02477 – Bayesian Machine Learning: Lecture 1

Michael Riis Andersen

Technical University of Denmark,
DTU Compute, Department of Applied Math and Computer Science

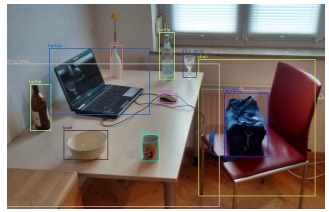
Outline for today

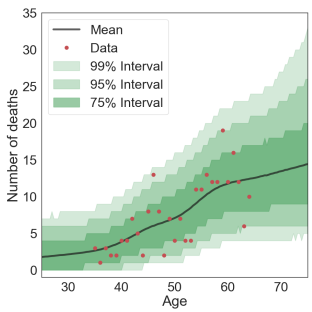
1. Introduction and course formalities
2. Bayesian machine learning
3. Bayesian inference for the beta-binomial model
4. Introduction to the exercises

Introduction and course formalities

Bayesian Machine Learning

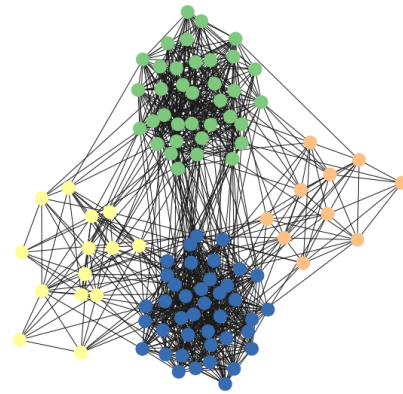
- Machine learning
  - Making predictions
  - Understanding data
  - Finding patterns
  - Making machines learn from data
  - AI
- A multitude of applications
  - Object detection
  - Speech recognition and natural language processing
  - Self-driving cars
  - Spam & fraud detection
  - Recommender systems
  - Brain imaging
  - ...
- Bayesian statistics
  - Mathematical framework for reasoning with uncertainty
  - Thomas Bayes (1702 - 1761)

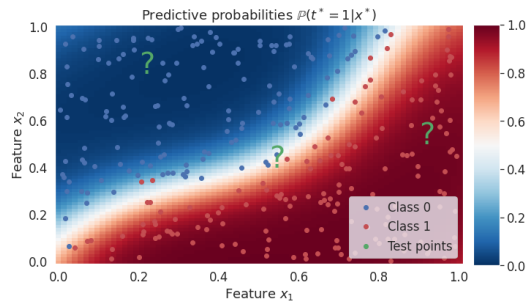


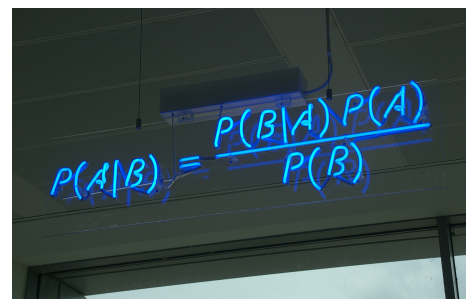


What to expect from this course?

- We will work with different machine learning problems, e.g.
  - Regression
  - Classification
  - Clustering
  - Change point detection
  - Community detection
  - A/B testing
  - ...
- A probabilistic modelling approach
  - Linear models
  - Gaussian processes
  - Neural networks
  - Probabilistic graphical models
  - Parametric vs nonparametric
  - Generative vs discriminative
- Inference algorithms and optimization
  - Exact Bayesian inference
  - Laplace approximations
  - Variational inference
  - Markov chain Monte Carlo methods
  - Stochastic optimization
- Bayesian perspective: understanding how and why
  - Deeper insights into fundamentals
  - Incorporating prior knowledge into models
  - Uncertainty quantification

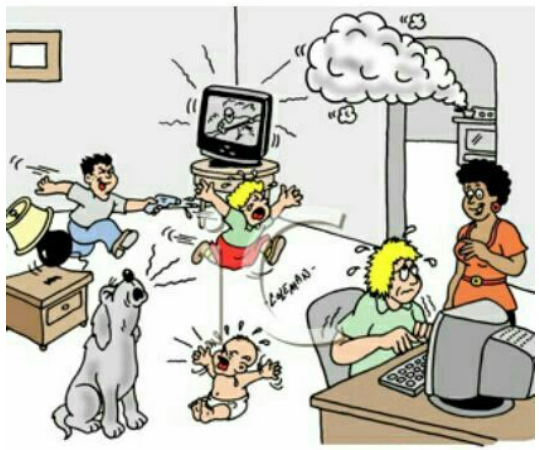






Course formalities

- When?
  - Lectures: Monday 13-15
  - Exercises: Monday 15-17
- Where?
  - Week 1-4: Online
  - Week 5-13: Lyngby Campus (subject to change)
- Exam
  - Oral exam based on exercises
  - Curriculum: All course materials, e.g. slides, book chapter, exercises etc.
  - Details T.B.A.
- Teachers
  - Michael Riis Andersen (Building 321, room 216, miri@dtu.dk)
  - TA: Bo Li (blia@dtu.dk)
  - TA: Federico Bergamin (fedbe@dtu.dk)



Learning objectives

1. Discuss fundamental concepts in Bayesian machine learning: Priors, likelihood functions, approximate inference, and evaluation.

2. Perform and analyse approximate inference by sampling and variational methods.

3. Perform and analyse tests to obtain un-biased performance estimates

4. Discuss uncertainty quantification and probability calibration in Bayesian models

5. Design systems based on semi-parametric and non-parametric Bayesian models.

6. Apply deep learning models for Bayesian machine learning in audio, image and text data

7. Implement Bayesian machine learning and evaluation methods in Python.

8. Give a verbal presentation of results obtained in hands-on exercises

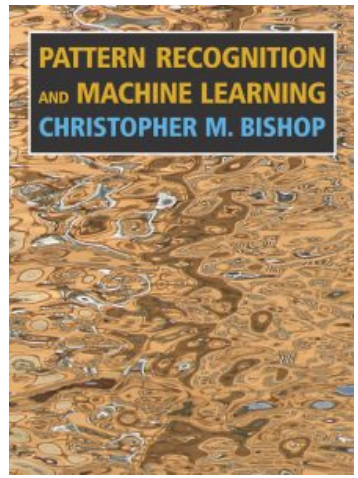
Course plan

The plan is subject to change!

Week	Topic
1	Intro, basic concepts, Beta-Binomial model
2	Bayesian Linear Regression and marginal likelihoods
3	Bayesian Classification and Laplace approximations
4	Distributions on function spaces, Gaussian Processes
5	Generalized linear models, non-linear extensions
6	Generalization, decision theory, calibration
7	Monte Carlo & Markov Chain Monte Carlo methods
8	Convergence diagnostics for MCMC, change point detection
9	Variational inference, mixture models
10	Black-box variational inference
11	Stochastic optimization
12	Network models, stochatic block models, Chinese restaurant processes
13	Bayesian neural networks

Literature

- Textbook: Pattern Recognition and machine learning by Christopher Bishop

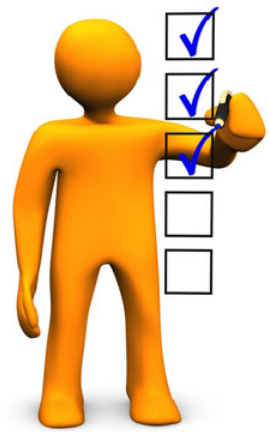


- Link: https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/
- We will supplement with other resources in the second half of the course, e.g. book chapters, research papers etc
- Other useful resources (freely available):
  - Bayesian Data Analysis by Gelman et al (link)
  - Bayesian reasoning and machine Learning by D. Barber (link)
  - Information Theory, Inference, and Learning Algorithms by D. MacKay (link)
  - Probabilistic machine learning: An introduction by K. Murphy (link)
  - Mathematics for Machine Learning by Deisenroth et al (link)

Ideal course prerequisites

- 02450 Introduction to machine learning

- Math
  1. Probability theory & statistics (Bishop chap. 1, 2, appendix B)
  2. Linear algebra (Bishop appendix C)
  3. Calculus



- Programming
  - Python
  - Mostly numpy, scipy, matplotlib etc

- It is indeed possible to complete the course without the ideal prerequisites, but you should expect an increased workload

Continuous feedback

- The course is relatively new and I need your help to improve the course!

