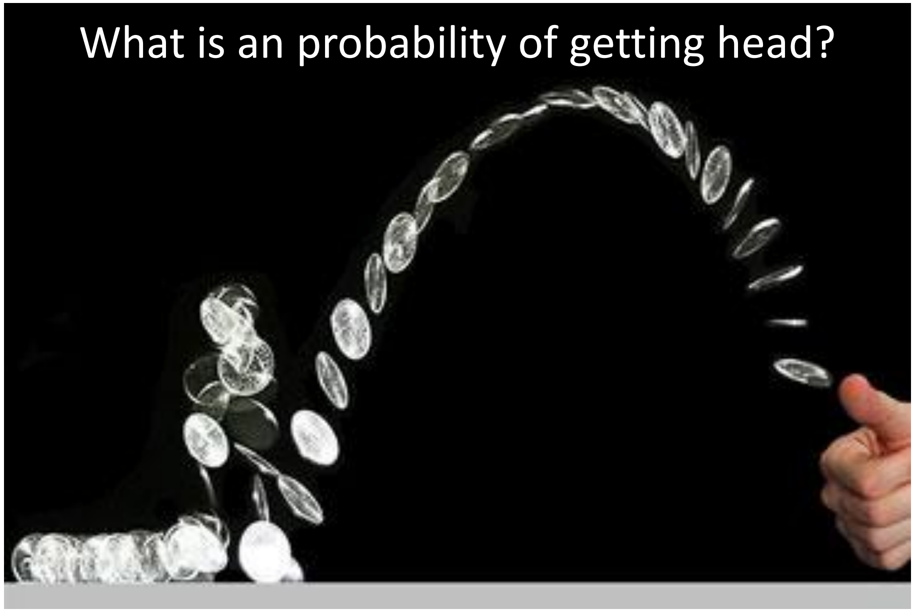
L2. Fundamentals of Bayesian Statistics



Statistics infer the causes that generated the observed data.

Model	→	data
θ (parameters) : characteristics of a model		$y = (y_1, \ldots, y_n)$: Observed consequence
θ : Probability of having a head for each coin tossing		(Head, Head, Tail, …)

View on Probability

Frequentists :

- probability only has a meaning in terms of a limiting case of repeated measurements.
- probabilities are fundamentally related to frequencies of events.

Bayesian :

- degrees of certainty about statements
- probabilities are fundamentally related to our own knowledge about an event.

View on Statistics

Approaches to Statistics

Frequentists :

- Data are a repeatable random sample → there is a frequency of occurrence
- Underlying parameters remain constant during this repeatable process
- Parameters are fixed and unchanging under all realistic circumstances

The true parameters are fixed, and a subset of data are realized from these parameters.
Then, we randomly sample a subset of data (varying) to estimate the fixed parameters.

Bayesian :

- Data are observed from the realized sample
- Parameters are unknown and described probabilistically

  (View the world probabilistically)
- Data are fixed

We collected data, thus the data is given to us (fixed).
Then, we try to estimate model parameters that can best describe the collected data.

Frequentist vs Bayesian : Coin Tossing

Frequentist

- θ = relative frequency of head in a "large number" of "identical flip"
- Statistical results assume that data were from a controlled experiment
- Nothing is more important than repeatability

(e.g., same experimental conditions)

Try

1. Estimate the parameter θ by conducting experiments
2. Give me estimates

Frequentist vs Bayesian : Coin Tossing

Frequentist

- θ = relative frequency of head in a “large number” of “identical flip”
- Statistical results assume that data were from a controlled experiment (e.g., same experimental conditions)
- Nothing is more important than repeatability

Issues

- When the number of trials n is small, estimation is biased $\frac{\#\text{Success}}{\#\text{Trials}}: \frac{1}{3}, \frac{5}{6}, \frac{5}{13}, \frac{129}{313}, \frac{61423}{123400}$
- Identical flip (controlled experiment) is unrealistic

The Artist might be a frequentist

http://www.dotmancando.info/index.php?/projects/coin-flipper/

Frequentist vs Bayesian : Coin Tossing

Frequentist

- θ = relative frequency of head in a "large number" of "identical flip"
- Statistical results assume that data were from a controlled experiment
- Nothing is more important than repeatability

Issues

- When the number of trials n is small, estimation is biased $\frac{\#\text{Success}}{\#\text{Trials}}: \frac{1}{3}, \frac{5}{6}, \frac{5}{13}, \frac{129}{313}, \frac{61423}{123400}$
- Identical flip (controlled experiment) is unrealistic

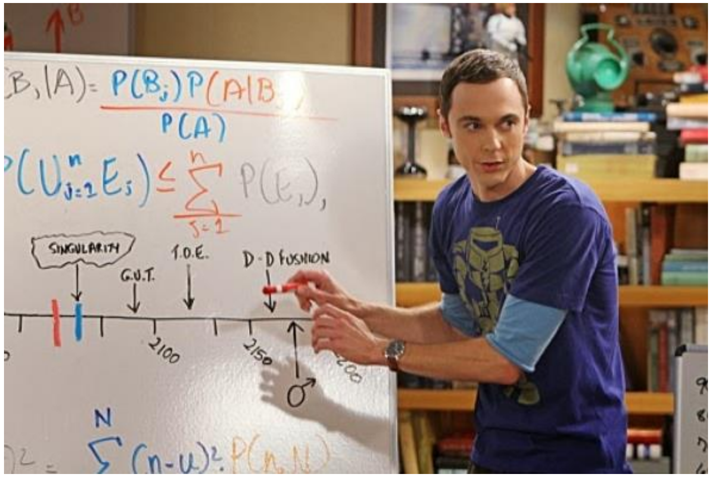
Bayesian

- Parameters θ are varied (uncertain)← Core part of Bayesian approach
- Use probability concept to provide our belief on θ : $p(\theta)$
- Each parameter θ can be associated with different conditions, i.e., orientation, force, etc.
- Data are fixed

Issues

- Subjective on $p(\theta)$
- How to specify $p(\theta)$

Bayes' rule



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

Bayes' rule will play fundamental role in proceeding Bayesian statistical analysis

Derivation of Bayes' rule :

