Inference of Probability Distributions for Trust and Security applications

Motivations

- Inferring the probability distribution of a random variable
- Examples of applications in trust & security
  - How much we can trust an individual or a set of individuals
  - Input distribution in a noisy channel to compute the Bayes risk
  - Application of the Bayesian approach to hypothesis testing (anonymity, information flow)
  - ...

Setting and assumptions

- For simplicity we consider only binary random variables
  - honest/dishonest, secure/insecure, ...
- Goal: infer (an approximation of) the probability of success
- Means: Sequence of *n* trials. Observation (*Evidence*) : *s* , *f*

$$X = \{succ, fail\}$$

$$Pr(succ) = \theta$$

$$s = \#succ$$
$$f = \#fail = n - s$$

Using the evidence to infer θ

- The Frequentist method:

$$F(n, s) = \frac{s}{n}$$

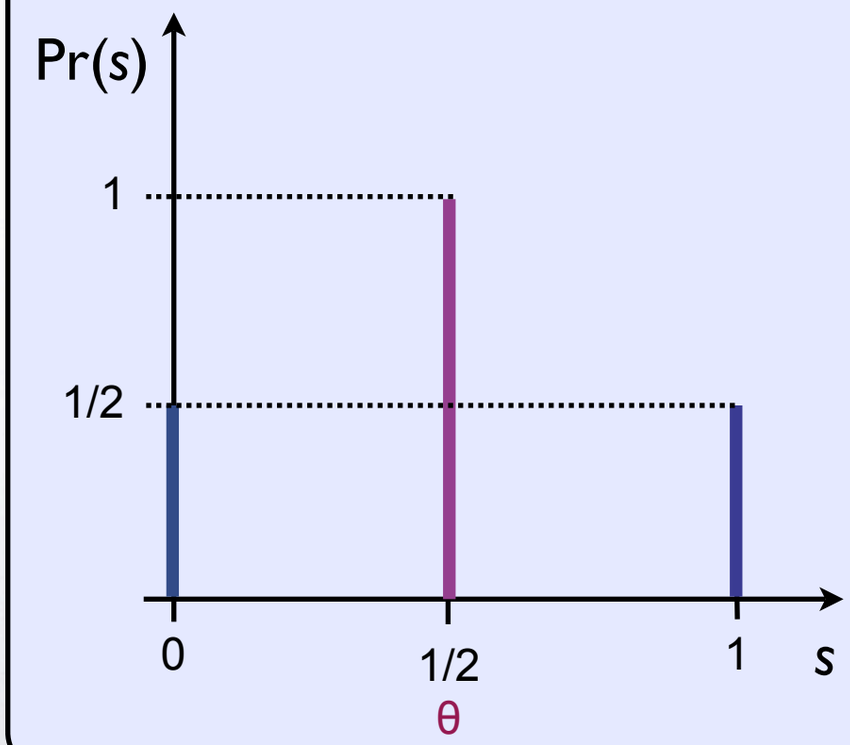
- The Bayesian method:

Assume an *a priori* probability distribution for θ (representing your partial knowledge about θ, whatever the source may be) and combine it with the *evidence*, using Bayes' theorem, to obtain the *a posteriori* distribution

Bayesian vs Frequentist

The surprising thing is that the Frequentist approach can be worse than the Bayesian approach even when the trials give a "good" result, or when we consider the average difference (from the "true" θ) wrt all possible results

