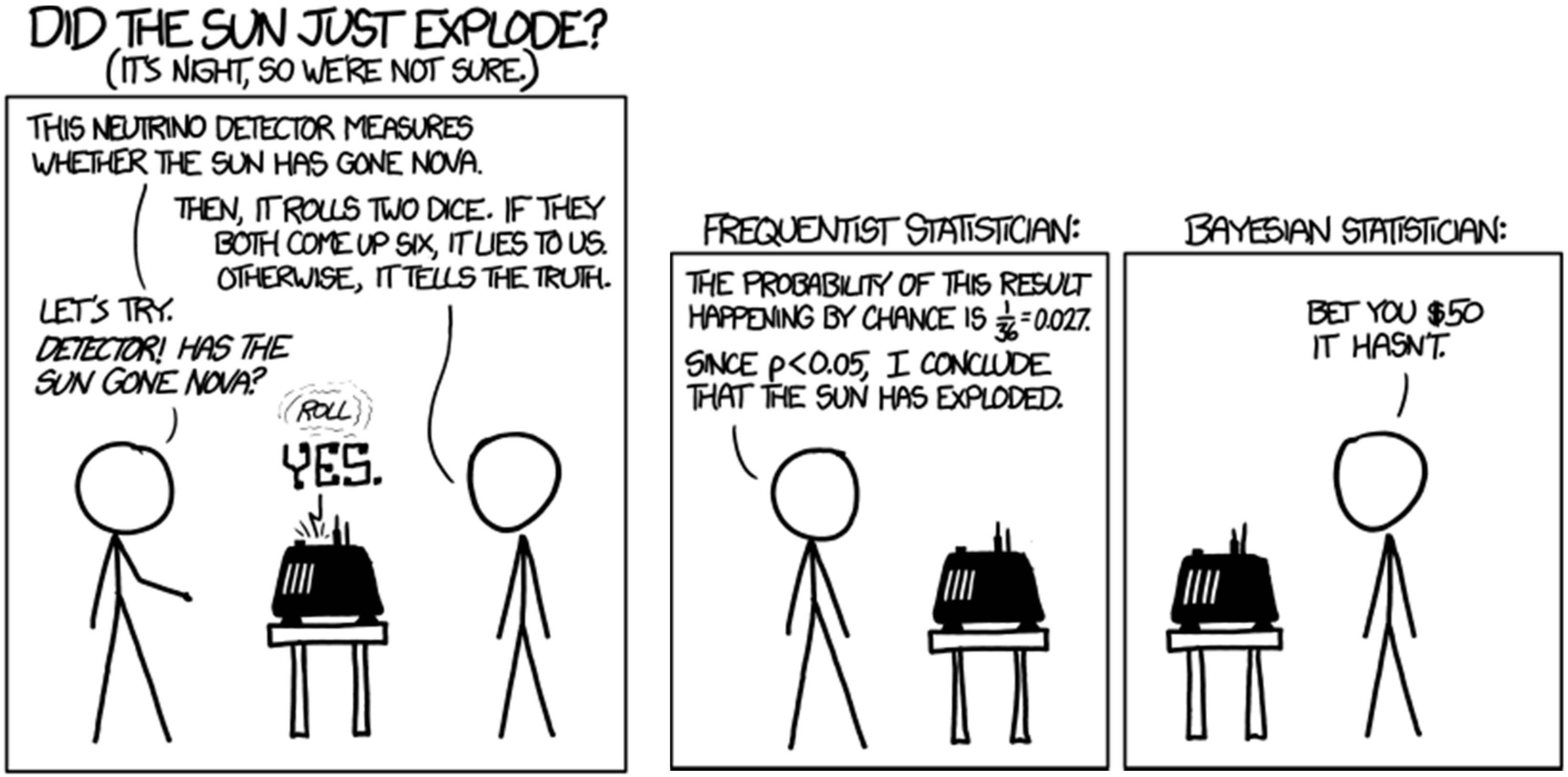


Classical or *frequentist* probability theory:

- Probabilities are *frequencies* of random repeatable experiments
- Probabilities quantify *variability*.
- Parameters are (unknown) *constants*.

Bayesian probability theory:

- Probabilities correspond to *reasonable expectation* of an event.
- Probabilities quantify *uncertainty*.
- Unknown parameters are treated as *random variables*.


DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)
THIS NEUTRINO DETECTOR MEASURES WHETHER THE SUN HAS GONE NOVA.
THEN, IT ROLLS TWO DICE. IF THEY BOTH COME UP SIX, IT LIES TO US. OTHERWISE, IT TELLS THE TRUTH.
LET'S TRY.
DETECTOR! HAS THE SUN GONE NOVA?
ROLL
YES.
FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}$ = 0.027.
SINCE $p < 0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Bayes' rule interpretation

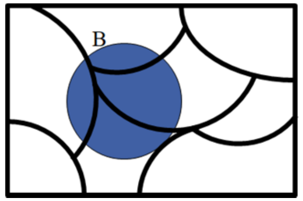
likelihood

posterior

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.