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

Since $P(A \wedge B) = P(B \wedge A)$,

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

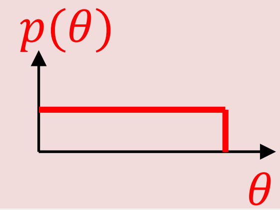
$$\Rightarrow P(A|\mathrm{B}) = \frac{P(B|A)P(A)}{P(B)}$$

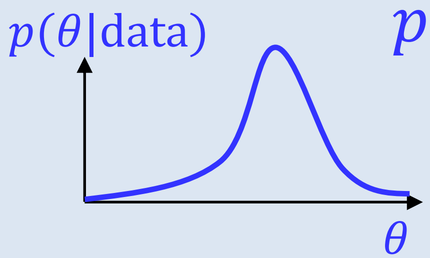
Bayes' rule in Statistics

The posterior :
The probability on the parameter θ given the evidence (data)

The evidence
The probability of getting this evidence given the parameter

The Prior
The belief (uncertainty) about the parameter θ





$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta)p(\theta)}{p(\text{data})}$$

The Marginal probability
The probability of the evidence, Probability of data over all possibilities

- Parameter instantiates one model among a model class

$$p(\theta|\text{data}, \text{model}) = \frac{p(\text{data}|\theta, \text{model})p(\theta|\text{model})}{p(\text{data}|\text{model})}$$

Approaches

1. Select a model for data (structure)
2. Specify a prior on model parameters
3. Calculate the likelihood
4. Construct posterior
5. If necessary, predict unobserved value given the updated information on the parameters

Estimating Model Parameters

Frequentist approach

1. Construct a likelihood function

$$L(\theta) = p(y|\theta)$$

2. Select the parameters θ that maximize the likelihood function:

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, L(\theta)$$

(Maximum likelihood estimation (MLE))

Bayesian approach

1. Construct posterior distribution

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

$$= \frac{p(y|\theta)p(\theta)}{\int_\theta p(y|\theta)p(\theta)d\theta}$$

$$\propto p(y|\theta)p(\theta)$$

($p(y)$ serves as a normalizing constant in terms of θ)

2. Use posterior distribution as a estimation

$$p(\theta|y)$$

(Bayesian Posterior estimation)

3. Or, select the parameters that maximize $p(\theta|y)$

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, p(\theta|y)$$

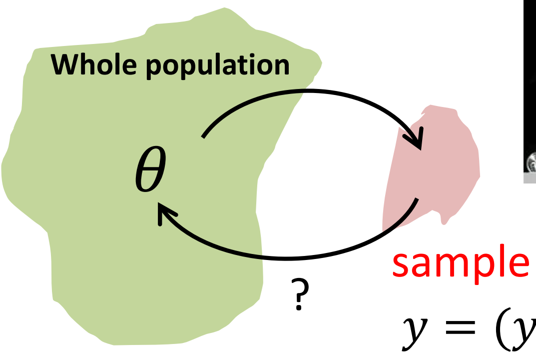
(Maximum a posteriori estimation (MAP))

Maximum Likelihood Estimation for Coin Flipping Probability

$Y_i \sim \mathrm{B}(\theta)$ $\quad Y_i = \begin{cases} 1 & \text{if Head} \\ 0 & \text{if Tail} \end{cases}$

$$p(y_i) = \mathrm{B}(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

$\theta \in [0,1]$: Probability of having a head



$$y = (y_1, y_2, \dots, y_n)$$

- Likelihood of a single observation :

$$L(\theta) = P(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

- Likelihood of a three-observations $y = (y_1, y_2, y_3) = (1,0,1)$:

$$L(\theta) = p(y_1 = 1, y_2 = 0, y_3 = 1|\theta)$$
$$= p(1|\theta)P(0|\theta)P(1|\theta) = \theta^2(1-\theta)^1$$

i.i.d.→ Exchangeability

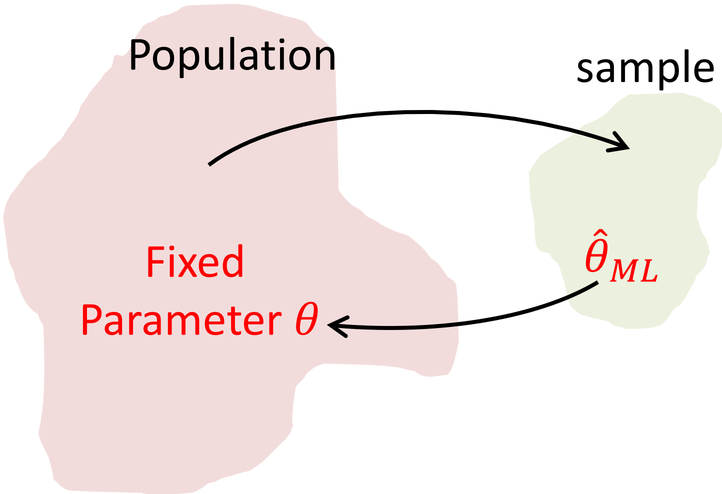
- Likelihood of n −observations :

$$L(\theta) = P(y_1, y_2, \dots, y_n|\theta) = \theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i}$$

$\Sigma_{i=1}^{n} y_i$: Number of Heads in n trials (sufficient statistics)

A statistic $t = T(x)$ is sufficient for underlying parameter θ if $p(x|t,\theta) = p(x|t)$

Maximum Likelihood Estimation for Coin Flipping Probability



Maximum likelihood estimation :

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, L(\theta) = p(y_1, y_2, \dots, y_n|\theta) = \theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i}$$

$$\frac{\partial L(\theta)}{\partial \theta} = \Sigma y_i \theta^{\Sigma y_i - 1}(1-\theta)^{n-\Sigma y_i} - \theta^{\Sigma y_i}(n - \Sigma y_i)(1-\theta)^{n-\Sigma y_i - 1} = 0$$

$$\Rightarrow \hat{\theta}_{ML} = \frac{\Sigma y_i}{n}$$

MLE estimation gives a relative frequency

Bayesian Approach for Estimating Model Parameters

The essential characteristics of Bayesian methods
= explicit use of probability for **quantifying uncertainty** in the statistical models