Example: "true θ" = 1/2, n = 1



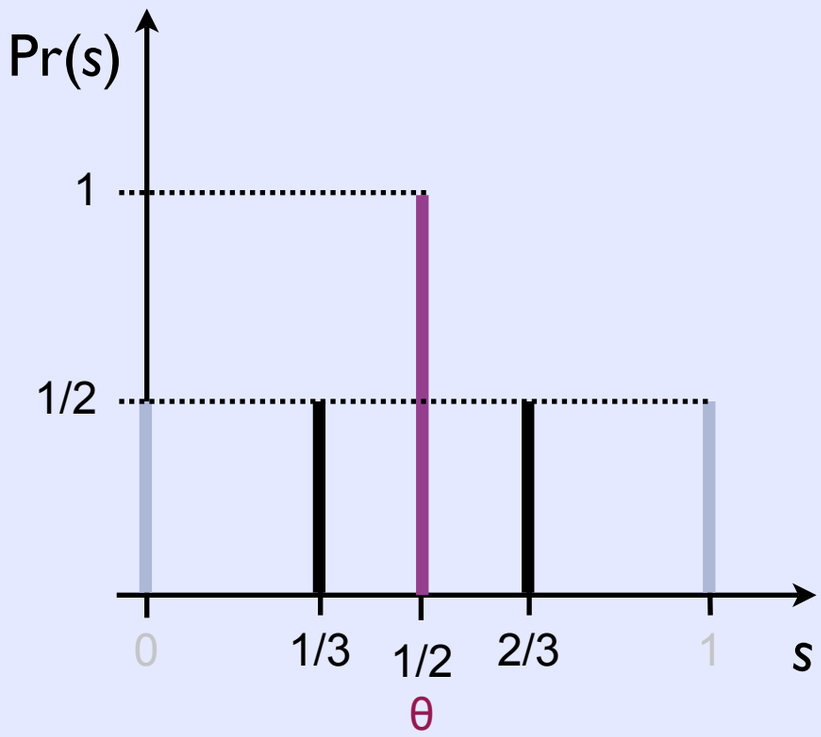
$$F(n,s) = \frac{s}{n} = \begin{cases} 0 & s = 0 \\ 1 & s = 1 \end{cases}$$

The difference from the true distribution is 1/2

Bayesian vs Frequentist

The surprising thing is that the Frequentist approach can be worse than the Bayesian approach even when the trials give a "good" result, or when we consider the average difference (from the "true" θ) wrt all possible results

Example: "true θ" = 1/2, n = 1



$$F(n,s) = \frac{s}{n} = \begin{cases} 0 & s = 0 \\ 1 & s = 1 \end{cases}$$

The difference from the true distribution is 1/2

A better function would be

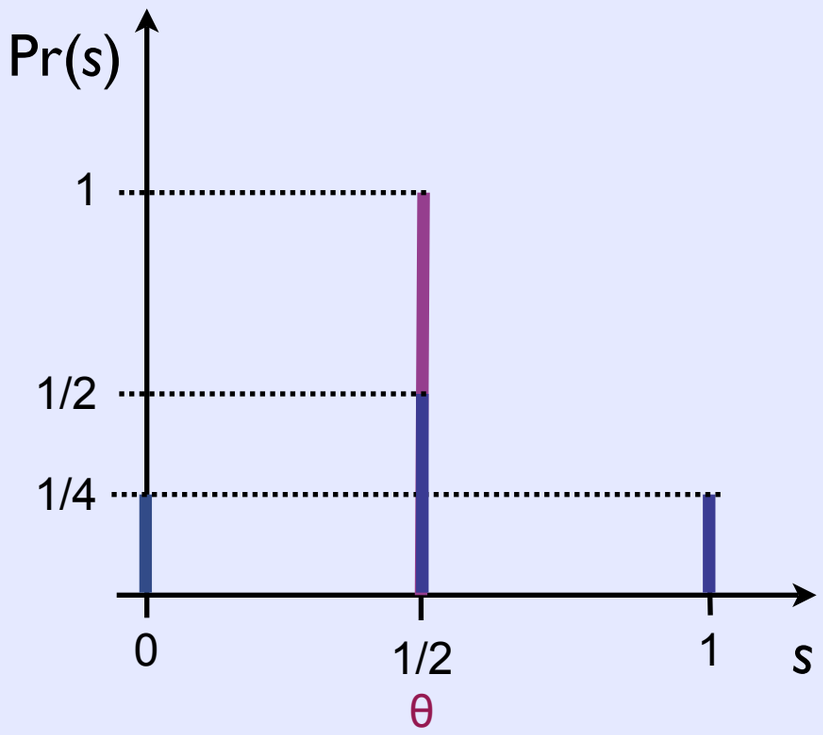
$$F_c(n,s) = \frac{s+1}{n+2} = \begin{cases} \frac{1}{3} & s = 0 \\ \frac{2}{3} & s = 1 \end{cases}$$

