Data Analytics
EEE 4774 & 6777

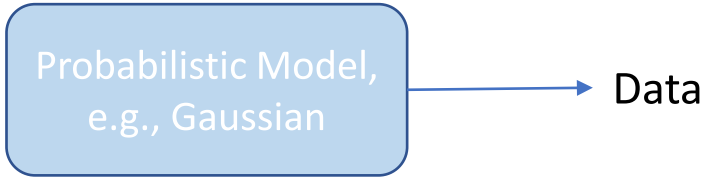
Module 2

Parameter Estimation

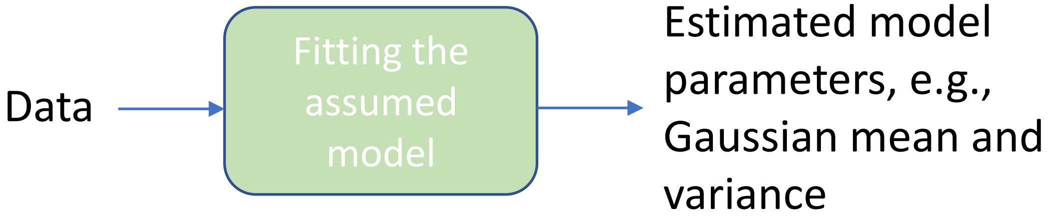
Spring 2022

Parameter Estimation for Model Fitting

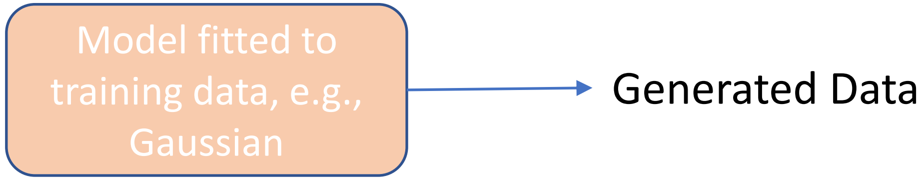
- Generative model assumption



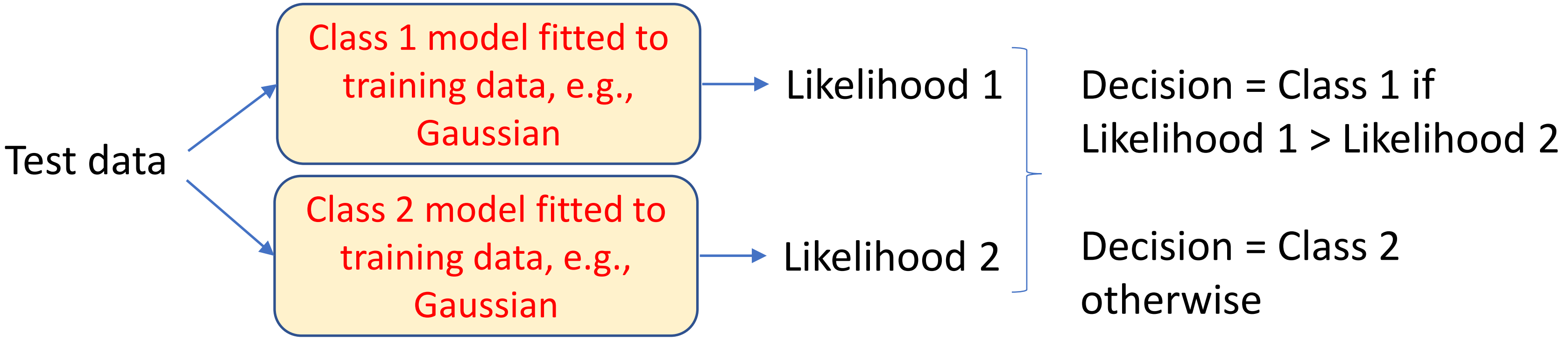