Bayes' rule:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

$$= \frac{p(y|\theta)p(\theta)}{\int_\theta p(y|\theta)p(\theta)d\theta} \qquad \left(\because p(y) = \int_\theta p(y|\theta)p(\theta)d\theta\right)$$

$$\propto p(y|\theta)p(\theta)$$

Posterior ∝ Likelihood X Prior

$p(\theta)$: Prior - subjective belief about θ

$p(y|\theta)$: Likelihood – observation (data) regarding θ

$p(\theta|y)$: Posterior - Updated belief about θ with the data

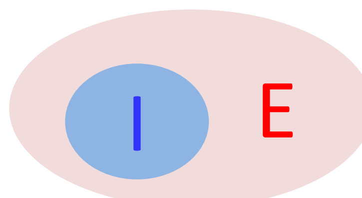
Exchangeability

$P($  $) = P($  $)$

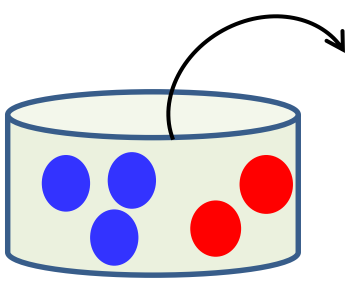
Definition (*Infinite exchangeability*). We say that $(y_1, y_2, \ldots)$ is an infinitely exchangeable sequence of random variables if, for any n, the joint probability $p(y_1, y_2, \ldots y_n)$ is invariant to permutation of the indices. That is, for any permutation π,

$$p_Y(y_1, y_2, \ldots y_n) = p_Y\left(y_{\pi_1}, y_{\pi_2}, \ldots, y_{\pi_n}\right)$$

I R.V.s are independent and identically distributed (i.i.d)



Random variables are infinitely exchangeable **E**



$$P(R, R, B, B, B) = P(B, R, B, B, R)$$

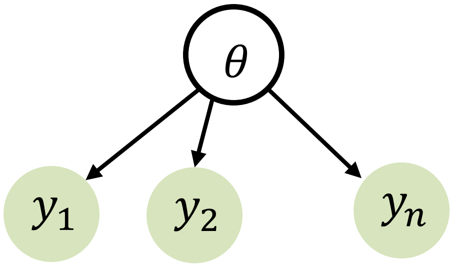
Exchangeable

Check!!

$$P(t_2 = R | t_1 = R) \neq P(t_2 = R | t_1 = B)$$

Not independent

Exchangeability



$y_1, y_2, \ldots y_n$ are conditionally independent given θ

Theorem (De Finetti, 1930s). A sequence of random variables $(y_1, y_2, \ldots)$ is infinitely exchangeable *iff, for all n,*

$$p(y_1, y_2, \ldots y_n) = \int \prod_{i=1}^{n} p(y_i|\theta)p(\theta)d\theta$$

e.g., for coin tossing example: $p(y_1, y_2, \ldots y_n) = \int \theta^{\Sigma y_i}[1-\theta]^{N-\Sigma y_i} p(\theta)d\theta$

The theorem says that if we have exchangeable data,

- There must exist a parameter θ
- There must exist a likelihood $p(y|\theta)$
- There must exist a distribution $p(\theta)$
- The above quantities must exist so as to render the data $\mathrm{y} = (y_1, y_2, \ldots, y_n)$ conditionally independent

⇒Prior (Bayesian approach) is suggested by the data being exchangeable

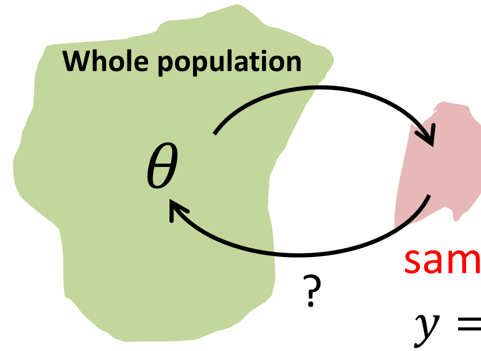
A and B are conditionally independent if $A \perp B|\theta \Rightarrow p(A,B|\theta) = p(A|\theta)p(B|\theta)$

Maximum a posteriori estimation (Bayesian Estimation) for Coin Flipping Probability

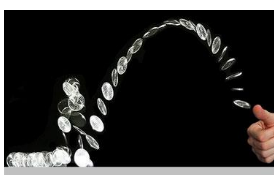
$$Y_i \sim \mathrm{B}(\theta) \qquad Y_i = \begin{cases} 1 & \text{if Head} \\ 0 & \text{if Tail} \end{cases}$$

$$p(y_i) = \mathrm{B}(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

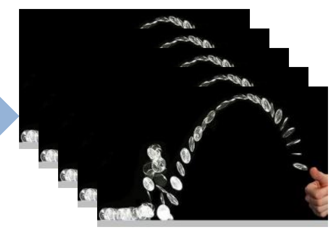
$\theta \in [0,1]$: Probability of having a head



$$y = (y_1, y_2, \dots, y_n)$$



n times



Bernoulli distribution

$$Y_i \sim \mathrm{B}(\theta) \qquad Y_i = \begin{cases} 1 & \text{if Head} \\ 0 & \text{if Tail} \end{cases}$$

$$p(y_i) = \mathrm{B}(y_i|\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

$\theta \in [0,1]$: Probability of having a head

Binomial distribution

$$Y \sim \mathrm{Bin}(n, \theta)$$

$$p(y|\theta) = \mathrm{Bin}(y|n,\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y}$$

$y \in \{0,1,\dots,n\}$:
The number of successes in a sequence of n independent yes/no experiments

Maximum a posteriori estimation (Bayesian Estimation) for Coin Flipping Probability

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int_0^1 p(y,\theta)d\theta} = \frac{p(y|\theta)p(\theta)}{\int_0^1 p(y|\theta)p(\theta)d\theta}$$

Likelihood :

$$Y\sim\mathrm{Bin}(n,\theta) \rightarrow p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

y = # of success among n trial

Prior:

$$\theta\sim\mathrm{Beta}(\alpha,\beta) \rightarrow p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Numerator

$$\begin{aligned}
p(y|\theta)p(\theta) &= \mathrm{Binomial}(y|n,\theta) \times \mathrm{Beta}(\theta|\alpha,\beta) \\
&= \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}
\end{aligned}$$