Example

Random number $p \in (0,1)$
Random numbers $q_1, q_2, q_3, \ldots$

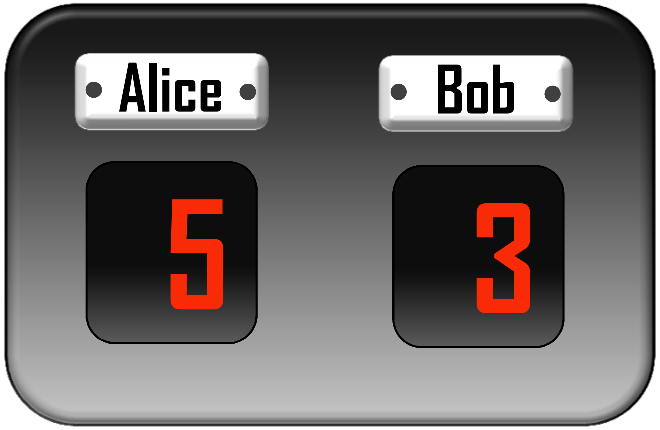
- If $q_i < p$ Alice wins
- If $q_i > p$ Bob wins

First to 6 wins the game.

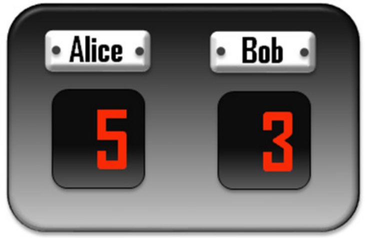
Only the scores are visible!

Example

What is the probability that Alice wins?



Example



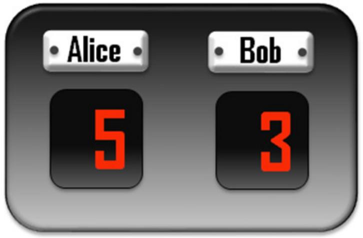
For known p:

- $P(\text{Bob}) = (1 - p)^3$
- $P(\text{Alice}) = 1 - P(\text{Bob})$

$p = 0.5 \Rightarrow P(\text{Alice}) = 7/8$

CHALMERS
UNIVERSITY OF TECHNOLOGY

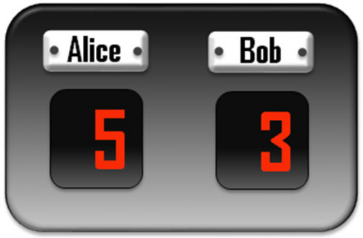
Example



Frequentists approach (ML):
$\hat{p} = 5/8 \Rightarrow P(\text{Alice}) \approx 0.95$

Fair odds: 19:1

Example



- Bayesian approach
  - Consider p a random variable.
  - Let $D = \{n_A = 5, n_B = 3\}$ denote our observed data
  - The expected probability that Bob wins is given by

$$E_B = \int_0^1 (1-p)^3 P(p|D)\, dp$$

- Bayes' rule

$$P(p|D) = \frac{P(D|p)P(p)}{P(D)} = \frac{\overbrace{P(D|p)}^{\text{likelihood}}\,\overbrace{P(p)}^{\text{prior}}}{\int_0^1 P(D|p')P(p')\,dp'}$$

Example

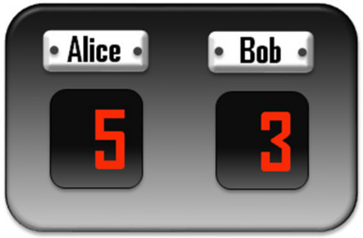
- The likelihood $P(D|p)$:
  - Let $X =$ the number of times Alice wins out of 8
  - Probability of winning $= p$

$$X \sim \mathrm{Bin}(n_A, p)$$

  - The likelihood of observing our data, given p becomes

$$P(X = 5) = \binom{8}{5} p^5 (1 - p)^3$$

Example



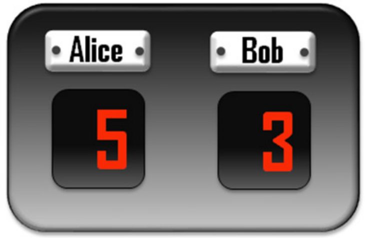
- **The prior $P(p)$:**
  - **Assume** $p \sim U(0,1) \Rightarrow P(p) = \text{constant}$

$$E_B = \int_0^1 (1-p)^3 P(D|p)dp = \frac{\int_0^1 p^5 (1-p)^6 dp}{\int_0^1 p^5 (1-p)^3 dp} = 1/11$$

$$E_A = 1 - 1/11 = 10/11$$

Beta-integral: $\int_0^1 p^{m-1}(1-p)^{n-1}dp = \frac{\Gamma(m)\Gamma(n)}{\Gamma(n+m)},\ \Gamma(n) = (n-1)!$

Example



Alice

Bob

Frequentist approach

- $P(\text{Alice}) \approx \mathbf{0.95}$
- **Fair odds: 19:1**

Bayesian approach

- $P(\text{Alice}) \approx \mathbf{0.91}$
- **Fair odds: 10:1**

Simulation confirms Bayesian computation!

Bayesianism versus frequentism

What is the probability of an event?