The difference from the true distribution is 1/6

Bayesian vs Frequentist

The surprising thing is that the Frequentist approach can be worse than the Bayesian approach even when the trials give a "good" result, or when we consider the average difference (from the "true" θ) wrt all possible results

Example: "true θ" = 1/2, n = 2



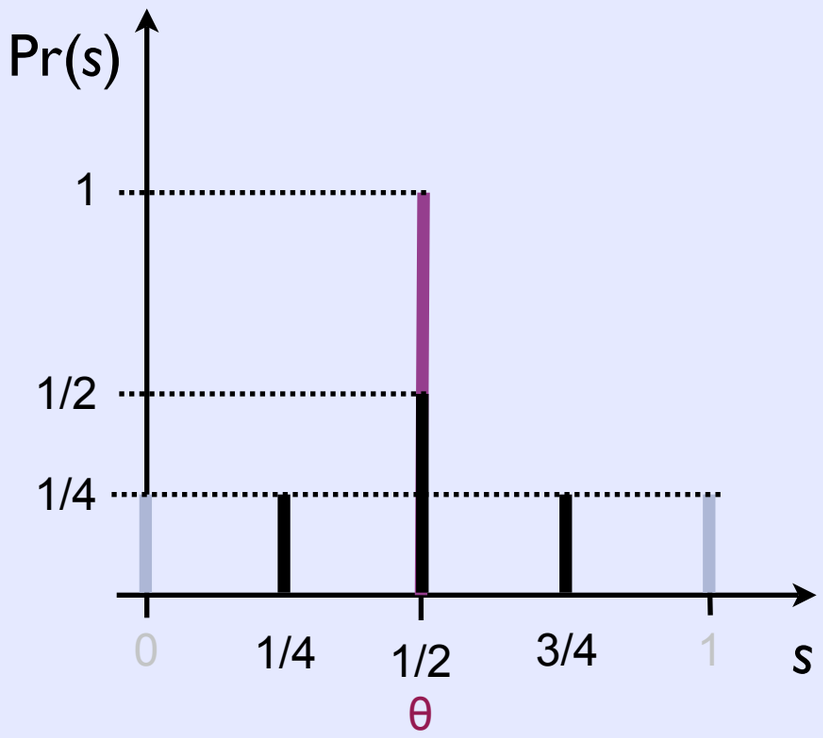
$$F(n,s) = \frac{s}{n} = \begin{cases} 0 & s = 0 \\ \frac{1}{2} & s = 1 \\ 1 & s = 2 \end{cases}$$

The average difference from the true distribution is 1/4

Bayesian vs Frequentist

The surprising thing is that the Frequentist approach can be worse than the Bayesian approach even when the trials give a "good" result, or when we consider the average difference (from the "true" θ) w.r.t. all possible results

Example: "true θ" = 1/2, n = 2



$$F(n,s) = \frac{s}{n} = \begin{cases} 0 & s = 0 \\ \frac{1}{2} & s = 1 \\ 1 & s = 2 \end{cases}$$

The average distance from the true distribution is 1/4

Again, a better function would be

$$F_c(n,s) = \frac{s+1}{n+2} = \begin{cases} \frac{1}{4} & s = 0 \\ \frac{1}{2} & s = 1 \\ \frac{3}{4} & s = 2 \end{cases}$$