Maximum a posteriori estimation (Bayesian Estimation) for Coin Flipping Probability

denominator

$$\binom{n}{y} = \frac{n!}{y!\,(n-y)!} = \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}$$

$$\begin{aligned}
p(y) &= \int_0^1 p(y|\theta)p(\theta)d\theta \\
&= \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\
&= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta \\
&= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}d\theta \\
&= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)}\int_0^1 \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}d\theta \\
&= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)}\int_0^1 \mathrm{Beta}(\theta|y+\alpha,n-y+\beta)d\theta \\
&= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)} \\
&= \text{Beta−Binomial}(y|n,\alpha,\beta)
\end{aligned}$$

Maximum a posteriori estimation (Bayesian Estimation) for Coin Flipping Probability

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_0^1 p(y|\theta)p(\theta)d\theta} = \frac{\text{Numerator}}{\text{Denominator}}$$

$$= \frac{\dfrac{\cancel{\Gamma(n+1)}\cancel{\Gamma(\alpha+\beta)}}{\cancel{\Gamma(x+1)}\cancel{\Gamma(n-x+1)}\cancel{\Gamma(\alpha)}\cancel{\Gamma(\beta)}}}{\dfrac{\cancel{\Gamma(n+1)}\cancel{\Gamma(\alpha+\beta)}\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\cancel{\Gamma(y+1)}\cancel{\Gamma(n-y+1)}\cancel{\Gamma(\alpha)}\cancel{\Gamma(\beta)}\Gamma(n+\alpha+\beta)}} \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}$$

$$= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} \theta^{\alpha+y-1}(1-\theta)^{\beta+n-y-1}$$

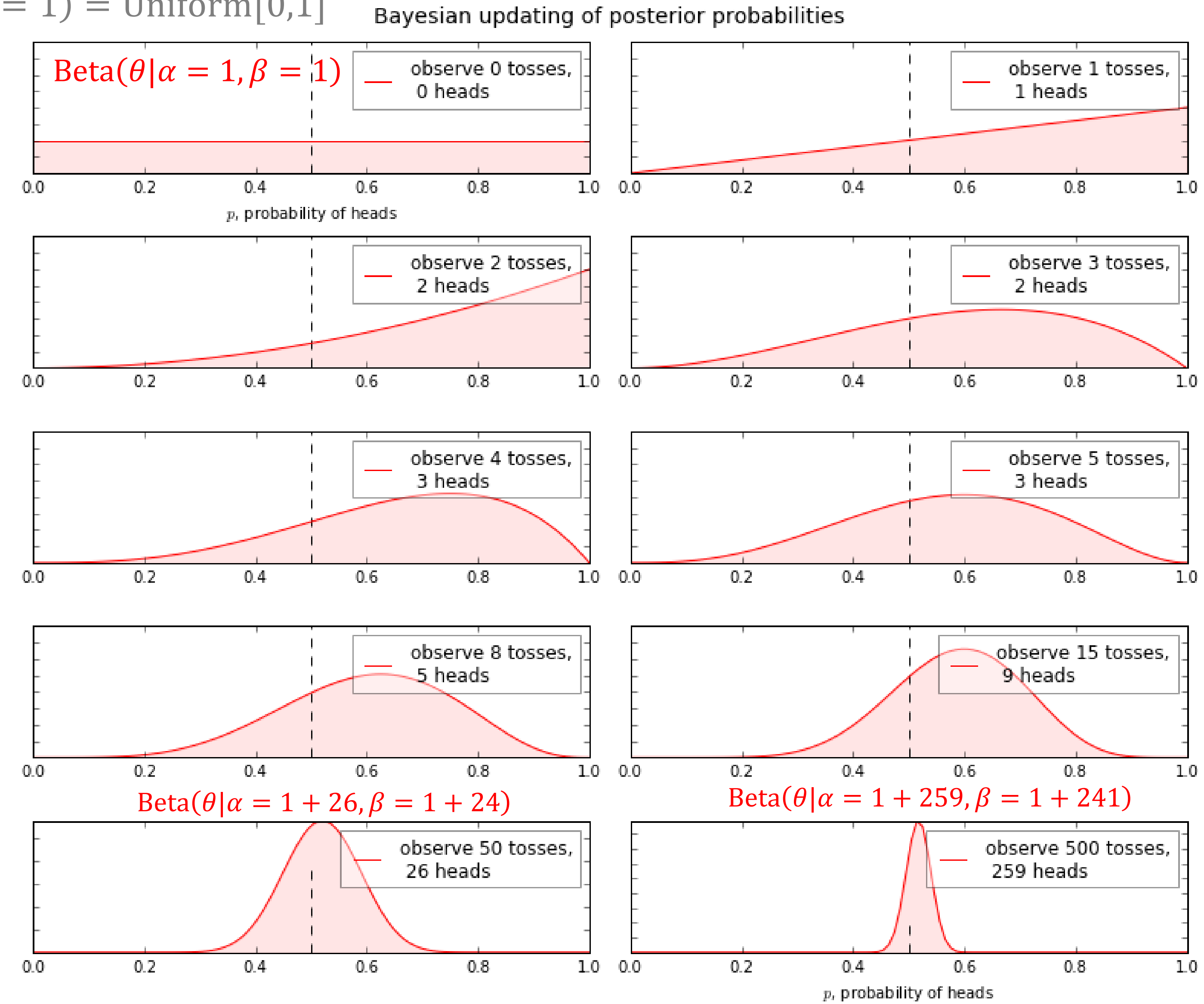
$$= \text{Beta}(\theta|\alpha+y, \beta+n-y)$$

$$\text{p}(\theta) = \text{Beta}(\theta|\alpha,\beta) \xrightarrow{\text{data}} \text{p}(\theta|y) = \text{Beta}(\theta|\alpha+y, \beta+n-y)$$

Maximum a posteriori estimation (Bayesian Estimation) for Coin Flipping Probability