- **Frequentists: the *relative frequency* of the event in a large number of trials.**
- **Bayesians: a *reasonable expectation*, quantifying personal beliefs and prior knowledge, and including the degrees of certainty in these beliefs.**

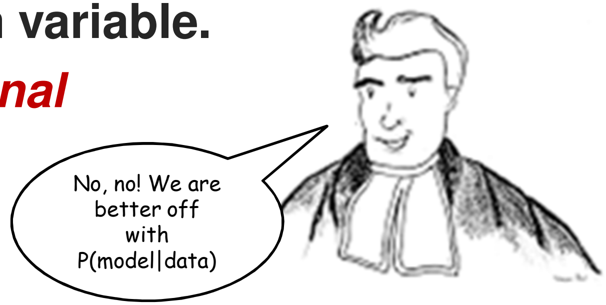
Bayesianism versus frequentism

Frequentists:

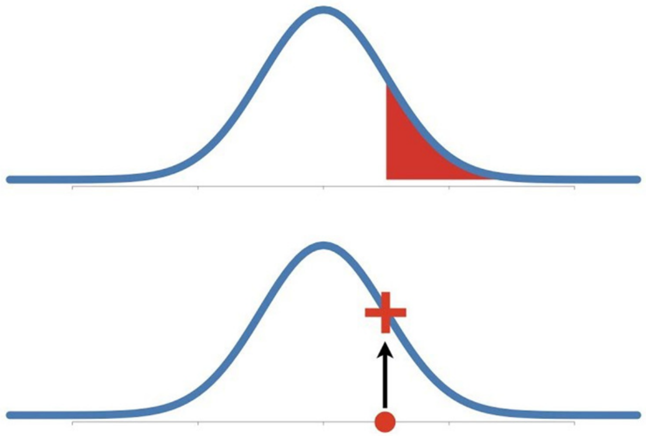
- **A distribution parameter θ is an (unknown) *constant*.**
- **$P(\theta = a) = ?$ becomes meaningless.**
- **The density of a random variable X: $f_\theta(X)$**

Bayesians:

- **An unknown parameter θ is treated as a random variable.**
- **The density of a random variable X is a *conditional probability*: $f(X|\theta)$**



The likelihood function



The ***likelihood function*** introduces a third view

- The density of X as a function of θ: $L_x(\theta)$
- Same thing, different names

$$L_\theta(x) = f_\theta(x) = f(x|\theta)$$

- But with Bayesian statistics we can use Bayes' theorem on θ

$$f(\theta|x) = \frac{f(x,\theta)}{f(x)} = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta')f(\theta')d\theta'}$$

Bayesianism versus frequentism

Frequentists:

- X is random, but θ is not.

Bayesians:

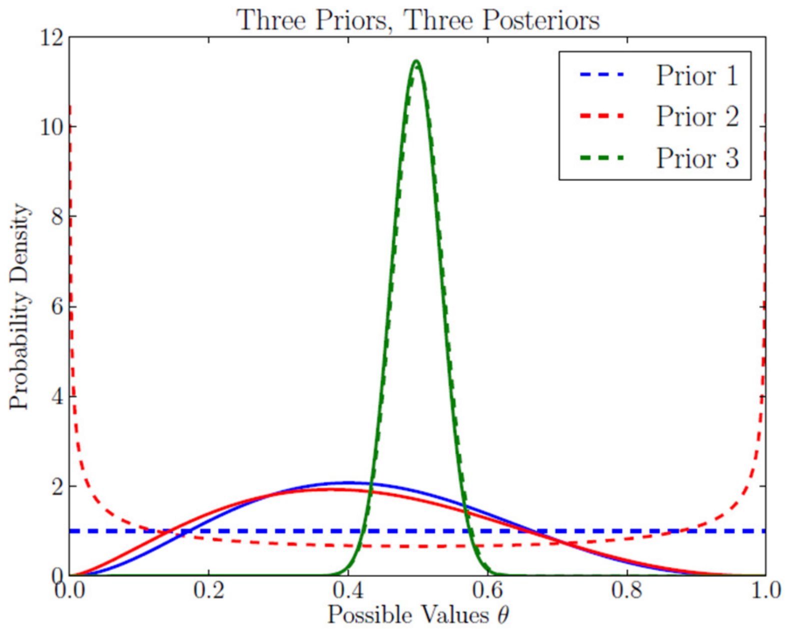
- θ is random, but after having seen data, x is not

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	

The effect of different priors

Bayesian feature: different priors will give different posteriors



Example

- Alice has moved to a new city
- She takes the bus to work
- Out of 5 attempts:
  - 2 got her to the right place
  - 3 forced her to walk another 20 min

What is the proportion of ”good” buses for her to take?

Example

Let θ = the fraction of ”good” buses.