- Feedback persons for today
  1. Alejandro Rodriguez Salamanca
  2. Aleksander Nagaj
  3. Anastas Magdych
  4. Anders Bredgaard Thuesen
  5. Andreas Lnstrup Ammitzbll
  6. Anton Baht

- We will meet in the Zoom breakout room 1 at 16:45

- Ideas for feedback
  - Was the internet connection OK? Did I speak clearly?
  - Were you able to follow the lecture?
  - How far did you make it in the exercise?
  - What was easy, what was difficult?
  - What did you like?
  - What can be improved?
  - etc etc.

Bayesian machine learning

Classical machine learning: supervised learning for regression

Common steps to fit a model to a given dataset $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$

1. Choose a model

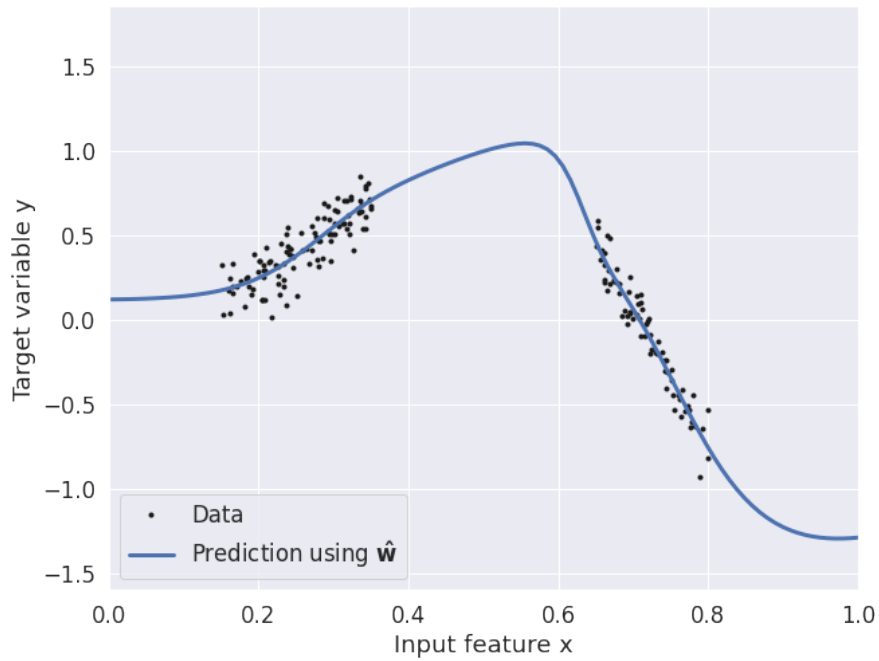
$$y_i = f(\boldsymbol{x}_i|\boldsymbol{w}) + e_i$$

2. Choose a loss function

$$\mathcal{L}(y, \hat{y}) = (y - \hat{y})^2$$

3. Find parameters $\boldsymbol{w}$ that minimizes the average loss $\mathcal{L}$ for the data set

$$\hat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} \sum_{i=1}^{N} [y - f(\boldsymbol{x}_i|\boldsymbol{w})]^2$$

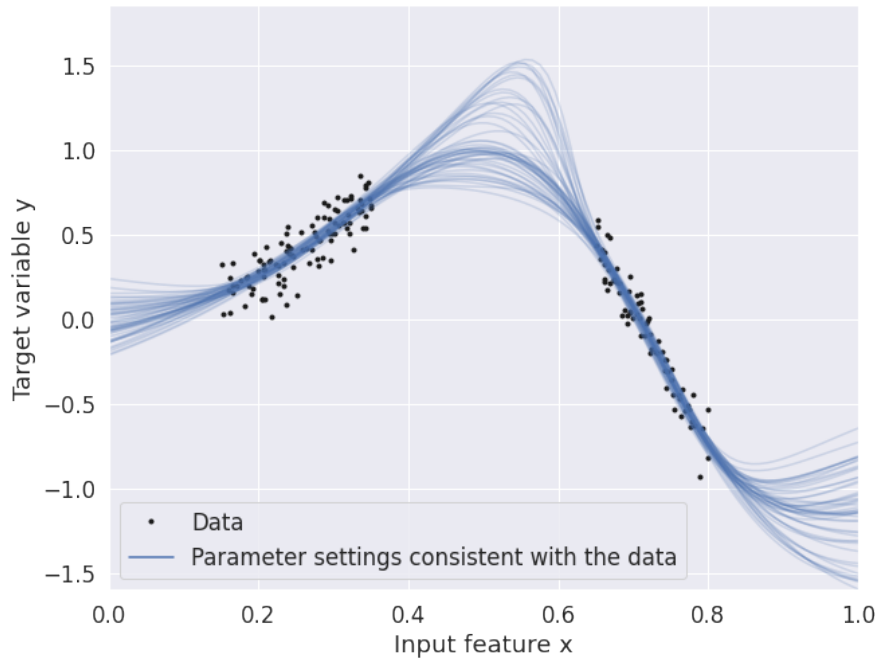


4. Make predictions for $\boldsymbol{x}^*$ using estimated parameters $\boldsymbol{w}$

$$y^* = f(\boldsymbol{x}^*|\hat{\boldsymbol{w}})$$

Model ambiguity due to finite data

- For a given model, there may be several sets of model parameters consistent with the data
- Model parameters $\hat{\boldsymbol{w}}_1$, $\hat{\boldsymbol{w}}_2$, and $\hat{\boldsymbol{w}}_3$ are all consistent with the data (as measured by training loss)
- Often many parameter settings consistent with data, but each can lead to very different predictions
- Classical machine learning: we choose *one* of these sets of parameters
- Bayesian machine learning: take *all* sets of model parameters consistent with the data into account



Bayesian inference and marginalization

- The *posterior distribution* $p(\boldsymbol{w}|\boldsymbol{y})$ measures how much weight to assign to each parameter set

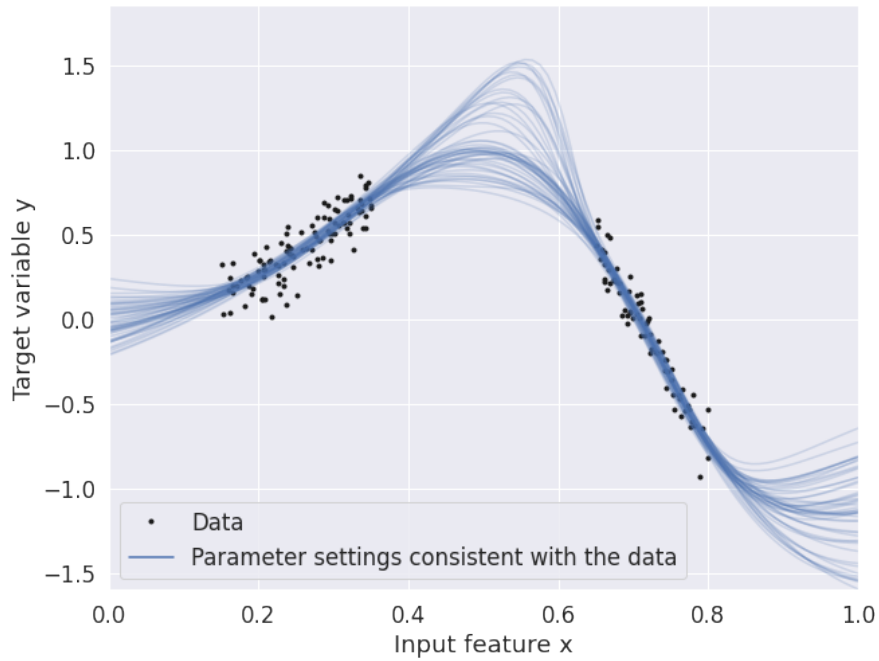
- Making predictions using a weighted average of all possible parameter sets

$$y^* = \sum_{i=1}^{M} f(\boldsymbol{x}^*|\boldsymbol{w}_i)p(\boldsymbol{w}_i|\boldsymbol{y}),$$

- Often, we have infinitely many parameter settings

$$y^* = \int f(\boldsymbol{x}^*|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{y})\mathrm{d}\boldsymbol{w}$$

- The process takes the uncertainty about the parameters into acoount and is called *marginalization*