$\mathrm{Beta}(\alpha=1,\beta=1)=\mathrm{Uniform}[0,1]$

$\mathrm{Beta}(\theta|\alpha,\beta)$



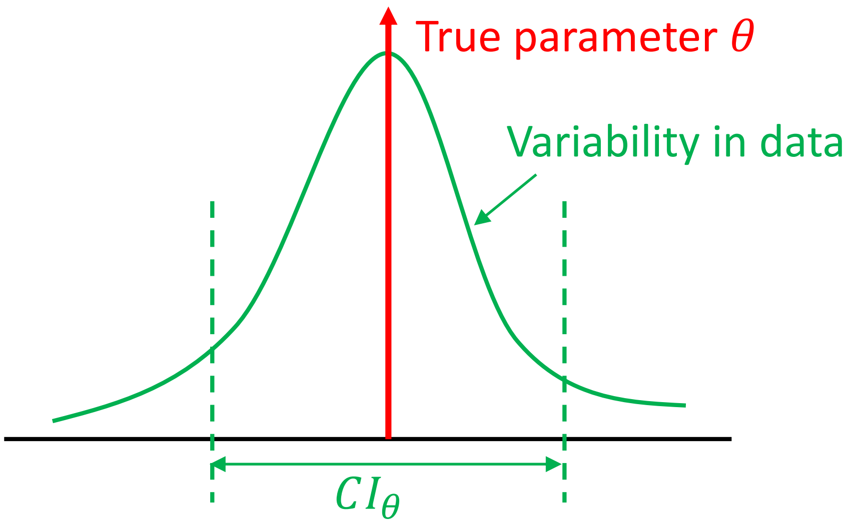
Maximum a posteriori estimation (Bayesian Estimation) for Coin Flipping Probability

Jupyter Demo Simulation

Confidence Interval vs. Credible region

Frequentist approach

- Describe variability in data given the fixed parameter θ

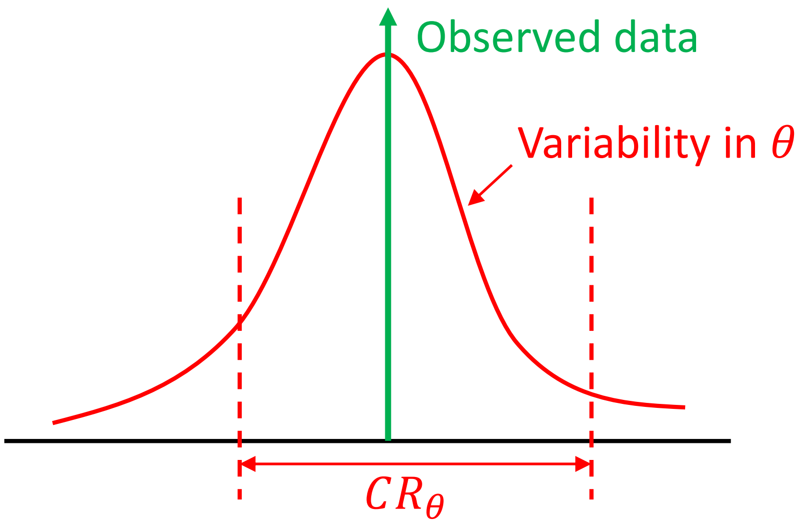


"There is a 95% probability that when I compute CI_θ from a current data, the computed CI contains θ_{true}

→ From current data set,
We can only say that $\theta \in \mathrm{CI}$ or $\theta \notin \mathrm{CI}$

Bayesian approach

- Describe variability in θ for fixed data



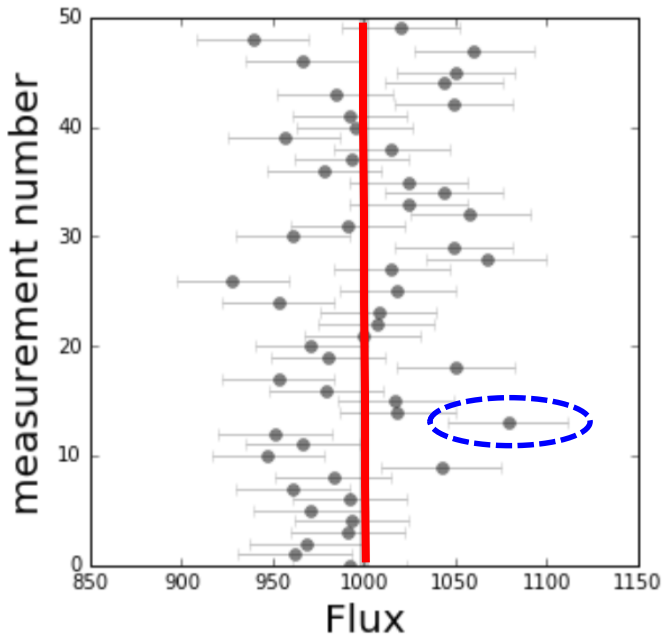
"Given our observed data, there is a 95% probability that the true value of θ falls within Credible region CR_θ"

→ From current data set,
We can make probabilistic statement such as $\Pr(\theta \in CR) = 0.95$

Confidence Interval vs. Credible region

Frequentist approach

- Describe variability in data given the fixed parameter θ

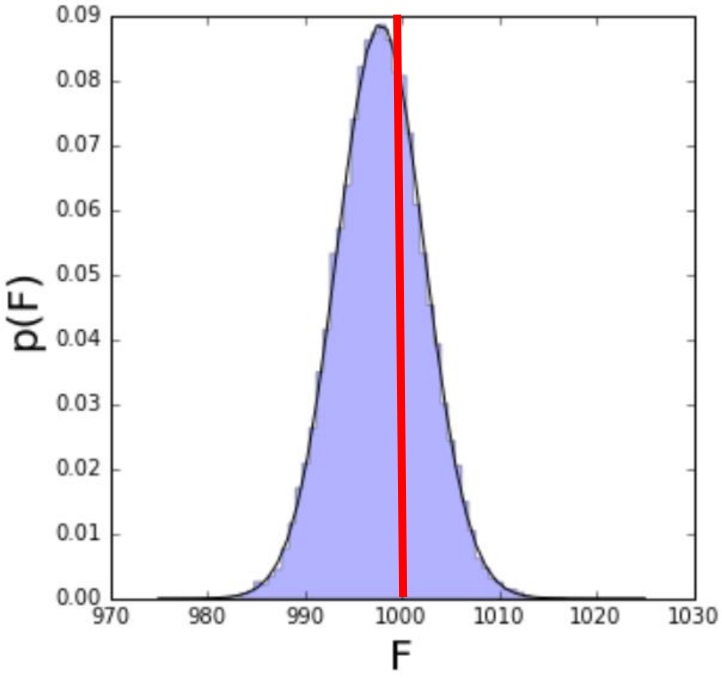


$\theta \in \mathrm{CI}$ or $\theta \notin \mathrm{CI}$

If the experiment were repeated an infinite number of times, 95% of the calculated intervals would contain θ .