- Prior $f(\theta)$: $\mathrm{Uniform}(0, 1)$

Let $X =$ the number of good buses of n

- Likelihood $f(x|\theta)$: $\mathrm{Bin}(n, \theta)$

Observed data:

- $\hat{\theta} = 2/5 = 0.4$

Parameter update, given observed data

- Posterior $\propto$ likelihood $\times$ prior
- $f(\theta|x) \propto f(x|\theta)f(\theta)$

Example

Assume for simplicity $\theta \in \{0.0, 0.1, 0.2, \dots, 0.9, 1.0\} = 11$ values

θ-values	prior	likelihood	prior × likelihood	posterior
0	0.0909	0	0	0
0.1	0.0909	0.0729	0.0066	0.0437
0.2	0.0909	0.2048	0.0186	0.1229
0.3	0.0909	0.3087	0.0281	0.1852
0.4	0.0909	0.3456	0.0314	0.2074
0.5	0.0909	0.3125	0.0284	0.1875
0.6	0.0909	0.2304	0.0209	0.1383
0.7	0.0909	0.1323	0.0120	0.0794
0.8	0.0909	0.0512	0.0047	0.0307
0.9	0.0909	0.0081	0.0007	0.0049
1	0.0909	0	0	0
Totals:	1		0.1515	1

Example

We can predict new values

$$\boldsymbol{P}(\text{good bus tomorrow}|\boldsymbol{x}) =$$

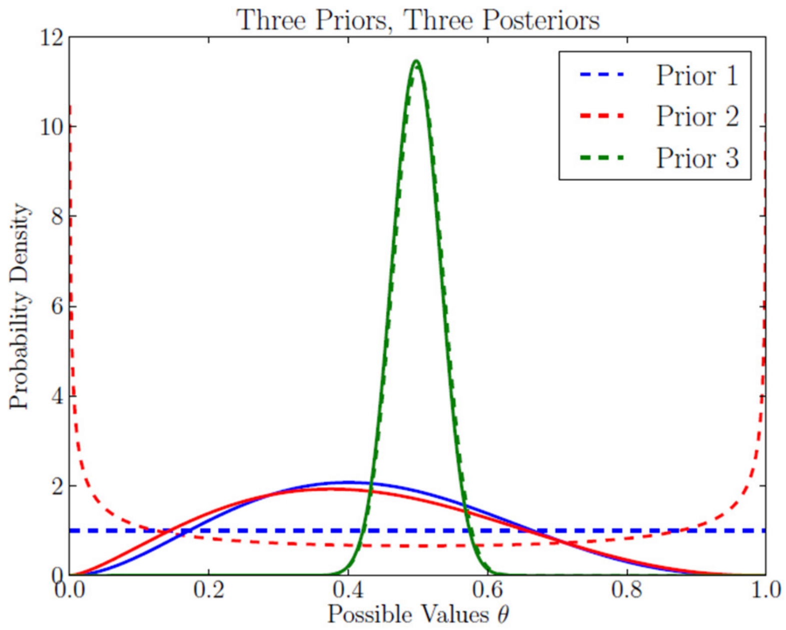
$$= \sum_{\boldsymbol{\theta}} \boldsymbol{P}(\text{good bus tomorrow}|\boldsymbol{\theta}, \boldsymbol{x})\boldsymbol{p}(\boldsymbol{\theta}|\boldsymbol{x})$$

$$= \sum_{\boldsymbol{\theta}} \boldsymbol{\theta} \cdot \boldsymbol{p}(\boldsymbol{\theta}|\boldsymbol{x})$$

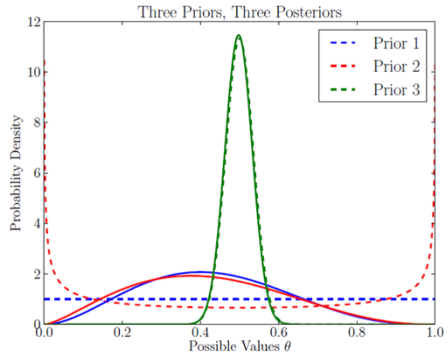
$$= \mathbf{0.429}$$

The effect of different priors

- **Prior 1: $U(0, 1)$**
  - $p(\theta) = \text{const}$
- **Prior 2:**
  - $p(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$
  - **more weight on extreme values**
- **Prior 3:**
  - $p(\theta) \propto \theta^{100}(1-\theta)^{100}$
  - **most weight in the centre $\theta = 0.5$**

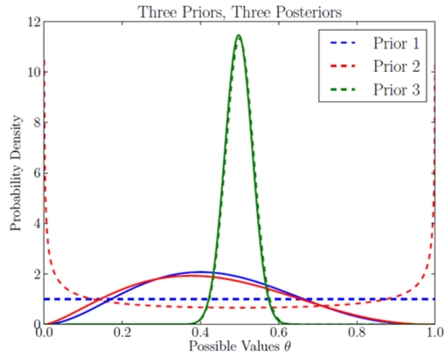


The effect of different priors



- **Prior 1:** $\boldsymbol{U(0,1)}$ $\sim \text{Beta}(\mathbf{1},\mathbf{1})$
  - $\boldsymbol{p(\theta)} = \text{const}$
- **Prior 2:** $\sim \text{Beta}\left(\frac{\mathbf{1}}{\mathbf{2}},\frac{\mathbf{1}}{\mathbf{2}}\right)$
  - $\boldsymbol{p(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}}$
  - **more weight on extreme values**
- **Prior 3:** $\sim \text{Beta}(\mathbf{101},\mathbf{101})$
  - $\boldsymbol{p(\theta) \propto \theta^{100}(1-\theta)^{100}}$
  - **most weight around** $\boldsymbol{\theta = 0.5}$

The effect of different priors



- **Prior 1:** $\mathrm{Beta}(\mathbf{1},\mathbf{1})$
- **Prior 2:** $\mathrm{Beta}\left(\frac{\mathbf{1}}{\mathbf{2}},\frac{\mathbf{1}}{\mathbf{2}}\right)$
- **Prior 3:** $\mathrm{Beta}(\mathbf{101},\mathbf{101})$