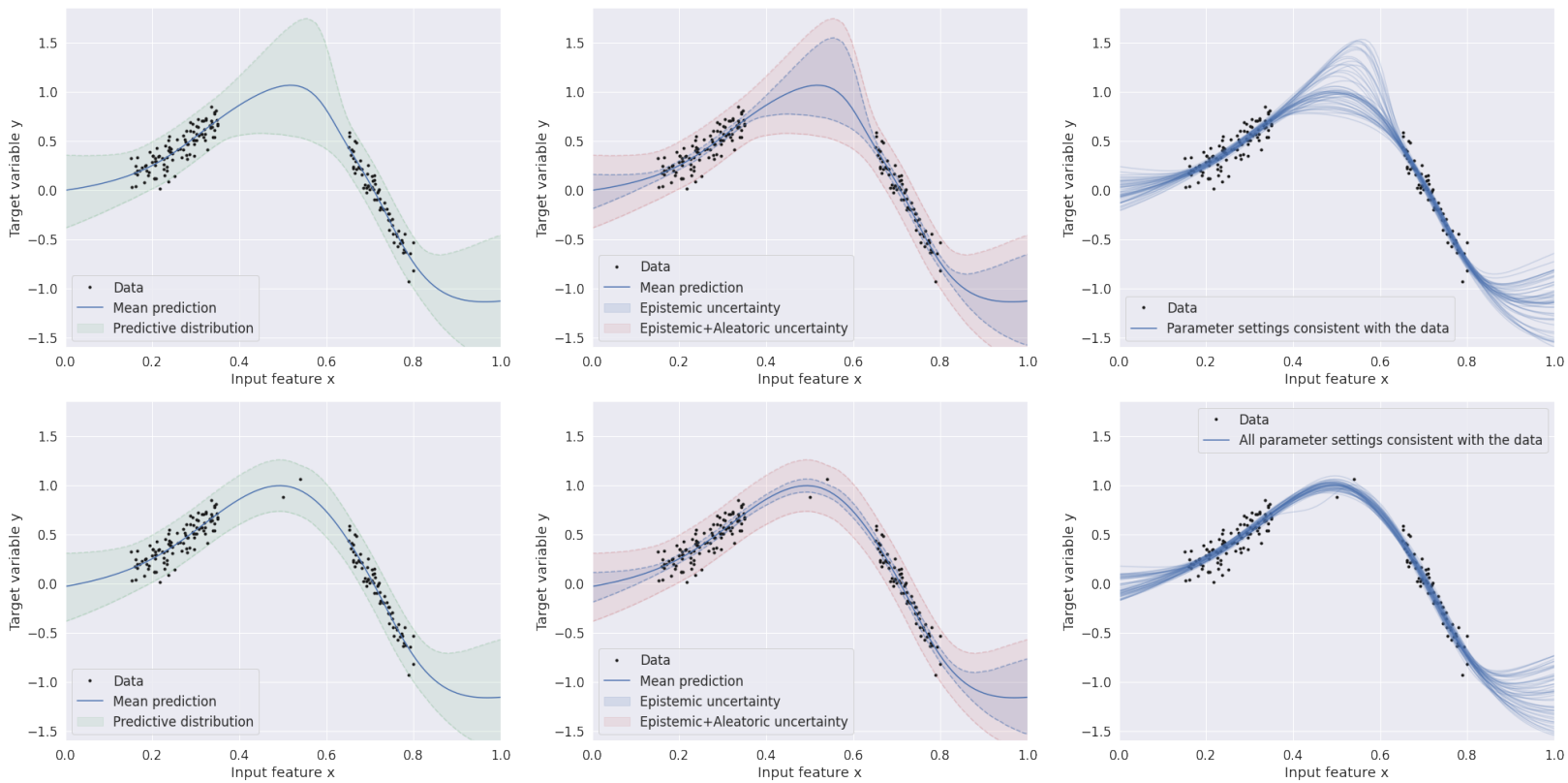


Uncertainty quantification

Two sources of uncertainty

1. *Epistemic uncertainty* is due to lack of knowledge (e.g. often due to a limited data set). Also sometimes called the *reducible* uncertainty.

2. *Aleatoric uncertainty* refers to the inherent randomness (e.g. measurement noise). Also sometimes called the *irreducible* uncertainty.



Bayesian machine learning

- In Bayesian methods, *all variables* (e.g. both parameters and data) are represented using *probability distributions*
- *Bayes' rule* provides a systematic way to combine data with prior knowledge

$$p(\boldsymbol{w}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y})}$$

- *Likelihood* $p(\boldsymbol{y}|\boldsymbol{w})$ - distribution of data $\boldsymbol{y}$ given a specific set of parameters $\boldsymbol{w}$
- *Prior* $p(\boldsymbol{w})$ - prior belief about parameters $\boldsymbol{w}$ before seeing any data
- *Posterior* $p(\boldsymbol{w}|\boldsymbol{y})$ - contains all knowledge about parameters after seeing data $\boldsymbol{y}$
- Why use priors?
  1. can encode domain knowledge
  2. can help prevent overfitting
  3. can generate artificial datasets from the model
- Why distributions rather than point estimates?
  1. Easy uncertainty quantification
  2. Avoid making predictions when uncertain is too large
  3. Better decision making

Probability notation and terminology

	Discrete distributions	**Continuous distributions**
Sample space	$\{0, 1\}$, $\{1, 2, 3, 4\}$, $\{\text{cat}, \text{dog}\}$	$\mathbb{R}$, $\mathbb{R}_+$, $[0, 1]$
Representation	Probability mass function (PMF) $0 \leq p(x) \leq 1$	Probability density functions (PDF) $p(x) \geq 0$
	$\sum_x p(x) = 1$	$\int p(x)\mathrm{d}x = 1$
		$p(x \in [a, b]) = \int_a^b p(x)\mathrm{d}x$
Mean	$\mathbb{E}[x] = \sum_x x\, p(x)$	$\mathbb{E}[x] = \int x\, p(x)\, \mathrm{d}x$
Variance	$\mathbb{V}[x] = \sum_x (x - \mathbb{E}[x])^2\ p(x)$	$\mathbb{V}[x] = \int (x - \mathbb{E}[x])^2\ p(x)\, \mathrm{d}x$
General expectations	$\mathbb{E}[f(x)] = \sum_x f(x)\, p(x)$	$\mathbb{E}[f(x)] = \int f(x)\, p(x)\, \mathrm{d}x$
Joint distribution	$p(x, y)$	$p(x, y)$
Conditional distribution	$p(x\|y) = \frac{p(x,y)}{p(y)}$	$p(x\|y) = \frac{p(x,y)}{p(y)}$
Sum rule (marginalization)	$p(x) = \sum_y p(x, y)$	$p(x) = \int p(x, y)\, \mathrm{d}y$
Product rule	$p(x, y) = p(x\|y)p(y)$	$p(x, y) = p(x\|y)p(y)$
Independence	$p(x, y) = p(x)p(y)$	$p(x, y) = p(x)p(y)$

Bayesian inference for the beta-binomial model

Motivating example: A/B testing

Your company's website has two ads. Ad A has been shown $N_A = 123$ times and generated $m_A = 12$ clicks, and Ad B has been shown $N_B = 145$ times and generated $N_B = 20$ clicks.

- What can we say about the click-rates for the two ads? Which one is best?
- We can calculate the sample click-rates

$$\hat{\mu}_A = \frac{m_A}{N_A} = \frac{12}{123} \approx 0.098, \qquad \hat{\mu}_B = \frac{m_B}{N_B} = \frac{20}{145} \approx 0.138$$

- Should we trust these estimates or ask for more data?
- What is the probability that the population click-rate for ad B is below 10%?
- What is the probability that ad B generates mores clicks than ad A?

We can answer such questions using the *beta-binomial model*

Goal for the rest of the lecture

- The *beta-binomial model* is a Bayesian model for estimating proportions, e.g.
  1. What proportion of users clicked the banner?
  2. What proportion of subject recovered after a certain medical treatment?
  3. What proportion of test images is correct classified?
  4. ...

- We will first look at the probabilistic building blocks
  1. Bernoulli distribution
  2. Binomial distribution
  3. Beta distribution

- Maximum likelihood inference

- Bayesian inference

- How does all this apply to A/B testing?

The Bernoulli distribution

Our first building block

- For a binary random variable $x \in \{0, 1\}$, where 1 represents "success"

$$P(x = 1|\mu) = \mu \qquad P(x = 0|\mu) = 1 - \mu,$$

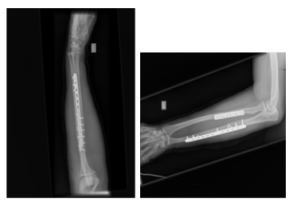
  where $0 \leq \mu \leq 1$

- *Bernoulli distribution* is a distribution over binary variables

$$\text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x}$$

Elbow

Forearm



- The mean of Bernoulli variable

$$\mathbb{E}[x] = \mu$$

- The variance is

$$\mathbb{V}[x] = \mu(1 - \mu)$$

Click me!

Binary variables and uncertainty quantification: example

- We model the user click behavior as Bernoulli distributed with probability μ, where $x = 1$ means click and $x = 0$ means no-click such that

$$P(\text{click}|\mu) = P(x = 1|\mu) = \mu$$

Click me!