Bayesian approach

- Describe variability in θ for fixed data



$\Pr(\theta \in \mathrm{CR}) = 0.95$

There is a 95% chance θ is in CR

Maximum a posteriori estimation (Bayesian Estimation) for Coin Flipping Probability

- From the previous result $p(\theta|y) = \text{Beta}(\theta|\alpha + y, \beta + n - y)$,

$$\mathbb{E}(\theta) = \frac{\alpha}{\alpha + \beta}$$

$$\mathbb{E}(\theta|\mathrm{y}) = \frac{\alpha + y}{\alpha + \beta + n} = \frac{\alpha}{\alpha + \beta}\frac{\alpha + \beta}{(\alpha + \beta + n)} + \frac{y}{n}\frac{n}{(\alpha + \beta + n)}$$

$$= \mathbb{E}(\theta)\frac{\alpha + \beta}{\alpha + \beta + n} + \hat{\theta}_{ML}\frac{n}{\alpha + \beta + n}$$

- As $n \to \infty$, $\mathbb{E}(\theta|\mathrm{y}) \to \frac{y}{n} = \hat{\theta}_{ML}$
- Large value of $\alpha + \beta$ signifies Posterior

→In the limit, the prior does not influence the results. That is, the results are dominated by the data (observation).

$$\mathrm{var}(\theta|y) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} = \frac{E(\theta|y)\big(1 - E(\theta|y)\big)}{\alpha + \beta + n + 1}$$

- As n and $(n - y) \to \infty$, $\mathrm{var}(\theta|y) \to \frac{1}{n}\frac{y}{n}\left(1 - \frac{y}{n}\right) = \frac{p(1-p)}{n}$

Posterior as compromise between data and prior information

- $\mathrm{E}(u) = \mathrm{E}\big(\mathrm{E}(u|v)\big)$

  $\Rightarrow \mathrm{E}(\theta) = \mathrm{E}\big(\mathrm{E}(\theta|y)\big)$

  The prior mean of θ is the average of all possible posterior means over the distribution of possible data

- $\mathrm{var}(u) = E\big(\mathrm{var}(u|v)\big) + \mathrm{var}\big(E(u|v)\big)$

  $\Rightarrow \mathrm{E}\big(\mathrm{var}(\theta|y)\big) = \mathrm{var}(\theta) - \mathrm{var}\big(\mathrm{E}(\theta|y)\big)$

  The posterior variance is on average smaller than the prior variance by an amount that depends on the variation in posterior means over the distribution of all the possible data

The posterior distribution is centered at a point that represents a compromise between the prior information and the data, and the compromise is controlled to a greater extent by the data as the sample size increases

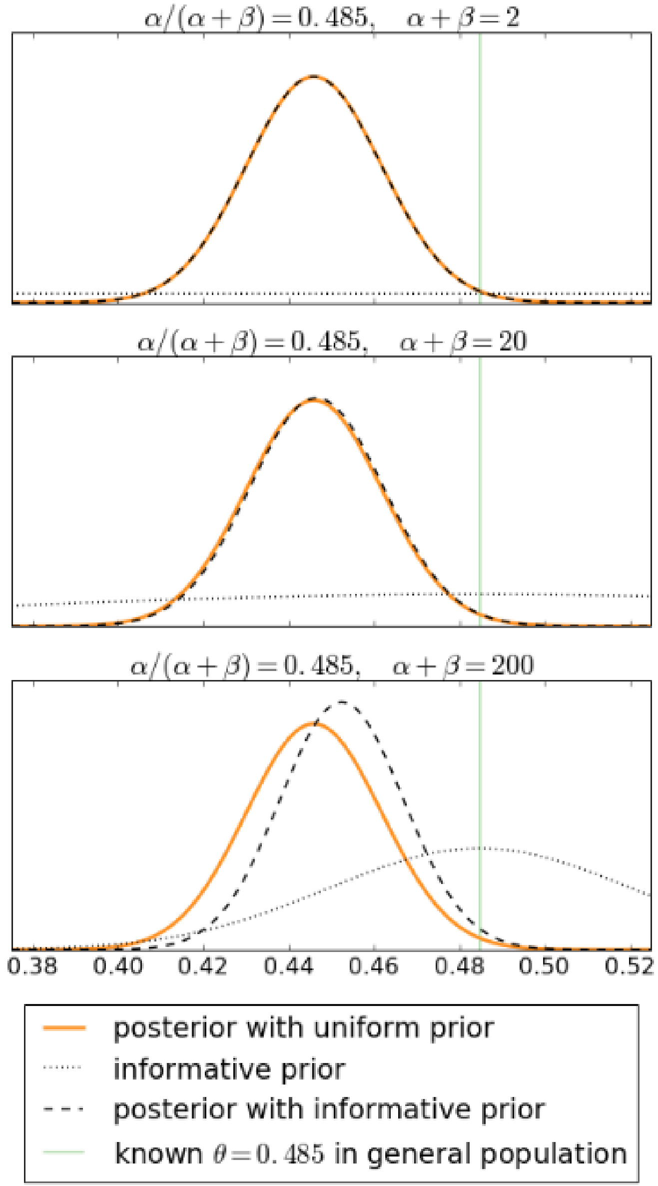
The Role of Prior

Example (BDA Ch.2.4)
Probability of girl birth given placenta Previa

- Among 980 births with placenta Previa, 437 are females
- How much evidence does this provide for the claim that the proportion of female birth in the population of placenta Previa birth is less than 0.485?

$$P(\theta) = \text{Beta}(\theta|\alpha, \beta)$$
$$P(\theta|y) = \text{Beta}(\theta|\alpha + 437, \beta + 543)$$
$$\mathbb{E}(\theta) = \frac{\alpha + 437}{\alpha + 437 + \beta + 543}$$

Parameters of the prior distribution		Summaries of the posterior distribution	
$\frac{\alpha}{\alpha+\beta}$	$\alpha+\beta$	Posterior median of θ	95% posterior interval for θ
0.500	2	0.446	[0.415, 0.477]
0.485	2	0.446	[0.415, 0.477]
0.485	5	0.446	[0.415, 0.477]
0.485	10	0.446	[0.415, 0.477]
0.485	20	0.447	[0.416, 0.478]
0.485	100	0.450	[0.420, 0.479]
0.485	200	0.453	[0.424, 0.481]