- Model inference/fitting from data



- Used for
  - Data generation
  - Missing value estimation

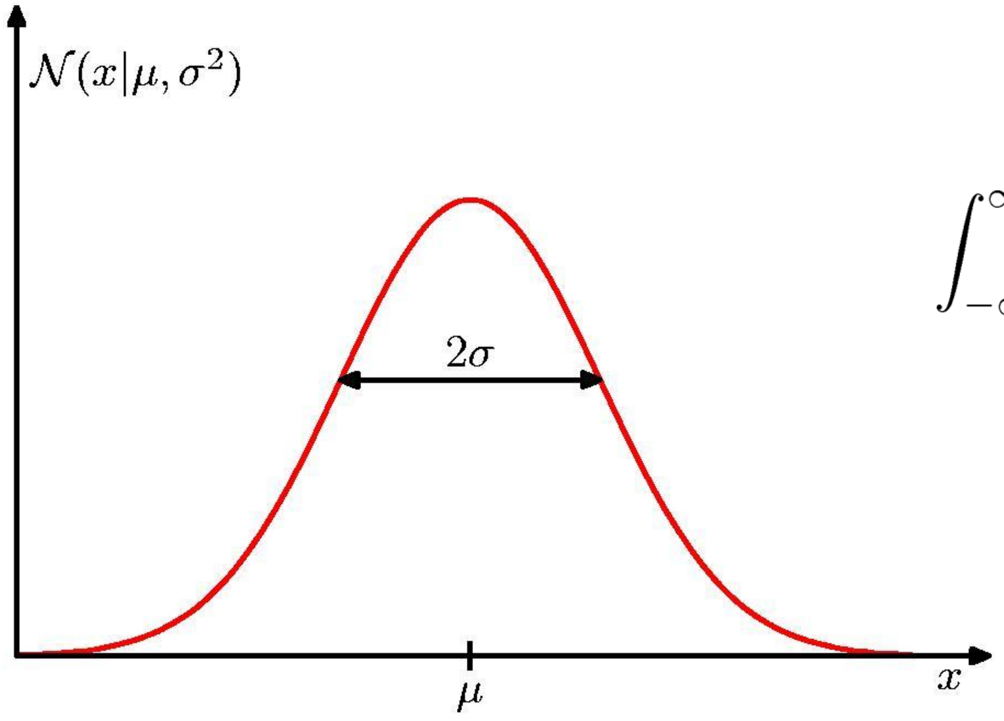


  - Classification



Gaussian Distribution

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



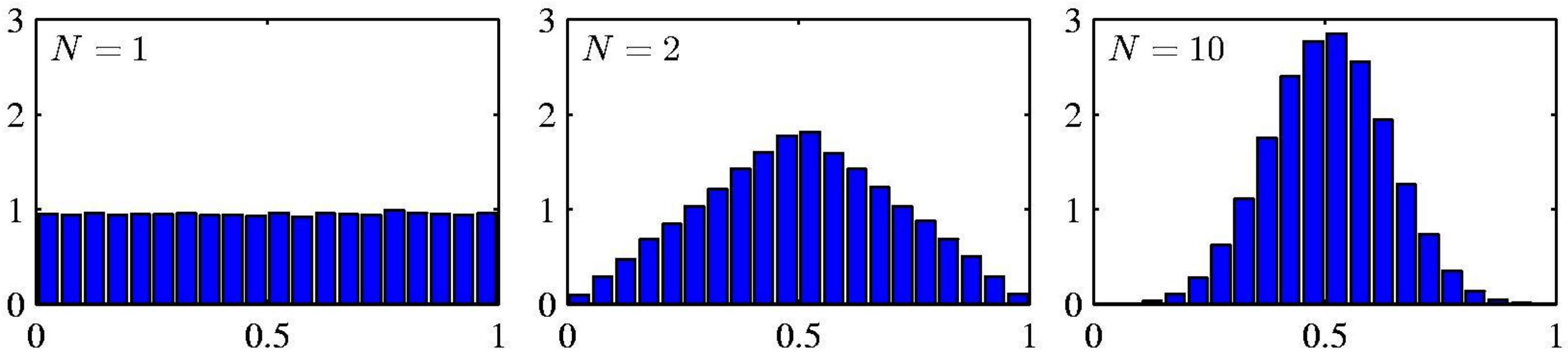
$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