The average distance from the true distribution is 1/8

A Bayesian approach

- **Assumption**: θ is the generic value of a continuous random variable Θ whose probability density is a *Beta distribution* with (unknown) parameters σ , φ

$$B(\sigma,\varphi)(\theta) = \frac{\Gamma(\sigma+\varphi)}{\Gamma(\sigma)\Gamma(\varphi)}\ \theta^{\sigma-1}(1-\theta)^{\varphi-1}$$

where Γ is the extension of the factorial function
i.e. $\Gamma(n) = (n-1)!$ for n natural number

- Note that the uniform distribution is a particular case of Beta distribution, with $\sigma = 1,\ \varphi = 1$

- B(σ , φ) can be seen as the a posteriori probability density of Θ given by a uniform a priori (principle of maximum entropy) and a trial sequence resulting in σ -1 successes and φ -1 failures.

The Bayesian Approach

- Following the approach, we have three probability density functions for Θ:
  - $B(\sigma, \varphi)$: the "real" distribution of Θ
  - $B(\alpha, \beta)$: the *a priori* (the distribution of Θ at the best of our knowledge)
  - $B(s+\alpha, f+\beta)$: the *a posteriori* (the distribution of Θ after the trials

- The result of the mean-based algorithm is :

$$A_{\alpha,\beta}(n,s) = E_{B(s+\alpha,f+\beta)}(\Theta) = \frac{s+\alpha}{s+f+\alpha+\beta} = \frac{s+\alpha}{n+\alpha+\beta}$$

Measuring the precision of Bayesian algorithms

- Define a "difference" $D(A(n,s), \theta)$ (not necessarily a distance)
  - non-negative
  - zero iff $A(n,s) = \theta$
  - what else?
- Consider the expected value $D_E(A,n, \theta)$ of $D(A(n,s), \theta)$ with respect to the likelihood (the conditional probability of s given θ)

$$D_E(A, n, \theta) = \sum_{s=0}^{n} Pr(s \mid \theta)\ D(A(n, s), \theta)$$

- *Risk of A* : the expected value $R(A,n)$ of $D_E(A,n, \theta)$ with respect to the "true"

$$R(A, n) = \int_0^1 Pd(\theta)\ D_E(A, n, \theta)\ d\theta$$

Measuring the precision of Bayesian Algorithms

We have considered the following candidates for $D(x,y)$ (all of which can be extended to the n-ary case):

- The norms:
  - $|x - y|$
  - $|x - y|^2$
  - ...
  - $|x - y|^k$
  - ...
- The Kullback-Leibler divergence

$$D_{KL}((y, 1-y) \parallel (x, 1-x)) = y \log_2 \frac{y}{x} + (1-y) \log_2 \frac{1-y}{1-x}$$

Measuring the precision of Bayesian algorithms

- *Theorem.* For the mean-based Bayesian algorithms, with *a priori* $B(\alpha, \beta)$, we have that the condition is satisfied (i.e. the Risk is minimum when α, β coincide with the parameters σ, φ of the "true" distribution), by the following functions:
  - The 2nd norm $(x - y)^2$
  - The Kullback-Leibler divergence
- We find it very surprising that the condition is satisfied by these two very different functions, and not by any of the other norms $|x - y|^k$ for $k \neq 2$

Possible applications (work in progress)

- We can use D_E to compare two different estimation algorithms; develop a measure of quality for "decision-making" algorithms
  - Mean-based vs other ways of selecting a θ
  - Bayesian vs non-Bayesian
  - In more complicated scenarios there may be different Bayesian mean-based algorithms. Example: noisy channel.
- D_E induces a metric on distributions. Bayes' equations define transformations on this metric space from the a priori to the a posteriori. We intend to study the properties of such transformations in the hope that they will reveal interesting properties of the corresponding Bayesian methods, independent of the a priori.

Possible applications (work in progress)

- Hypothesis testing (Privacy, Anonimity, Confidentiality, Information Flow Analysis, Input Distribution Analysis, ...) :

- determine (probabilistic) bounds as to what probability-distribution inference algorithm may determine about you, your online activity, your software