- We know $0 \leq P(\text{click}|\mu) \leq 1$, but when are we most uncertain about the outcome?

a) $\mu = 1$?
b) $\mu = 0.5$?
c) $\mu = 0$?

The Binomial distribution

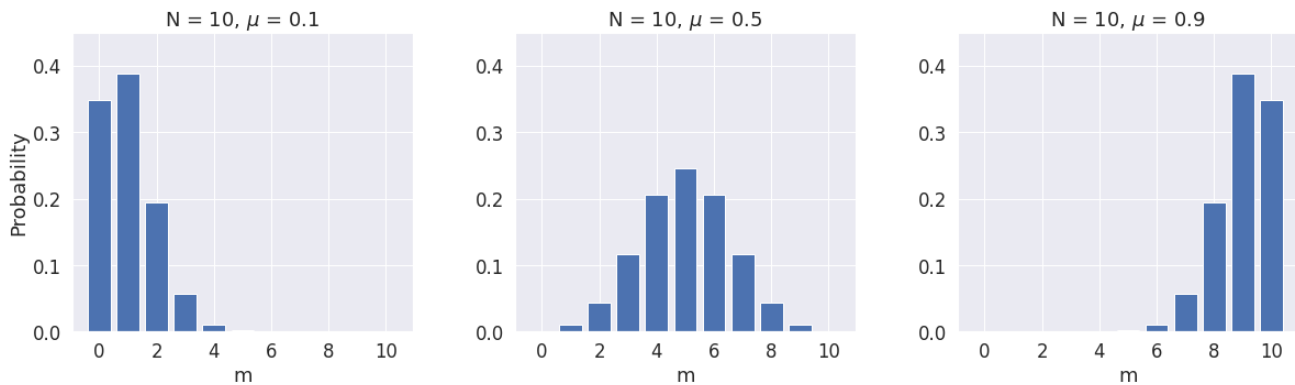
Modelling sequences of independent Bernoulli trials

- For a sequence of N independent Bernoulli trials $x_i \sim \text{Bern}(\mu)$ for $i = 1, \ldots, N$, the number of succeses $m = \sum x_i$ is said to follow a *Binomial distribution*

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

- Example: Suppose we flip a fair coin $N = 10$ times, the probability of getting $m = 4$ heads is given by

$$\text{Bin}(m = 4|N = 10, \mu = 0.5) = \binom{10}{4} 0.5^4 (1-0.5)^{10-4} \approx 0.21$$



- Calculating the mean of m

$$\mathbb{E}[m] = \mathbb{E}\left[\sum_{i=1}^{N} x_i\right] = \sum_{i=1}^{N} \mathbb{E}[x_i] = \sum_{i=1}^{N} \mu = N\mu$$

The Binomial distribution and maximum likelihood I

- Assume we collected a data set $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ of N independent Bernoulli trials with probability μ. How to estimate μ?

- Let $m = \sum x_i$ denote number of successes, then

$$P(m|\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \qquad \text{for} \quad m = 1, \ldots, N$$

- The *likelihood function* is defined as $\mathcal{L}(\mu) \equiv P(m|\mu)$

- The likelihood function measures: what is the probability of the observed data given the parameter value μ

- *Maximum likelihood:* We can estimate the parameters by maximing the likelihood function $\mathcal{L}$ wrt. μ

$$\hat{\mu}_{\text{ML}} \equiv \arg\max_{\mu} \mathcal{L}(\mu) = \arg\max_{\mu} \log \mathcal{L}(\mu) = m/N$$

- Rewriting

$$\mathcal{L}(\mu) = \log \left[\binom{N}{m} \mu^m (1-\mu)^{N-m} \right] = \log \left[\binom{N}{m} \right] + m \log(\mu) + (N-m) \log(1-\mu)$$

The Binomial distribution and maximum likelihood II

$$\mathcal{L}_{\mathcal{D}}(\mu) = \log\left[\binom{N}{m}\right] + m\log(\mu) + (N-m)\log(1-\mu)$$

- We *maximize* the log likelihood function by *differentiating, equating to zero and solving* for μ.

- Computing the derivative

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\mu}\log\mathcal{L}_{\mathcal{D}}(\mu) &= \frac{\mathrm{d}}{\mathrm{d}\mu}\log\left[\binom{N}{m}\right] + \frac{\mathrm{d}}{\mathrm{d}\mu}\log(\mu)m + \frac{\mathrm{d}}{\mathrm{d}\mu}\log(1-\mu)(N-m) \\ &= \frac{1}{\mu}m + \frac{1}{1-\mu}(N-m) = 0\end{aligned}$$

- Solving for μ yields the *maximim likelihood estimator*

$$\hat{\mu}_{\mathrm{ML}} = \frac{m}{N} = \frac{1}{N}\sum_{n=1}^{N} x_i$$

- The solution is equal to the empirical mean of the dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$

The Binomial distribution and maximum likelihood: Quiz

The *likelihood function* is defined as $\mathcal{L}(\mu) \equiv P(m|\mu)$ and the *maximum likelihood estimator* is

$$\hat{\mu}_{\text{ML}} = \arg\max_{\mu} \mathcal{L}(\mu) = \frac{m}{N}$$

Questions

1. The likelihood measures the probability of the specific parameter value given the observed data. True or False?

2. The likelihood measures the probability of the data given a specific parameter value. True or False?

3. The maximum likelihood estimator is the parameter value of μ that maximizes the likelihood. True or false?

4. What is the maximum likelihood estimate for μ if we observed $N = 10$ and $m = 2$?

The Binomial distribution and maximum likelihood III

- *Small data*: suppose an ad is shown $N = 3$ times and observe $m = 0$ clicks, then

$$\hat{\mu}_{\text{ML}} = 0/3 = 0$$

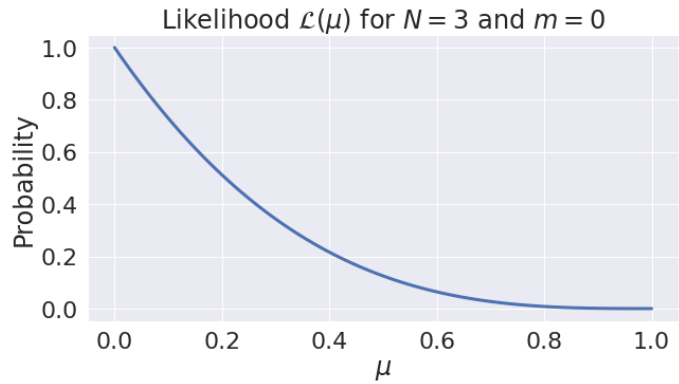
- What is the probability of exactly 0 clicks in the next 50 views?

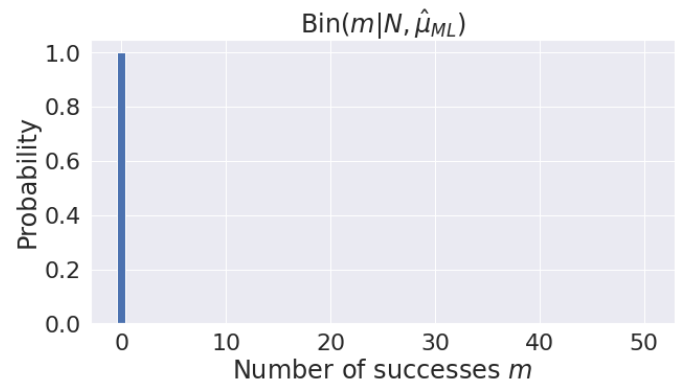
$$\begin{aligned} P[m = 0|\mu_{\text{ML}}] &= \text{Bin}(0|N = 50, \mu_{\text{ML}}) \\ &= \binom{50}{0} \mu_{\text{ML}}^0 (1 - \mu_{\text{ML}})^{50-0} \\ &= 1 \end{aligned}$$

- According to the model, we are *absolutely sure* that there will be exactly 0 heads in the next 50 views based on information from *only 3 observations*.

- Does this seem like a reasonable conclusion?

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$





Bayesian inference for μ

- Example continued: We observed $N = 3$ views and $m = 0$ clicks, then $\mu_{\text{ML}} = 0/3 = 0$. This is *overfitting* and this is a common problem when using maximum likelihood for small data sets.