Central Limit Theorem

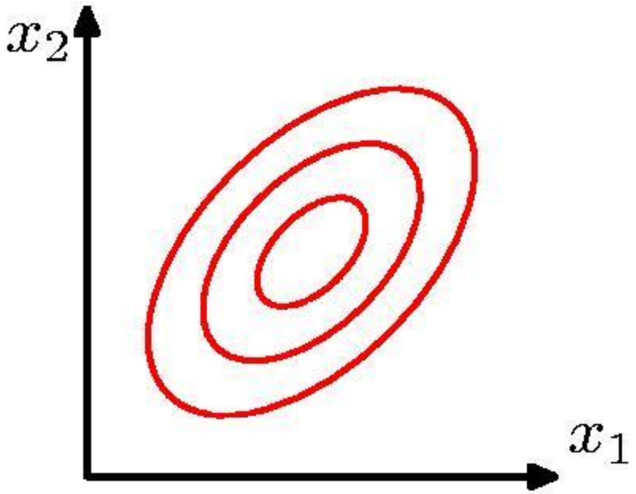
- The distribution of the sum of N i.i.d. random variables becomes increasingly Gaussian as N grows.

- Example: N uniform [0,1] random variables.



Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$



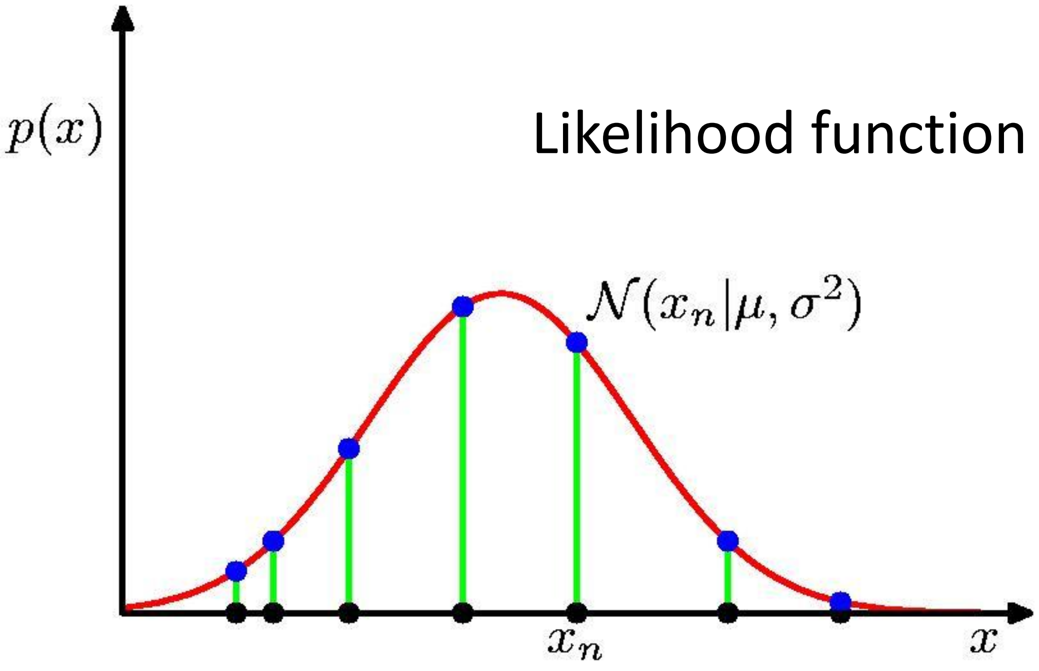
Gaussian Mean and Variance

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

Gaussian Parameter Estimation



Likelihood function

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

Maximum Likelihood (ML) Estimation

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

Properties of μ_{ML} and σ^2_{ML}

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2$$

$$\begin{aligned}\widetilde{\sigma}^2 &= \frac{N}{N-1}\sigma^2_{\mathrm{ML}} \\ &= \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2\end{aligned}$$

Parameter estimation: Bayesian for Gaussian

- Gaussian prior $p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right).$

- posterior $p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu).$ $p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$