- **Posterior 1:** $\mathrm{Beta}(\mathbf{3},\mathbf{4})$
- **Posterior 2:** $\mathrm{Beta}(\mathbf{2.5},\mathbf{2.5})$
- **Posterior 3:** $\mathrm{Beta}(\mathbf{103},\mathbf{104})$

Beta-prior + binomial likelihood ⇨ Beta-posterior

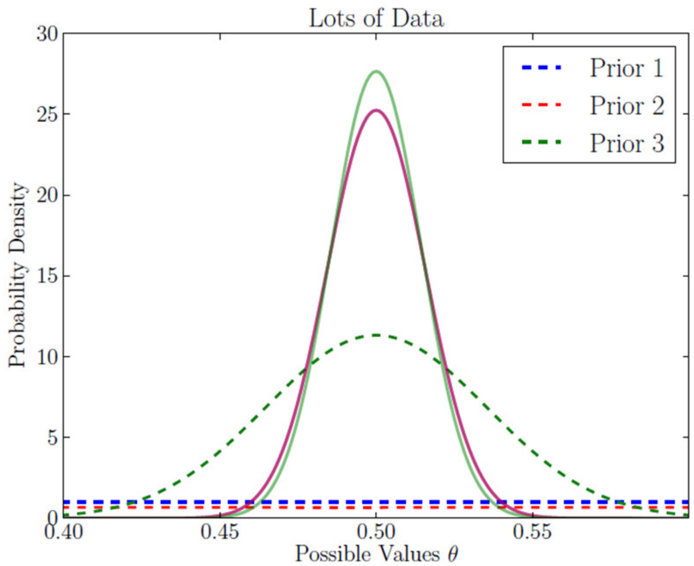
$\mathrm{Beta}(\boldsymbol{\alpha},\boldsymbol{\beta})$ **+ ”$\boldsymbol{x}$ of $\boldsymbol{n}$ successes”** ⇨ $\mathrm{Beta}(\boldsymbol{\alpha}+\boldsymbol{x},\boldsymbol{\beta}+\boldsymbol{n})$

Example

- **Prior 1:** $P(\text{good bus tomorrow}|\boldsymbol{x}) \approx \mathbf{0.429}$
- **Prior 2:** $P(\text{good bus tomorrow}|\boldsymbol{x}) \approx \mathbf{0.417}$
- **Prior 3:** $P(\text{good bus tomorrow}|\boldsymbol{x}) \approx \mathbf{0.498}$

The effect of different priors

- **The more data, the less important the prior**



Conjugate priors

- We have a sample of observed data: $x_1, \dots, x_n$
- We have a corresponding likelihood function (or *samling distribution*): $f(x|\theta)$
- A prior $f(\theta)$ is called a *conjugate prior* if the corresponding posterior $f(\theta|x)$ belongs to the same family of distributions.

Bayes' theorem:

$$\boldsymbol{f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)}}$$

Conjugate priors

Likelihood	Parameter	Prior	Posterior
Bernoulli	p	Beta	Beta
Binomial			
Geometric			
Negative binomial			
Exponential	λ	Gamma	Gamma
$\text{Gamma}(\alpha, \beta), \alpha$ known	β		
Poisson	λ		
$\mathcal{N}(\mu, \sigma^2)$, μ known	σ^2		
$\mathcal{N}(\mu, \sigma^2)$, σ^2 known	μ	Normal	Normal
Multinomial	$p_1, \ldots, p_K$	Dirichlet	Dirichlet

Conjugate priors

- **Conjugacy is mutual, e.g.**

$$\text{Dirichlet} \propto \text{Multinomial} \times \text{Dirichlet}$$

$$\text{Multinomial} \propto \text{Dirichlet} \times \text{Multinomial}$$

Bayes' theorem:

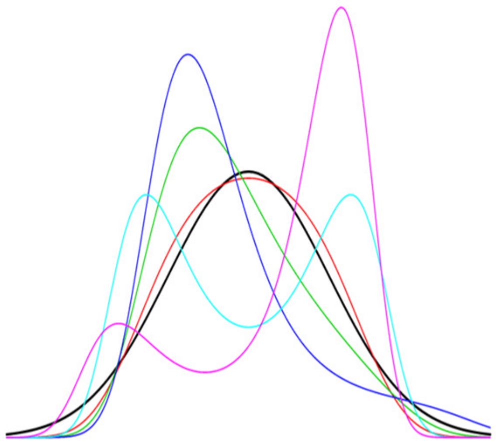
$$\boldsymbol{f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)}}$$

The exponential family of distributions

- The ***exponential family*** of distributions over x, given parameters η, takes the form

$$f(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}u(\boldsymbol{x})\right\}$$

- The function $u(x)$ is called a ***sufficient statistic*** for η, i.e. it contains all information needed to estimate η.
- All members of the exponential family has conjugate priors.
- **Products** of exponential family members also have conjugate priors.