- We can reduce this effect using Bayesian inference

$$p(\mu|m) = \frac{p(m|\mu)p(\mu)}{p(m)}$$

- *The prior* $p(\mu)$ represents our prior belief about μ **before** seeing the data

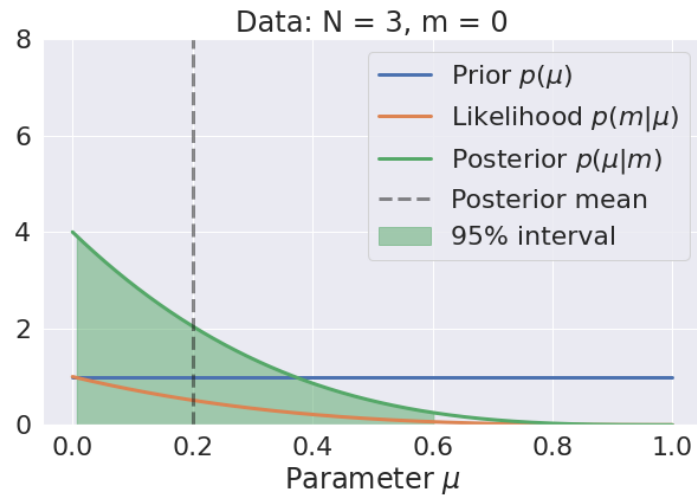
- The *likelihood* $p(m|\mu)$ represents our information from data

- After observing m, *the posterior distribution* $p(\mu|m)$ summarizes all our available information about μ

Bayesian inference in images

- Bayes' rule gives for a probability distribution for μ conditioned on the observed value for m

$$p(\mu|m) = \frac{p(m|\mu)p(\mu)}{p(m)}$$



- We can estimate μ using the mean of the posterior distribution

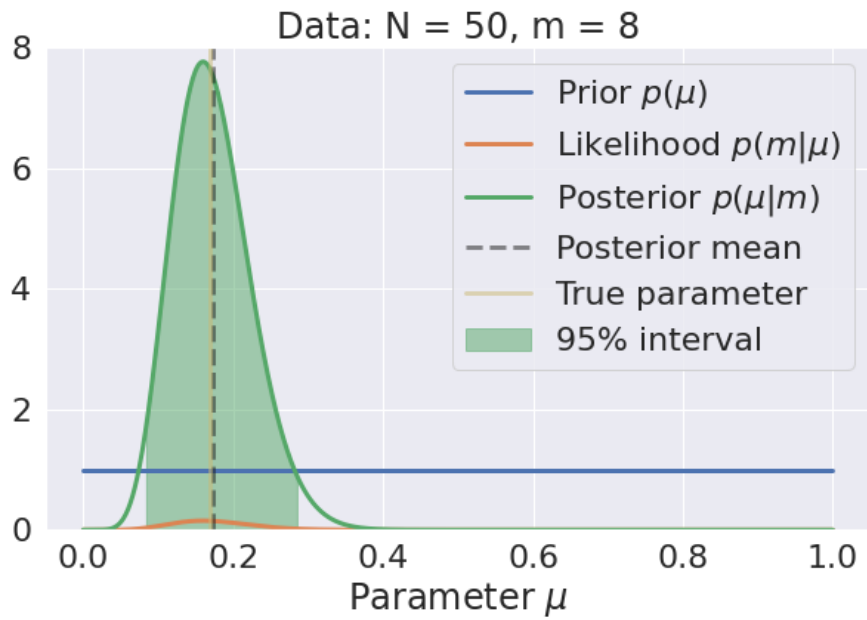
$$\mu_{\text{Bayes}} = \mathbb{E}\left[\mu|m\right] \equiv \int \mu\, p(\mu|m) \mathrm{d}\mu = 0.2$$

- and use *credibility intervals* of the posterior to quantify the uncertainty

$$P(\mu \in [0.01, 0.60]\,|m) = 0.95$$

What happens as we collect more data?

- Simulated data: $x_i \sim \text{Bern}(\mu_0)$ for $i = 1, \ldots, N$, where $\mu_0 = 0.17$



- The posterior concetrates as we collect more data

Bayesian analysis and probabilistic reasoning

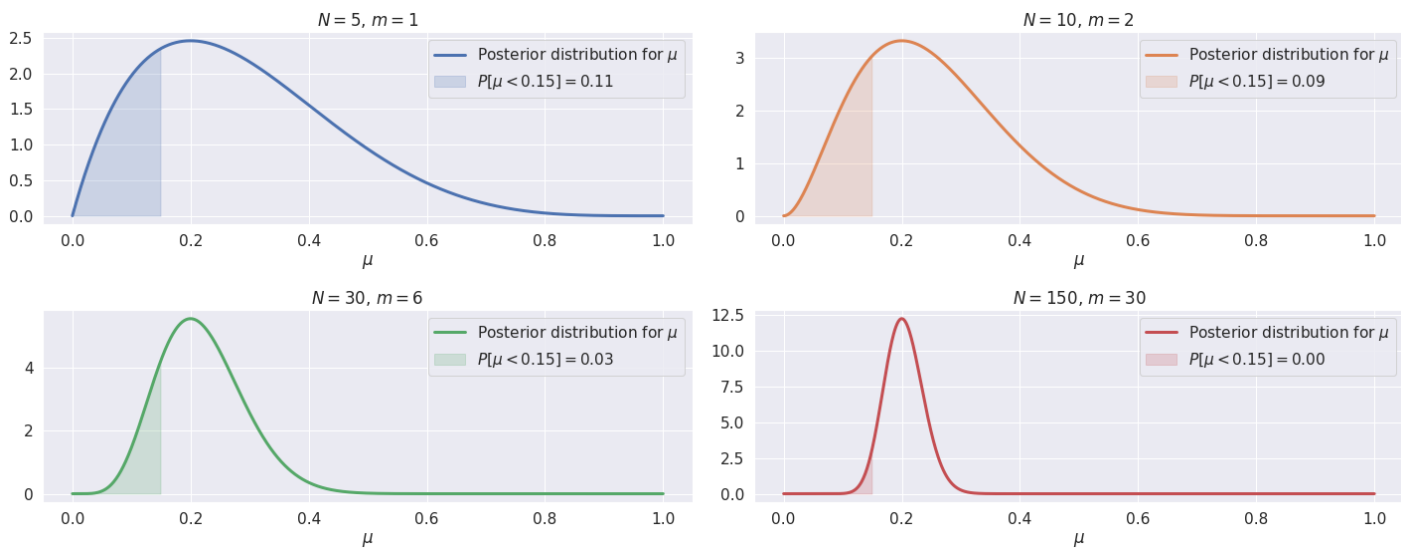
From point estimates to probability distributions

- Bayesian analysis provides a *probability distribution* summarizing our knowledge of μ rather than a *point estimate* like $\hat{\mu}_{\text{ML}}$

- Many common questions can be answered using *posterior summaries*, e.g. mode, mean, standard deviation, intervals, tail probabilities etc.

$$p(\mu < 0.15|m) = \int_0^{0.15} p(\mu|m)\,\mathrm{d}\mu$$

Examples



The prior distribution: how to choose?

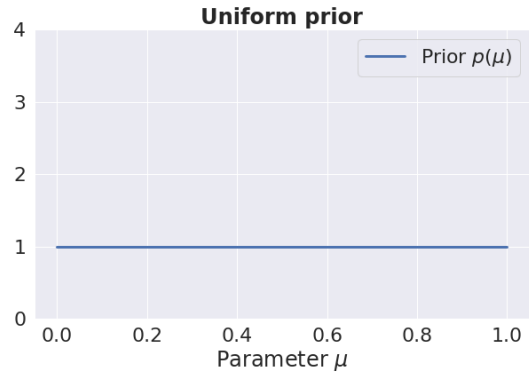
- The prior distribution $p(\mu)$ should reflect our prior assumptions before seeing the data

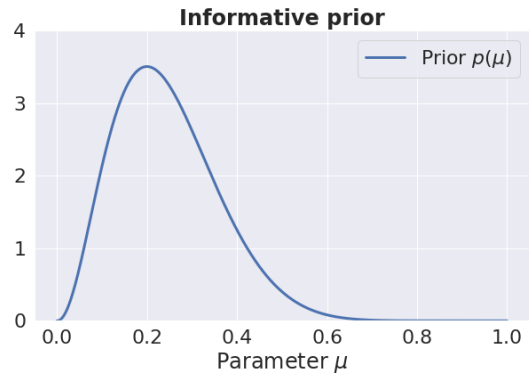
$$p(\mu|m) = \frac{p(m|\mu)p(\mu)}{p(m)} \propto p(m|\mu)p(\mu)$$