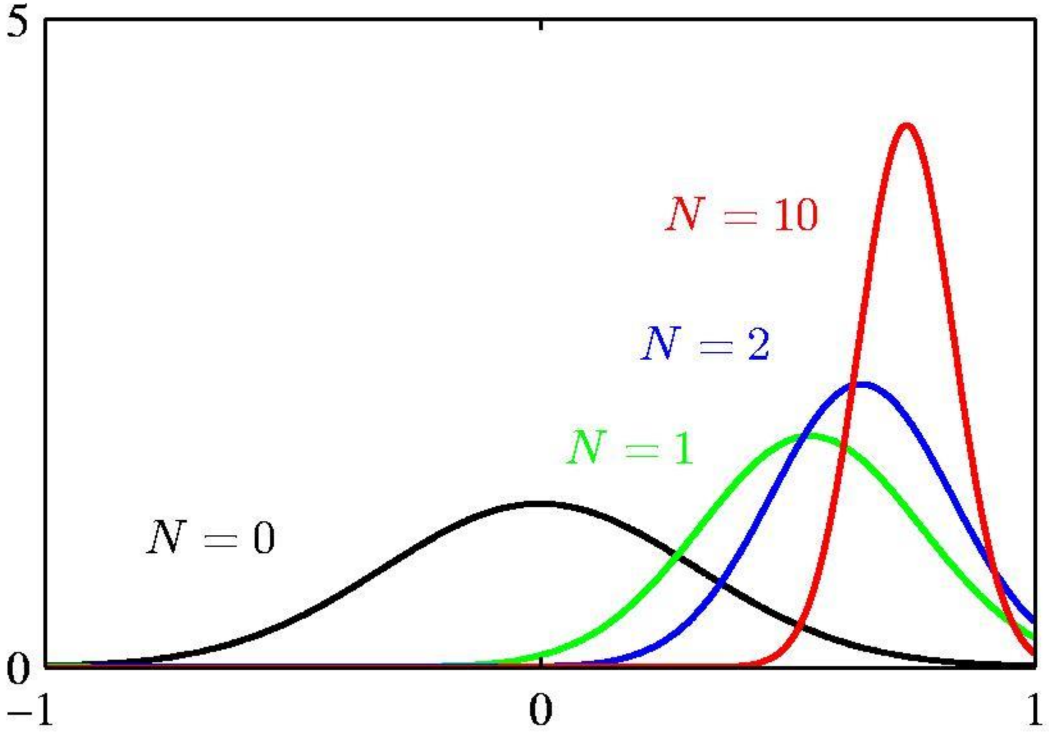
$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

	$N = 0$	$N \rightarrow \infty$
μ_N	μ_0	μ_{ML}
σ_N^2	σ_0^2	0

Parameter estimation: Bayesian for Gaussian

- Example: $p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$ for N = 0, 1, 2 and 10.
- True mean = 0.8



Python Example

- Dataset “adult.csv” includes information (attributes/features) of a number of people such as age, gender, occupation, and whether income less than or greater than 50K per year.
- Compute the mean and standard deviation of age for two groups of people:
  - Income <= 50K
  - Income > 50K
- Use the mean and standard deviation statistics to fit a Gaussian model to each group
- Compute the histogram to check if the Gaussian model assumptions are suitable.

Python Example

- Dataset “iris.csv” includes 4 measurements (attributes/features) from a number of iris plants.
- Each plant is from one of 3 classes.
- Fit a 4-dimensional multivariate Gaussian to data from each class.
- Compute the likelihood of each test instance under the 3 Gaussian models to make a classification decision.
- Choose the class (i.e., model) with the highest likelihood value as the class prediction for each test instance.
- Compute the accuracy by comparing the predicted class labels with the ground truth.

Bernoulli Distribution

- Coin flipping: heads=1, tails=0

$$p(x = 1|\mu) = \mu$$

- Bernoulli Distribution

$$\begin{aligned} \mathrm{Bern}(x|\mu) &= \mu^x (1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu) \end{aligned}$$

Binomial Distribution

- N coin flips:

$$p(m \text{ heads}|N, \mu)$$

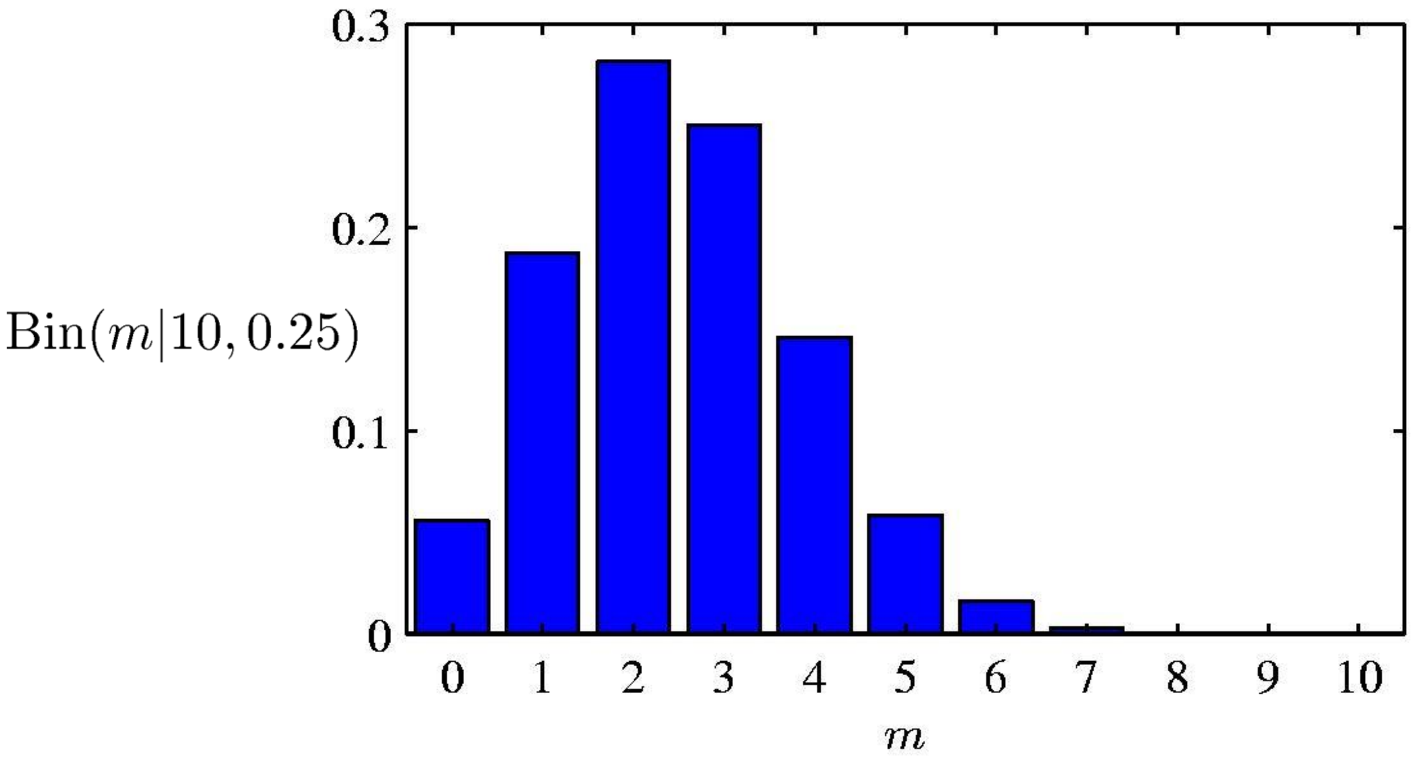
- Binomial Distribution

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

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\mathrm{Bin}(m|N, \mu) = N\mu$$

$$\mathrm{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \,\mathrm{Bin}(m|N, \mu) = N\mu(1-\mu)$$

Binomial Distribution



Parameter Estimation: ML for Bernoulli

Given: $\mathcal{D} = \{x_1, \ldots, x_N\}$, m heads (1), $N - m$ tails (0)