Example: the Bernoulli distribution

$$\boxed{f(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}u(\boldsymbol{x})\right\}}$$

$$\begin{aligned}
f(x|p) &= p^x(1-p)^{1-x} \\
&= \exp\{\ln(p^x(1-p)^{1-x}\} \\
&= \exp\{x\ln p + (1-x)\ln(1-p)\} \\
&= \exp\left\{x\ln\left(\frac{p}{1-p}\right) + \ln(1-p)\right\} \\
&= \exp\{x\eta - \ln(1+e^{\eta})\} \\
&= \left(\frac{1}{1+\exp(\eta)}\right)\exp\{x\eta\}
\end{aligned}$$

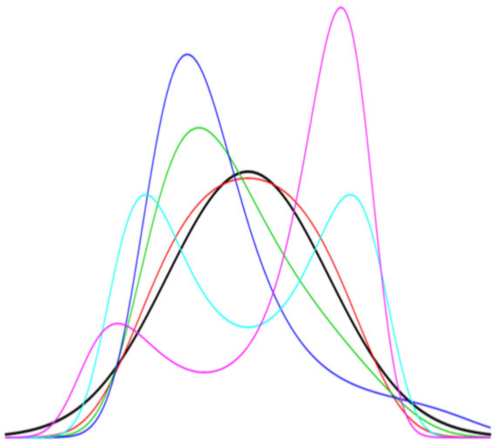
substitute $\left[\eta = \ln\left(\frac{p}{1-p}\right)\right]$

Examples of exponential family members

- Bernoulli
- Geometric
- Gamma
- Exponential
- Poisson
- Beta
- Normal
- Beta
- Dirichlet
- Chi-squared

Also:

- **Binomial**, with fixed number of trials
- **Multinomial**, with fixed number of trials
- **Negative binomial**, with fixed number of failures



Uninformative priors

- When nothing is known, we may want to play equal weights to all parameter values

  ⇒ Uniform distribution

+ Gives the same parameter estimate as Maximum Likelihood

- Not invariant under parameterization

$$X \sim U[a, b], Y = f(X) \nRightarrow Y \sim U[f(a), f(b)]$$

⇨ Large variation in posterior

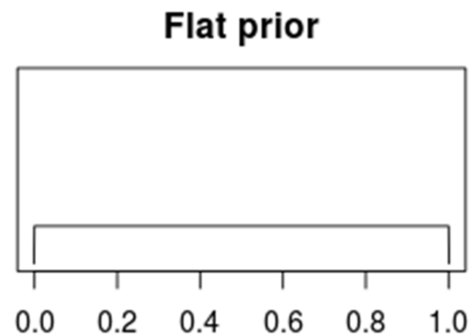
Jeffrey's prior

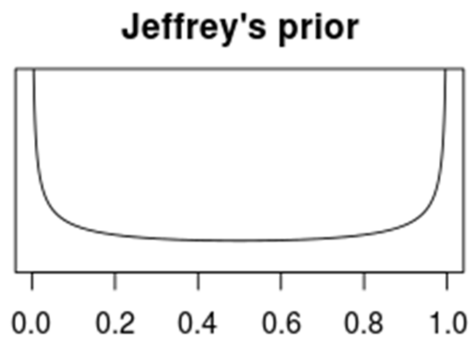
- **Uninformative prior**
- **Invariant under transformation**
- **Given by**

$$p(\boldsymbol{\theta}) \propto \sqrt{\det\left(\mathcal{I}(\boldsymbol{\theta})\right)}$$

where $\mathcal{I}(\boldsymbol{\theta})$ is the *Fisher information*

$$\mathcal{I}(\boldsymbol{\theta}) = -\boldsymbol{E}_{\boldsymbol{\theta}}\left[\frac{\boldsymbol{d}^2 \log \boldsymbol{f}(\boldsymbol{X}|\boldsymbol{\theta})}{\boldsymbol{d}\boldsymbol{\theta}^2}\right]$$





Fisher information

- For a random variable X with density $f(x|\theta)$:

The Fisher information =

= ”information content of X in terms of estimating θ”

Example: Jeffrey's prior

Let $X \sim \mathrm{Bin}(n, p)$. We want a prior for p.

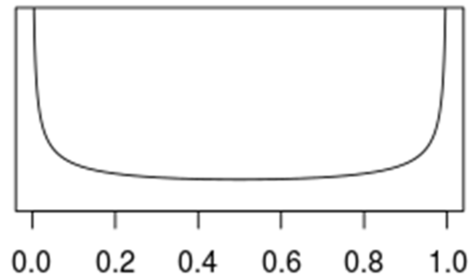
$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\log f(x|p) = x \log p + (n-x) \log(1-p)$$

$$\frac{d}{dp} \log f(x|p) = \frac{x}{p} - \frac{n-x}{1-p}$$

$$\frac{d^2}{dp^2} \log f(x|p) = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}$$

$$\mathcal{I}(p) = -E_p\left[\frac{d^2}{dp^2} \log f(x|p)\right] = -\frac{np}{p^2} - \frac{n-np}{(1-p)^2} = \frac{n}{p(1-p)}$$