Beyond Parameter Estimation: Prediction based on Bayesian Approach

- **Posterior distribution**

  the distribution of the unknown and an observable parameter θ given observed y

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{p(y|\theta)p(\theta)}{\int_\theta p(y|\theta)p(\theta)d\theta} \propto p(y|\theta)p(\theta)$$

- **Prior predictive distribution**

  Before the data y are considered, the distribution of the unknown but observable y is

$$p(y) = \int_\theta p(y,\theta)d\theta = \int_\theta p(y|\theta)p(\theta)d\theta$$

- **Posterior predictive distribution**

  Prediction for an observable $\hat{y}$ conditional on the observed y

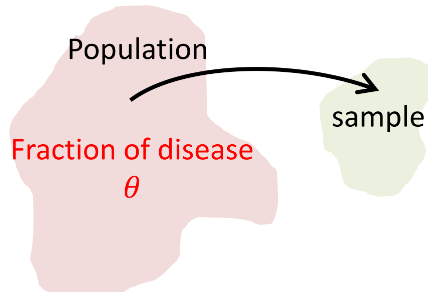
$$p(\hat{y}|y) = \int_\theta p(\hat{y},\theta|y)d\theta = \int_\theta p(\hat{y}|\theta,y)p(\theta|\mathrm{y})d\theta = \int_\theta p(\hat{y}|\theta)p(\theta|\mathrm{y})d\theta$$

- Posterior predictive distribution=an average of conditional predictions over the posterior dist. on θ

- **Prior predictive distribution**

Before the data y are considered, the distribution of the unknown but observable y is

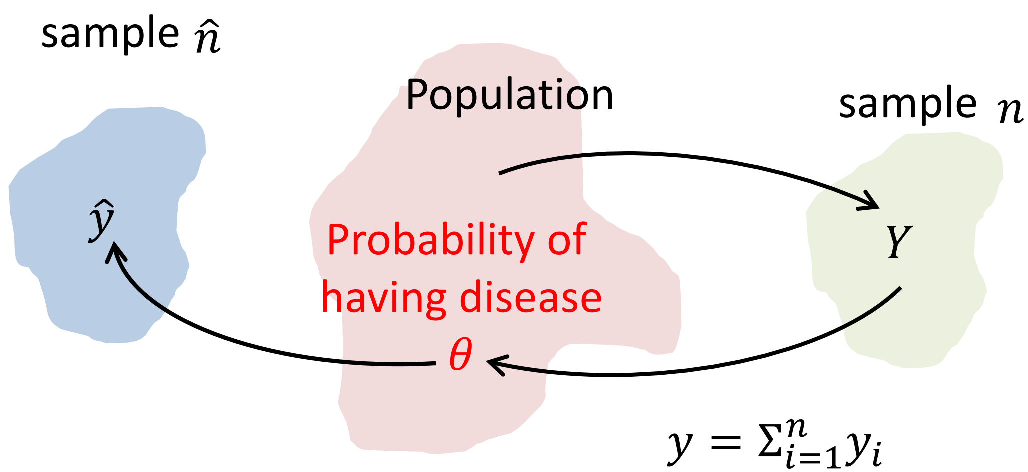
$y = \Sigma_{i=1}^{n} y_i$:the number of people with diseases among n



$$
\begin{aligned}
P(y) &= \int_0^1 P(y,\theta)d\theta \\
&= \int_0^1 P(y|\theta)p(\theta)d\theta \qquad P(y|\theta) = \mathrm{Bin}(y|n,\theta),\ \ p(\theta) = \mathrm{Beta}(\alpha,\beta) \\
&= \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta \qquad \binom{n}{y} = \frac{y!}{y!\,(n-y)!} = \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \\
&= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta^y (1-\theta)^{n-y} \theta^{\alpha-1}(1-\theta)^{b-1} d\theta \\
&= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} d\theta \\
&= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)} \int_0^1 \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} d\theta \\
&= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)} \int_0^1 \mathrm{Beta}(\theta|y+\alpha, n-y+\beta) d\theta \\
&= \frac{\Gamma(n+1)\Gamma(\alpha+\beta)\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(y+1)\Gamma(n-y+1)\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)} 1 \\
&= \text{Beta–Binomial}(y|n,\alpha,\beta)
\end{aligned}
$$

For$\alpha = 1, \beta = 1, P(y) = \frac{1}{n+1}$ (check!)

Posterior-Predictive Distribution



$y = \Sigma_{i=1}^{n} y_i$

the number of persons with diseases among n

- **Posterior predictive distribution**

  Prediction for an observable $\hat{y}$ conditional on the observed y

$$p(\hat{y}|y) = \int_{\theta} p(\hat{y}, \theta|y) d\theta = \int_{\theta} p(\hat{y}|\theta, y) p(\theta|\mathrm{y}) d\theta = \int_{\theta} p(\hat{y}|\theta) p(\theta|\mathrm{y}) d\theta$$

- Posterior predictive distribution=an average of conditional predictions over the posterior dist. on θ

Posterior-Predictive Distribution

Remember Prior Predictive distribution

$$p(y) = \int_0^1 p(y,\theta)d\theta = \int_0^1 p(y|\theta)p(\theta)d\theta$$
$$= \text{Beta−Binomial}(y|n,\alpha,\beta)$$

$p(\theta) = \mathrm{Beta}(\alpha,\beta)$

$p(\theta|y) = \mathrm{Beta}(\alpha+y, \beta+n-y)$

Using this result, the posterior predictive distribution is

$$p(\hat{y}|y) = \int_0^1 p(\hat{y},\theta|y)d\theta = \int_0^1 p(\hat{y}|\theta)p(\theta|\mathrm{y})d\theta$$
$$= \text{Beta−Binomial}(\hat{y}|n,\alpha+y,\beta+n-y)$$

When $\hat{y}=1$, $p(\hat{y}=1|y)$

$$= \int_0^1 \theta p(\theta|\mathrm{y})d\theta \qquad \because p(\hat{y}=1|\theta)=\theta$$
$$= \mathbb{E}[\theta|y]$$
$$= \frac{\alpha+y}{\alpha+y+\beta+n-y} \qquad \because \text{when } \theta\sim\mathrm{Beta}(\alpha,\beta),\ \mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta}$$
$$= \frac{\alpha+y}{\alpha+\beta+n}$$

For $\alpha=1, \beta=1$, $P(\hat{y}=1|y) = \frac{y+1}{n+2}$ (check!)