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n) \ln(1-\mu)\}$$

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{m}{N}$$

Parameter Estimation: ML for Bernoulli

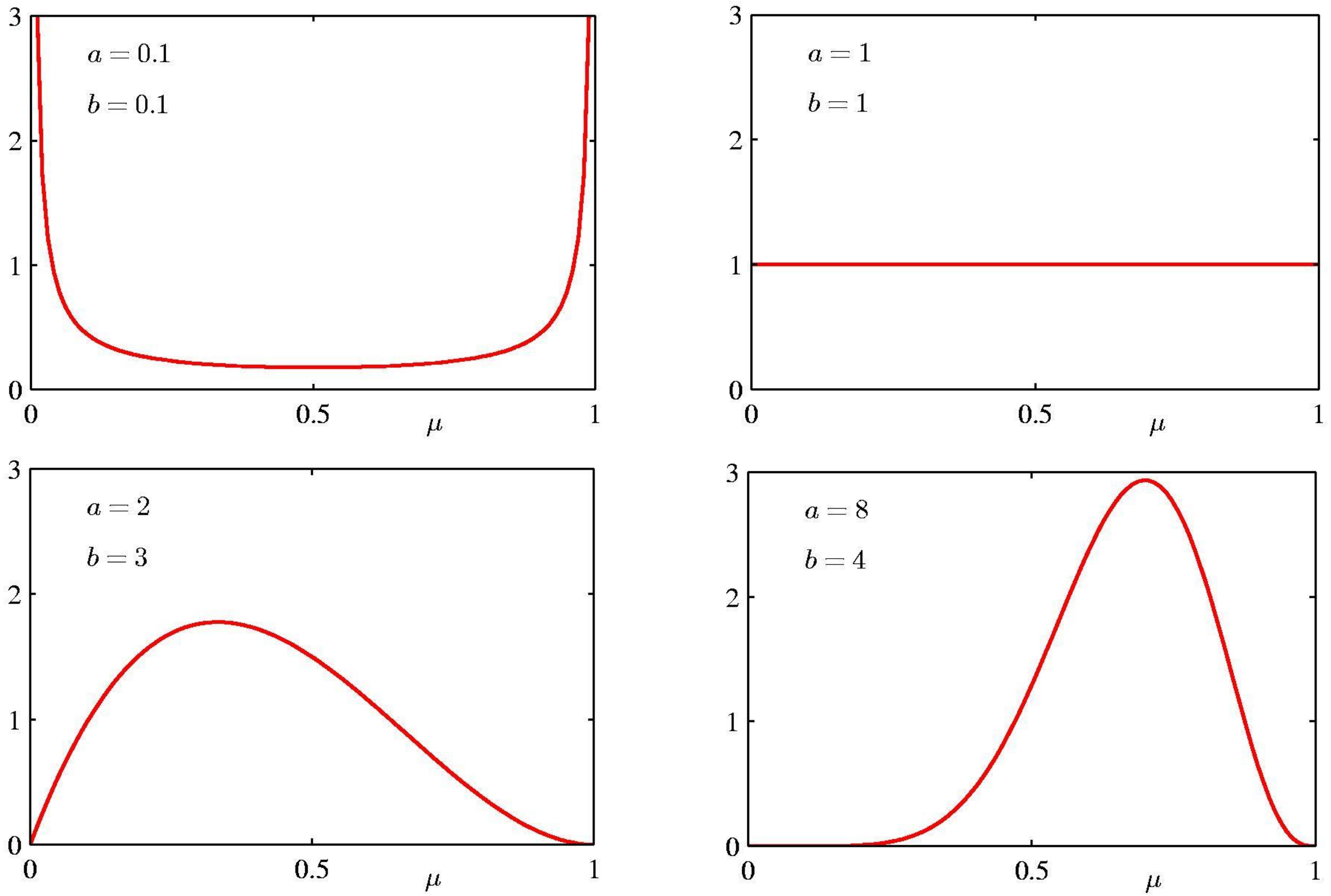
Example: $\mathcal{D} = \{1, 1, 1\} \rightarrow \mu_{\mathrm{ML}} = \frac{3}{3} = 1$

- Prediction: all future tosses will land heads up
- Overfitting to D

Beta Distribution

- Distribution over $\mu \in [0,1]$

$$
\begin{aligned}
\mathrm{Beta}(\mu|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
\mathbb{E}[\mu] &= \frac{a}{a+b} \\
\mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}
$$