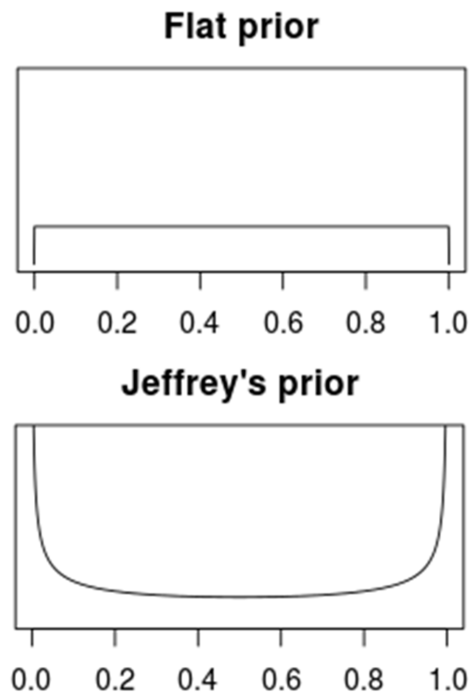


$$\boxed{E_p[X] = np}$$

Example: Jeffrey's prior

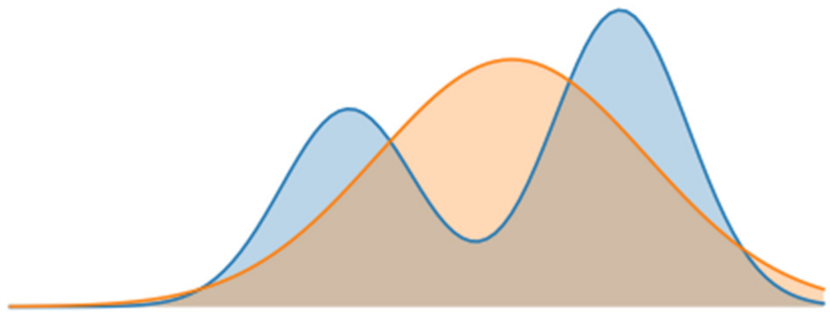
$$f(p) \propto \sqrt{\mathcal{J}(p)} \propto p^{-1/2}(1-p)^{-1/2} \sim \mathrm{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$$

Note: Jeffrey's prior is generally not conjugate.



Reference priors

- Maximize the distance between the prior and the posterior
  - Kullback-Leibler divergence
  - Hellinger distance



Inference in classical statistic

Based on a sample $x_1, \dots, x_n$ from some density $f_\theta(x)$.

Parameter estimation

Estimate the parameter θ using Maximum Likelihood

$$\hat{\theta} = \mathrm{argmax}_\theta \log L_\theta(x_1, \dots, x_n)$$

Confidence intervals

A 95% confidence interval for θ is an interval $\left(\theta^{\mathrm{lo}}, \theta^{\mathrm{up}}\right)$ such that

$$P\left(\theta^{\mathrm{lo}} \leq \theta \leq \theta^{\mathrm{up}}\right) = 0.95.$$

Note: θ^{lo} and θ^{up} are random variables, *not* θ.

Inference in classical statistic

Based on a sample $x_1, \dots, x_n$ from some density $f_\theta(x)$.

Hypothesis testing

We want to test the hypothesis

$$H_0: \theta = \theta_0$$
$$H_1: \theta \neq \theta_0$$

using som test statistic T (function of the sample). We reject H_0 on 5% significance level if $\hat{\theta} \geq \theta^{\text{up}}$ or $\hat{\theta} \leq \theta^{\text{lo}}$ where again

$$P\left(\theta^{\text{lo}} \leq \theta \leq \theta^{\text{up}}\right) = 0.95$$

Bayesian inference: parameter estimation

Based on a sample $D = \{x_1, \ldots, x_n\}$ from some density $f(x|\theta)$.

Parameter estimation

The most likely estimate of θ is the maximum of the posterior

$$\hat{\theta} = \mathrm{argmax}_\theta\, P(\theta|D)$$

Note: if the prior $\mathrm{P}(\theta)$ is the uniform distribution, then this estimate is the same as the Maximum Likelihood estimate.

Bayesian inference: credible intervals

Based on a sample $D = \{x_1, \dots, x_n\}$ from some density $f(x|\theta)$.

Credible intervals

A 95% *credible* interval for θ is an interval $\left(\theta^{\text{lo}}, \theta^{\text{up}}\right)$ such that the posterior

$$P\left(\theta^{\text{lo}} \leq \theta \leq \theta^{\text{up}} | D\right) = 0.95.$$

$$\int_{\theta^{\text{lo}}}^{\theta^{\text{up}}} p(\theta|D) d\theta = 0.95$$

Note: Now θ is a random variable with prior $\mathrm{P}(\theta)$.