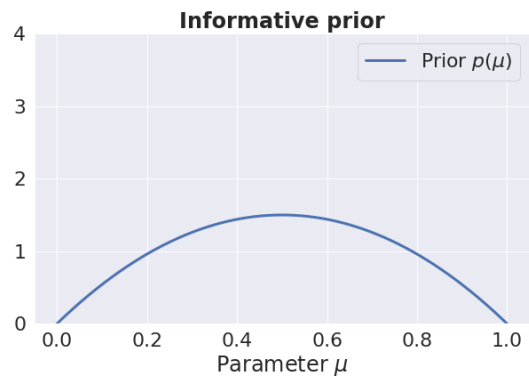
- Different types of priors
  - Uniform priors
  - Informative priors
  - Weakly informative priors
  - Priors for mathematical convinience

- Where does prior knowledge come from?
  - Previous experiments
  - Domain experts
  - Regularization

- Specifying a prior forces us to be explicit about our assumptions

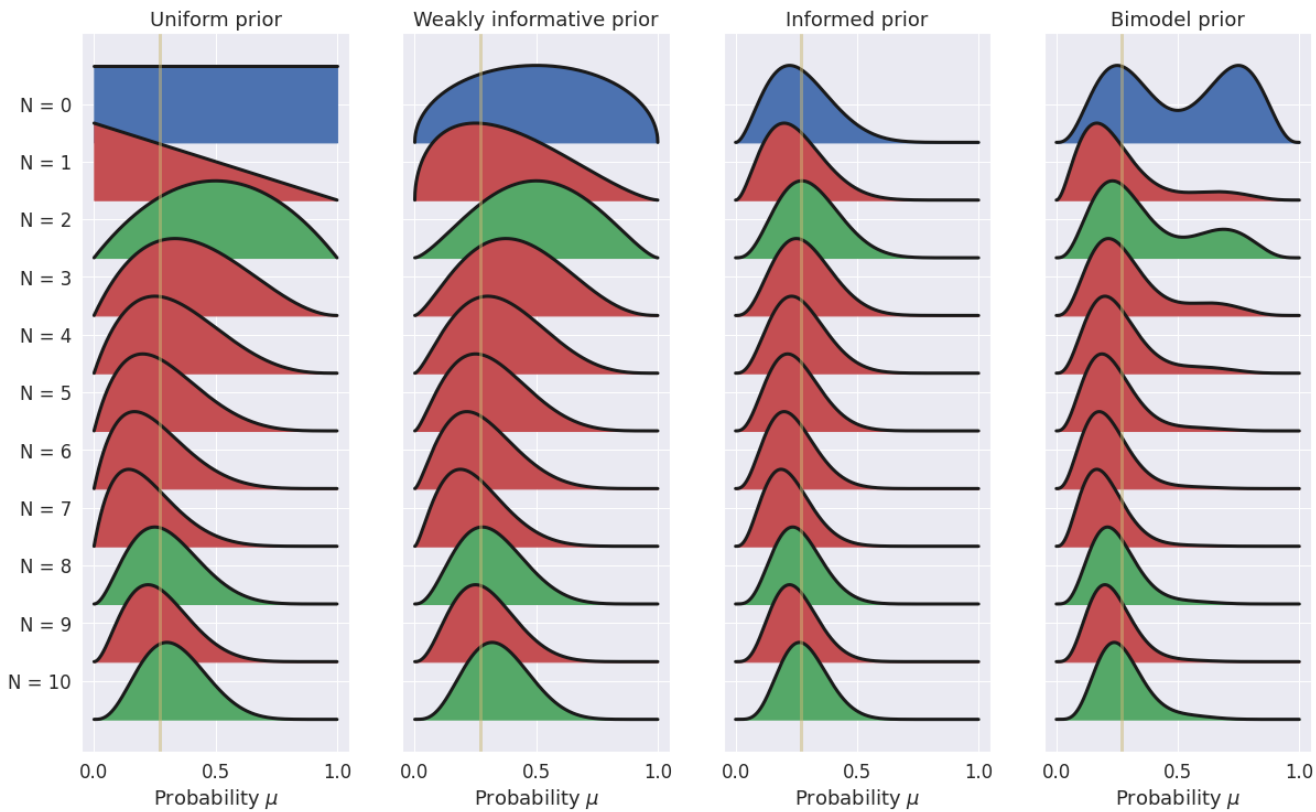






The effect of the prior distribution

- A Bayesian analysis starts with a *prior distribution* $p(\mu)$ representing our knowledge of μ *before* we observe any data



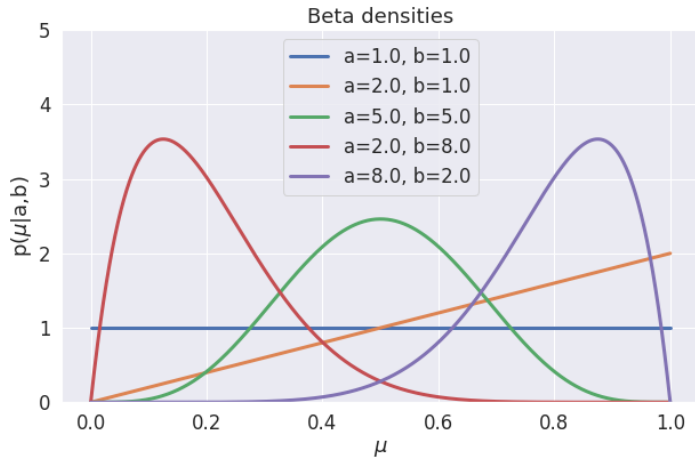
- The prior has a strong influence when the sample size is small, but its effect disappears when the sample size grows

Plots inspired by Rasmus Baath

The Beta distribution

A mathematically convinient prior distribution for μ

- *Beta distribution* is a family of distributions for a random variable $\mu \in [0, 1]$ in the unit interval



- The density of the Beta distribution is given by

$$p(\mu|a, b) = \frac{1}{B(a, b)} \mu^{a-1}(1 - \mu)^{b-1}$$

where $B(a, b)$ is a normalization constant given by

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$$

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x$$

- Mean and variance

$$\mathbb{E}[\mu] = \frac{a}{a + b}$$

$$\mathbb{V}[\mu] = \frac{ab}{(a + b)^2(a + b + 1)}$$

Deriving the analytical posterior for the Beta prior

- The beta distribution is a particular convinient choice for Binomial likelihoods

$$p(\mu|a_0, b_0) = \frac{1}{B(a_0, b_0)} \mu^{a_0 - 1}(1 - \mu)^{b_0 - 1}$$

- Bayes rule states

$$\begin{aligned} p(\mu|m) &= \frac{p(m|\mu)p(\mu)}{p(m)} \propto p(m|\mu)p(\mu) \\ &= \underbrace{\binom{N}{m} \mu^m (1 - \mu)^{N-m}}_{\text{Binomial PMF}} \underbrace{\frac{1}{B(a_0, b_0)} \mu^{a_0 - 1}(1 - \mu)^{b_0 - 1}}_{\text{Beta density}} \\ &\propto \mu^m (1 - \mu)^{N-m} \mu^{a_0 - 1}(1 - \mu)^{b_0 - 1} \\ &= \mu^{m + a_0 - 1}(1 - \mu)^{N - m + b_0 - 1} \\ &\propto \text{Beta}(\mu|m + a_0, N - m + b_0) \end{aligned}$$

- Key take-away: The posterior distribution is a another Beta distribution with parameters

$$\begin{aligned} a &= a_0 + m \\ b &= b_0 + N - m \end{aligned}$$

- The Beta distribution is said to be *conjugate* to the binomial distribution because the posterior is of the same *functional form* as the prior

The posterior mean

- The *key equations* for the beta-binomial model

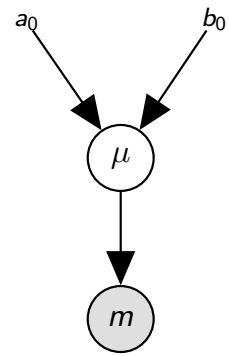
$$p(\mu) = \text{Beta}(\mu|a_0, b_0) \quad \text{(Prior)}$$

$$p(m|\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \quad \text{(Likelihood)}$$

$$p(\mu|m) = \text{Beta}(\mu|a_0 + m, b_0 + N - m) \quad \text{(Posterior)}$$

- The posterior mean is a compromise between the prior mean and the maximum likelihood solution

$$\mathbb{E}[\mu|m] = \frac{a}{a+b} = \frac{a_0 + m}{a_0 + b_0 + N}$$



- We can interpret a_0 and b_0 as *pseudo observations* of prior successes and failures, respectively

5 minutes exercise on your own

- The *key equations* for the beta-binomial model

$$p(\mu) = \text{Beta}(\mu|a_0, b_0) \qquad \text{(Prior)}$$

$$p(m|\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \qquad \text{(Likelihood)}$$

$$p(\mu|m) = \text{Beta}(\mu|a_0 + m, b_0 + N - m) \qquad \text{(Posterior)}$$

$$\mathbb{E}[\mu|m] = \frac{a_0 + m}{a_0 + b_0 + N} \qquad \text{(Posterior mean)}$$

Exercise
Assuming we have observed the following data $N = 20$ views and $m = 4$ click-rates and assume the prior is a Beta distribution with $a_0 = 2$ and $b_0 = 2$, compute ...