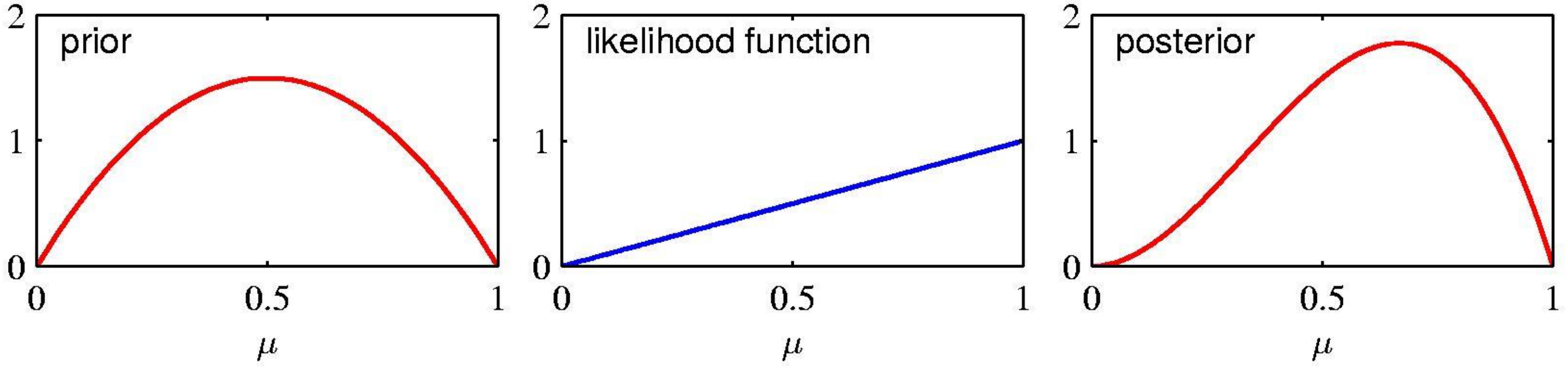
Parameter Estimation: Bayesian for Bernoulli

$$\begin{aligned} p(\mu|a_0, b_0, \mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0) \\ &= \left(\prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n} \right) \mathrm{Beta}(\mu|a_0, b_0) \\ &\propto \mu^{m+a_0-1}(1-\mu)^{(N-m)+b_0-1} \\ &\propto \mathrm{Beta}(\mu|a_N, b_N) \end{aligned}$$

$$a_N = a_0 + m \qquad b_N = b_0 + (N-m)$$

The **Beta** distribution provides the ***conjugate prior*** for the **Bernoulli** distribution.

Prior · Likelihood = Posterior



a=2, b=2
Beta

N=m=1

a=3, b=2
Beta

Properties of the Beta Posterior

As the size of the data set, N, increases

$$
\begin{aligned}
a_N &\rightarrow m \\
b_N &\rightarrow N - m \\
\mathbb{E}[\mu] &= \frac{a_N}{a_N + b_N} \rightarrow \frac{m}{N} = \mu_{\mathrm{ML}} \\
\mathrm{var}[\mu] &= \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \rightarrow 0
\end{aligned}
$$

Empirical Bayes

- In full Bayes a probability distribution is assumed/known for hyperparameters (a_0, b_0)

- In empirical Bayes, hyperparameters (a_0, b_0) are estimated from data
- Example:

Consider a baseball dataset with the total number of batting attempts A_i and the total number of hits H_i given for each player i.

The ML estimate for the probability of hit is given by the batting average H_i/A_i

To obtain a MAP estimate if you assume a Beta prior on the probability of hit, then you will either need to assume values for hyperparameters (a_0, b_0) such as $(100, 300)$ or estimate these values from the dataset.

Categorical Distribution

1-of-K coding scheme: $\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}}$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

$$\forall k : \mu_k \geqslant 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1$$

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}} = \boldsymbol{\mu}$$

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1$$

Multinomial Distribution

$$
\begin{aligned}
\mathrm{Mult}(m_1, m_2, \ldots, m_K|\boldsymbol{\mu}, N) &= \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k} \\
\mathbb{E}[m_k] &= N\mu_k \\
\mathrm{var}[m_k] &= N\mu_k(1-\mu_k) \\
\mathrm{cov}[m_j m_k] &= -N\mu_j\mu_k
\end{aligned}
$$