Beyond Parameter Estimation: Prediction based on Bayesian Approach

Jupyter Demo Simulation

Bayesian Approach Example : Is sea water contaminated with cholera

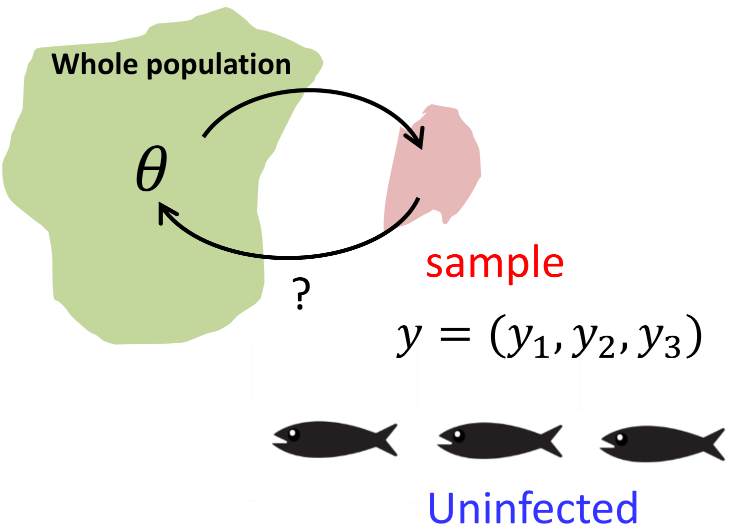
Prior

$$\theta = \begin{cases} 0 & \text{Sea without cholera} \\ 1 & \text{Sea with cholera} \end{cases}$$

$$p(\theta) = \begin{cases} 1/2 & \theta = 0 \\ 1/2 & \theta = 1 \end{cases}$$

Uninform prior

Discretized θ



Likelihood

$$y_i = \begin{cases} 1 & \text{fish with cholera} \\ 0 & \text{fish without cholera} \end{cases}$$

$$p(y_i = 1|\theta = 0) = 0 \longrightarrow p(y_i = 0|\theta = 0) = 1$$

$$p(y_i = 1|\theta = 1) = \frac{1}{2} \longrightarrow p(y_i 0|\theta = 1) = \frac{1}{2}$$

data $y = \{y_1 = 0, y_2 = 0, y_3 = 0\}$ y_i are i.i.d.

$$P(\mathrm{y}|\theta = 0) = (P(y = 0|\theta = 0))^3 = (1)^3 = 1$$

$$P(\mathrm{y}|\theta = 1) = (P(y = 0|\theta = 1))^3 = (1/2)^3 = 1/8$$

Denominator (Marginal likelihood)

$$P(y) = \sum_{\theta} P(\mathrm{y}|\theta)P(\theta)$$

$$= P(\mathrm{y}|\theta = 0)P(\theta = 0) + P(\mathrm{y}|\theta = 1\,)P(\theta = 1) = 1 \times \frac{1}{2} + \frac{1}{8} \times \frac{1}{2} = \frac{9}{16}$$

Bayesian Approach Example : Is sea water contaminated with cholera

Posterior distribution: $p(\theta|y) = \dfrac{p(y|\theta)p(\theta)}{p(y)}$

$$P(\theta = 0|y) = \frac{P(y|\theta = 0,)P(\theta = 0)}{P(y)} = \frac{1 \times 1/2}{9/16} = \frac{8}{9}$$

$$P(\theta = 1|y) = \frac{P(y|\theta = 1,)P(\theta = 1)}{P(y)} = \frac{1/8 \times 1/2}{9/16} = \frac{1}{9}$$

Update with data

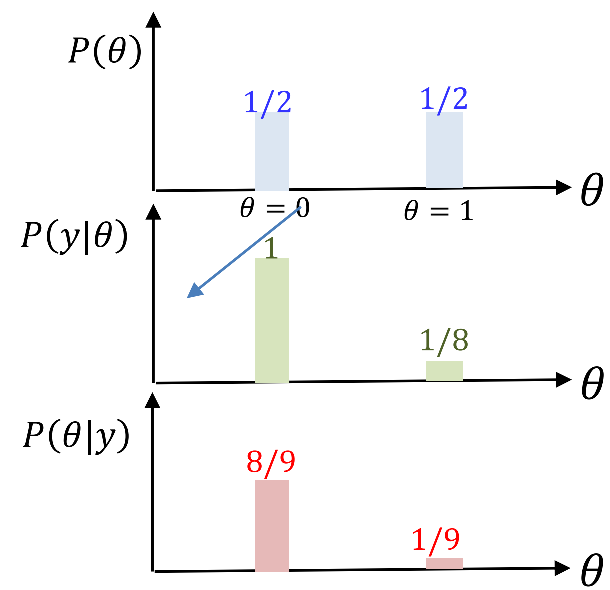
Prior

$$P(\theta = 1) = \frac{1}{2}$$

$$P(\theta = 0|y) \propto P(y|\theta = 0,)P(\theta = 0) = 1 \times 1/2$$

$$P(\theta = 1|y) \propto P(y|\theta = 1,)P(\theta = 1) = 1/8 \times 1/2$$

Bayes' factor = $\dfrac{P(\theta = 0|y)}{P(\theta = 1|y)} = \dfrac{8}{1}$



Bayes factors is a Bayesian alternative to classical hypothesis testing

$$P(\theta_1|D) = \frac{P(D|\theta_1)p(\theta_1)}{p(D)}$$

$$P(\theta_2|D) = \frac{P(D|\theta_2)p(\theta_2)}{p(D)}$$

Which parameter is good?

Bayes' factor = $\dfrac{P(\theta_1|D)}{P(\theta_2|D)}$

Three Steps in Bayesian Approaches

Step 1: Modeling

setting up a full probability model, a joint probability distribution for all observable and unobservable quantities in a target problem

Step 2: Inferencing

calculate and interpret the appropriate posterior distribution, the conditional probability distribution of the unobserved quantities of interests

Step 3: Checking

Evaluate the fit of the model and the sensitiveness of the assumption in step 1