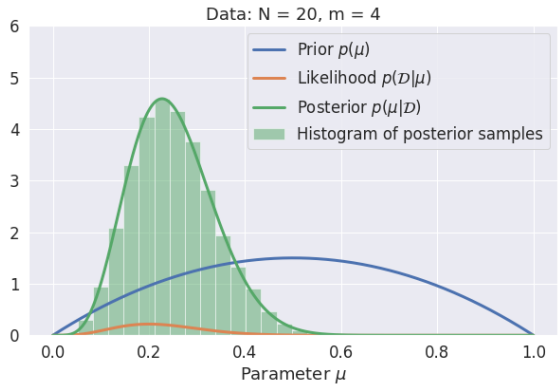
1. the prior mean
2. the parameters a and b for the posterior distribution
3. the posterior mean

Exercise - follow up I

Calculating posterior summaries in practice

- Suppose our posterior of interest is

$$p(\mu|m) = \text{Beta}(\mu|a = 6, b = 18)$$



- ... and our goal is to estimate $P(\mu < 0.15|m)$

- If we can generate samples from the posterior

$$\mu^{(i)} \sim \text{Beta}(\mu|a = 6, b = 18)$$

- We can estimate the probability by counting the fraction of samples below 0.15

$$P(\mu < 0.15|m) = \int_0^{0.15} p(\mu|m)\mathrm{d}\mu \approx \frac{1}{S}\sum_{i=1}^{S} \mathbb{I}\left[\mu^{(i)} < 0.15\right]$$

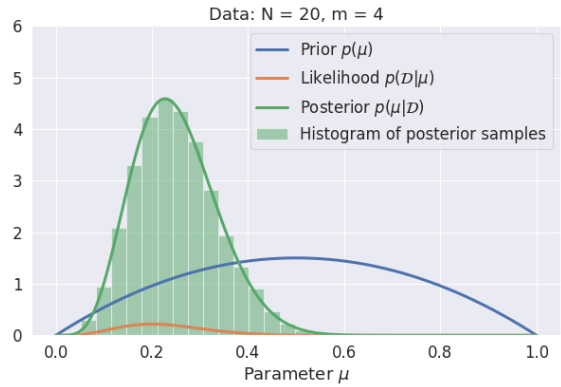
- Similarly, we can estimate the mean (variances, intervals, etc) using the sample

$$\mathbb{E}\left[g(\mu)|m\right] = \int g(\mu)\, p(\mu|m)\, \mathrm{d}\mu \approx \frac{1}{S}\sum_{i=1}^{S} g(\mu^{(i)})$$

Bayesian predictive distributions

Averaging over the posterior uncertainty

- How many clicks can we expect if we run the ad again and get $N^* = 50$ views?



- *Predictive likelihood*: what is the probability for observing m^* click for a given value of μ

$$p(m^*|N^*, \mu) = \text{Bin}(m^*|N^*, \mu)$$

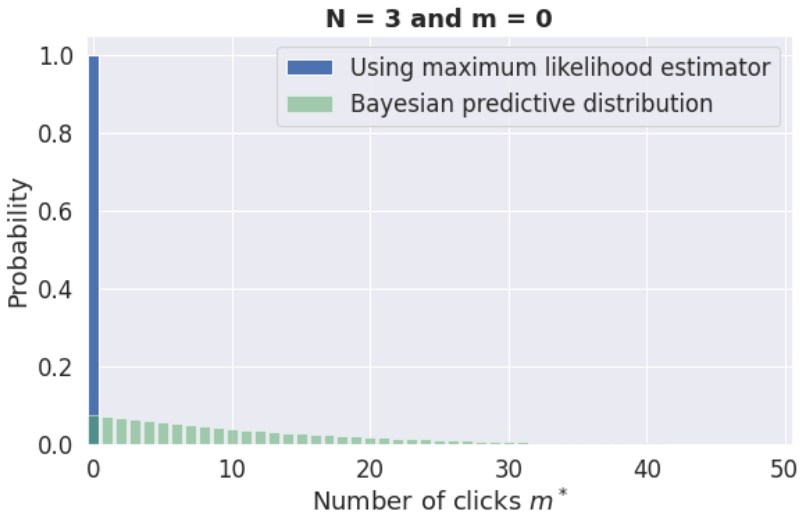
- The *predictive distribution* is obtained by *averaging over the posterior*

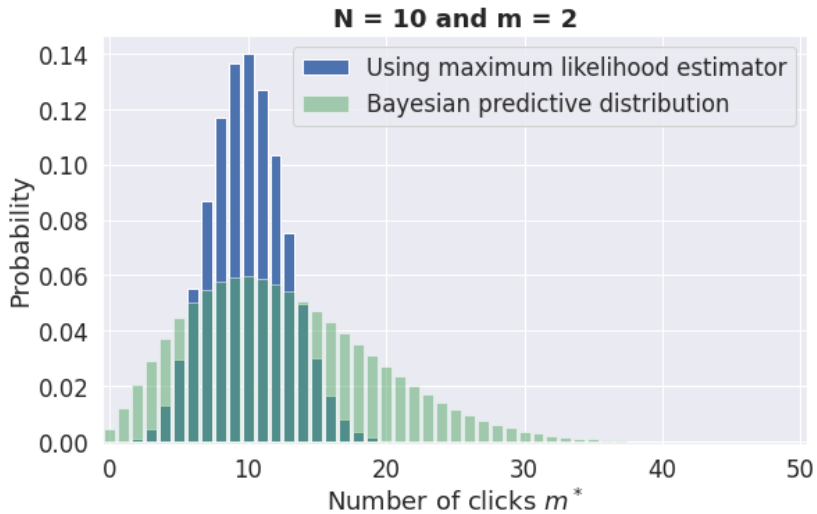
$$p(m^*|m) = \int p(m^*|N^*, \mu)p(\mu|m)\text{d}\mu$$

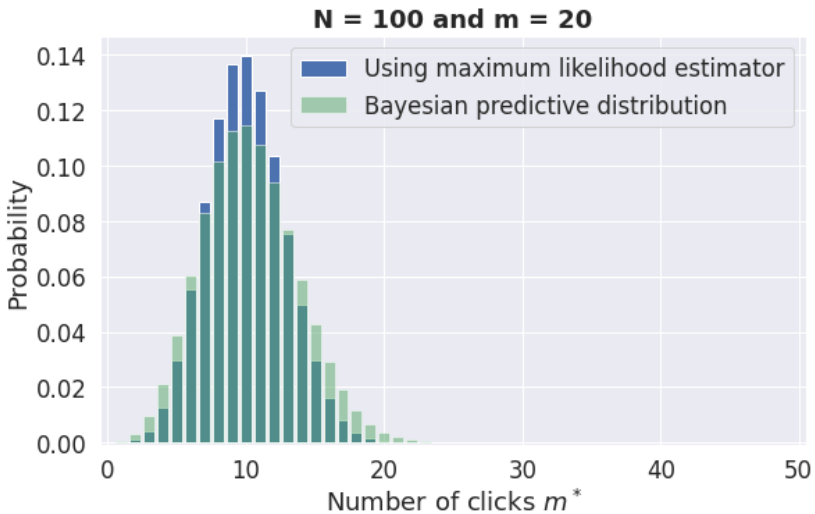
$$\approx \frac{1}{S}\sum_{i=1}^{S} \text{Bin}(m^*|N^*, \mu^{(i)})$$