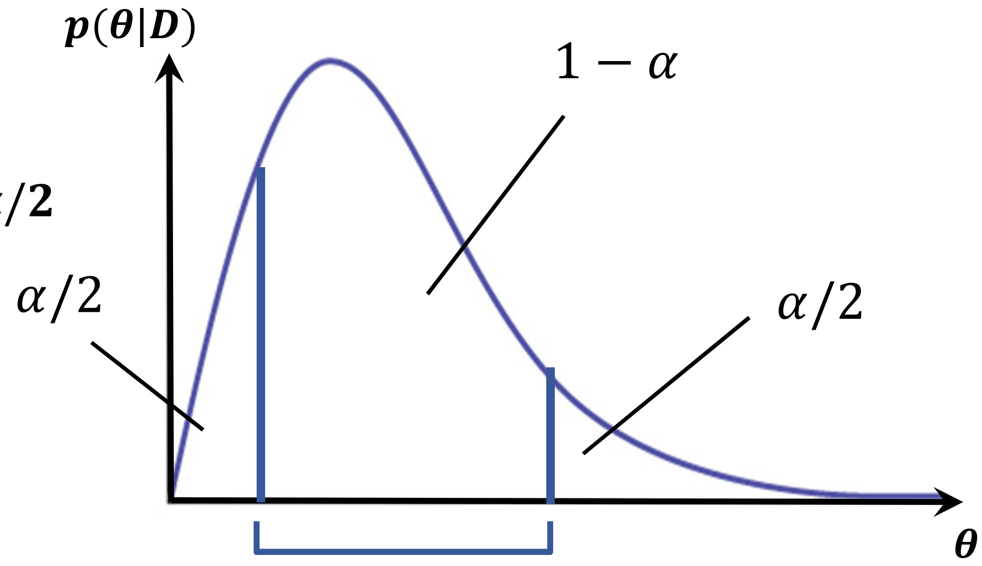
Bayesian inference: credible intervals

However, the credible interval is not unique. We need additional conditions.

For an $(1-\alpha)$-interval

- **Equal-tailed interval (ETI)**

$$P\left(\theta \le \theta^{\mathrm{lo}} \middle| D\right) = P\left(\theta \ge \theta^{\mathrm{up}}\right) = \alpha/2$$



Bayesian inference: credible intervals

However, the credible interval is not unique. We need additional conditions.

For an $(1-\alpha)$-interval

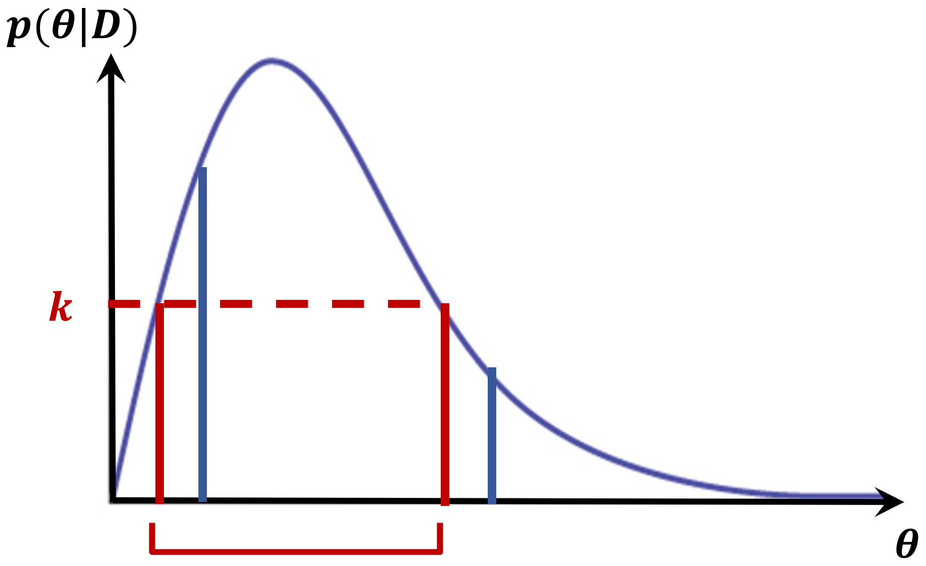
- **Equal-tailed interval (ETI)**

$$P\left(\theta \leq \theta^{\mathrm{lo}} \middle| D\right) = P\left(\theta \geq \theta^{\mathrm{up}}\right) = \alpha/2$$

- **Highest density interval (HDI)**

$\mathcal{C} = \{\theta : p(\theta|D) \geq k\}$ where

$$\int_{\theta : p(\theta|D) \geq k} p(\theta|D)\, d\theta = 1 - \alpha$$



Confidence intervals vs credible intervals

- A 95% *credible interval* contains the true value $\boldsymbol{\theta}$ with probability 95%.
    - i.e. based on data there is a 95% probability that the interval contains $\boldsymbol{\theta}$
    - Statement *after* data is collected
- A 95% *confidence interval* contains the true value of $\boldsymbol{\theta}$ 95% of the time.
    - i.e. 95% of the samples we draw will cover the true value of $\boldsymbol{\theta}$
    - Statement *before* data is collected

Bayesian inference: hypothesis testing

Bayes factor

We want to test the hypothesis

$$H_0: \theta = \theta_0$$
$$H_1: \theta \neq \theta_0$$

The Bayes factor is the ratio of the posteriors

$$\frac{P(H_1|D)}{P(H_0|D)} = \boxed{\frac{P(D|H_1)}{P(D|H_0)}} \cdot \frac{P(H_1)}{P(H_0)}$$

Bayes factor

Bayes factor	Evidence against H_0
1-3	Very weak
3-20	Positive
20-150	Strong
> 150	Very strong

Summary

- Bayesianism versus frequentism
- The choice of priors
  - Conjugate priors
  - Uninformative priors
  - Jeffrey's prior
  - Reference priors
- Exponential family
- Frequentist versus Bayesian inference
  - Parameter estimation
  - Confidence intervals – credible intervals
  - Hypothesis testing – Bayes factor