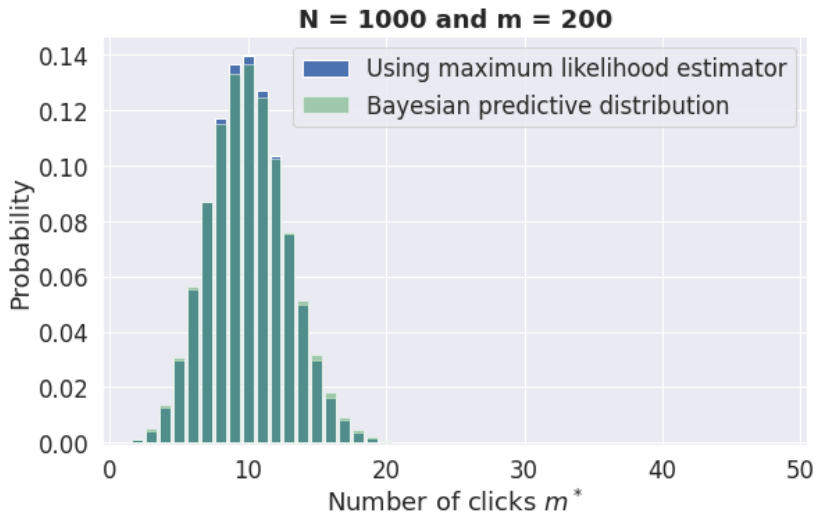
where $\mu^{(i)} \sim p(\mu|m)$ for $i = 1, \ldots, S$

Predictive distributions as we get more data









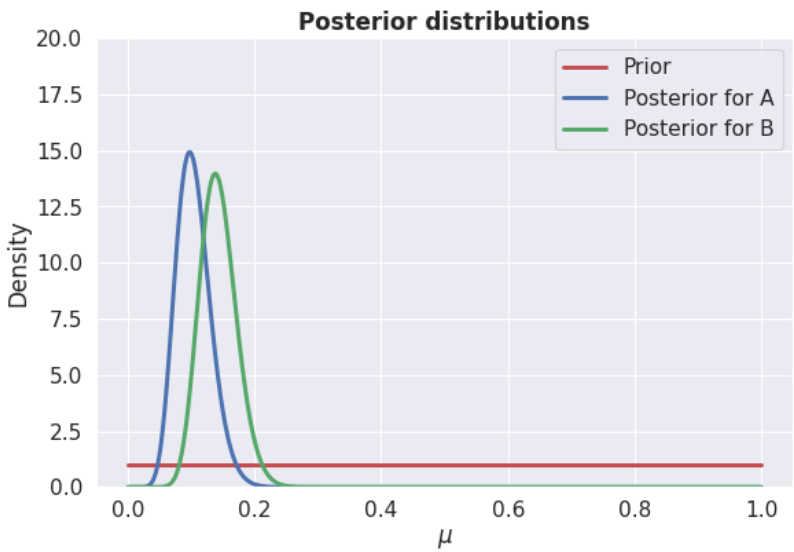
Posterior summaries for A/B testing example

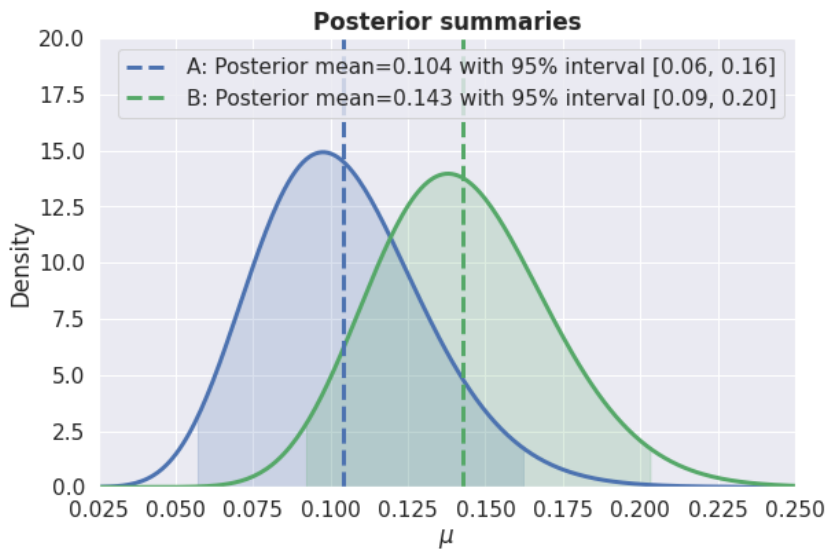
- Ad A has been shown $N_A = 123$ times and generated $m_A = 12$ clicks, and Ad B has been shown $N_B = 145$ times and generated $N_B = 20$ clicks.

- We will use uniform Beta-priors with $a_0 = b_0 = 1$ for both

- We compute posterior distribution for each ad (using $\mathcal{D}$ to denote observed data)

$$p(\mu_A|\mathcal{D}_A) = \text{Beta}(\mu_A|124, 13) \qquad p(\mu_B|\mathcal{D}_B) = \text{Beta}(\mu_B|146, 21)$$





Yes, but which one is better? A or B?

- Let's introduce the difference of the click rates

$$\mu_D = \mu_B - \mu_A$$

- We compute μ_D for each pair of samples

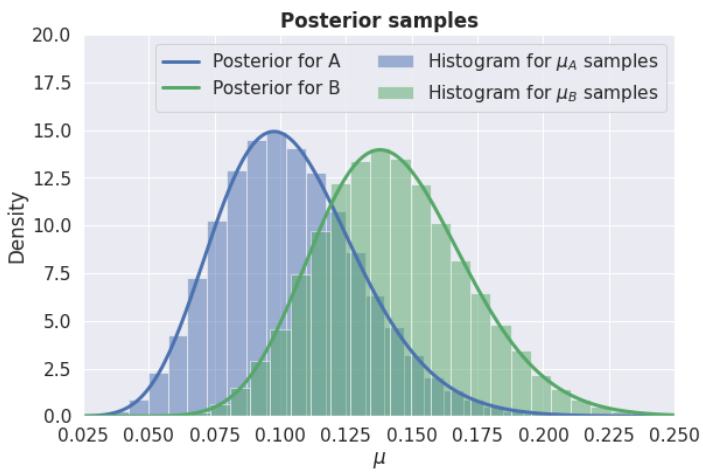
$$\mu_D^{(i)} = \mu_B^{(i)} - \mu_A^{(i)}$$

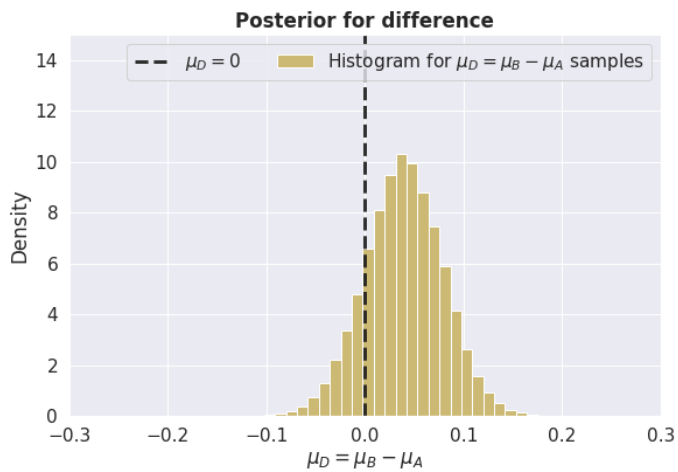
- Calculating posterior mean and credibility interval from posterior samples $\mu_D^{(i)}$

$$\mathbb{E}\left[\mu_D|\mathcal{D}\right] = 0.039 \qquad P\left(\mu_D \in [-0.038, 0.116]\,|\mathcal{D}\right) \approx 0.95$$

- Posterior probability that B is A than A

$$P\left(\mu_B > \mu_A|\mathcal{D}\right) = P\left(\mu_D > 0|\mathcal{D}\right) \approx 0.85$$





Main takeaways for today

1. Bayesian inference represents all variables using *probability distributions*
2. We update *from prior to posterior* belief using Bayes' rule

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})}$$

3. The posterior summarizes all information about μ after we observed data
4. The predictive distribution is obtained by *averaging over the uncertainty of the posterior*

$$p(m^*|\mathcal{D}) = \int P(m|N^*, \mu)p(\mu|\mathcal{D})\mathrm{d}\mu$$

5. Bayesian methods tend to produce wider distributions because the posterior uncertainty is taken into account
6. As we collect more and more data, the posterior distribution concetrates and become more and more independent of the prior
7. The Beta-binomial model is Bayesian approach for estimating proportions that uses the binomial distribution as likelihood and the beta distributon as prior

Introduction to the exercises

Intro to exercise

- Exercise is available as notebook on DTU Learn
- The purpose of the exercise is to get familiar with
  - Basic Bayesian terminology
  - The Beta-binomial model
  - Application to A/B testing
- A few exercises are optional
- Recall that the exercises will form the basis for the oral exam
- Feel free to collaborate with your peers using zoom - Work in breakout rooms in groups
- Ask for help: Use "ask for help"-button or- go to main room to ask for help and one of the teachers will come to your breakout room
- Feedbacks persons: Meet at 16:45 in the breakout room 1