Parameter estimation: ML for Multinomial

- Given: $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k}$$

- Ensure $\sum_k \mu_k = 1$, use a Lagrange multiplier,

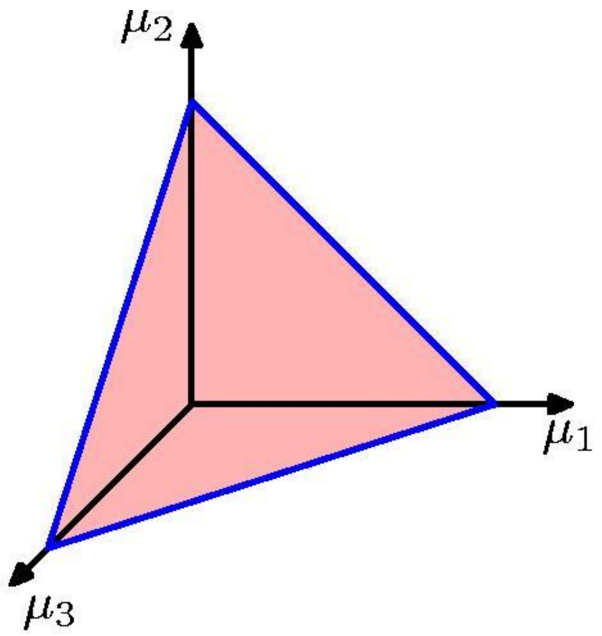
$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

$$\mu_k = -m_k/\lambda \qquad \mu_k^{\mathrm{ML}} = \frac{m_k}{N}$$

Dirichlet Distribution

$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

$$\alpha_0 = \sum_{k=1}^{K} \alpha_k$$



Conjugate prior for the **multinomial** distribution.

Parameter estimation: Bayesian for Multinomial

$$p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1}$$

$$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) &= \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}+\mathbf{m}) \\ &= \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\cdots\Gamma(\alpha_K+m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1} \end{aligned}$$

Exponential Family

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

- where η is the *natural parameter* and

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \, \mathrm{d}\mathbf{x} = 1$$

- so g(η) can be interpreted as a normalization coefficient.

- For any member of the exponential family, there exists a ***conjugate prior,*** which makes the posterior the same distribution as itself

Nonparametric Methods

- Parametric distribution models are restricted to specific forms,
  - which may not always be suitable;
  - for example, consider modelling a multimodal distribution with a single, unimodal model.
- Nonparametric approaches
  - make few assumptions about the overall shape of the distribution being modelled.

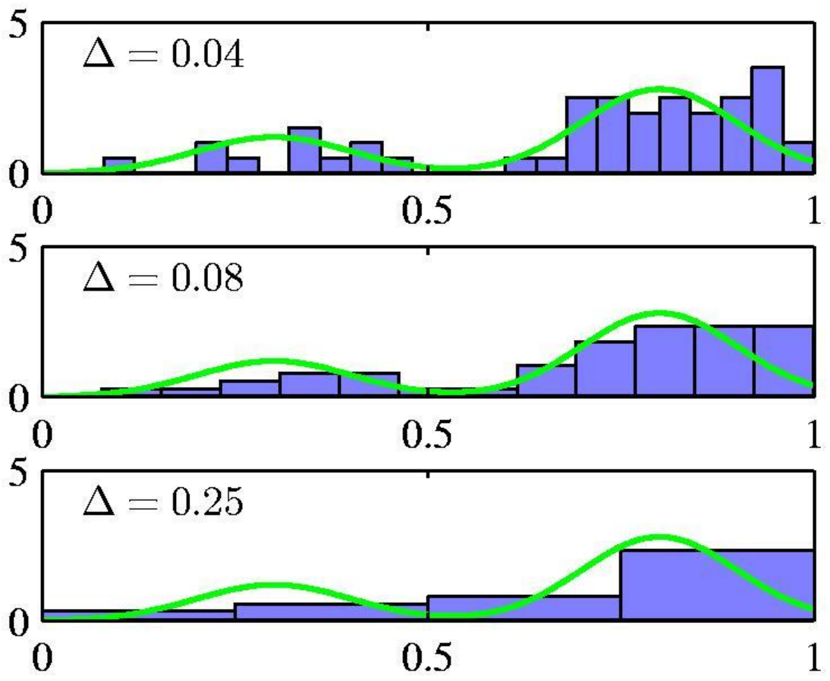
Histogram

Histogram methods partition the data space into distinct bins with widths Δ_i and count the number of observations, n_i, in each bin.

$$p_i = \frac{n_i}{N\Delta_i}$$

- Often, the same width is used for all bins, $\Delta_i = \Delta$.

- Δ acts as a smoothing parameter.



- In a D-dimensional space, using M bins in each dimension will require M^D bins!

Curse of dimensionality!

Kernel Methods

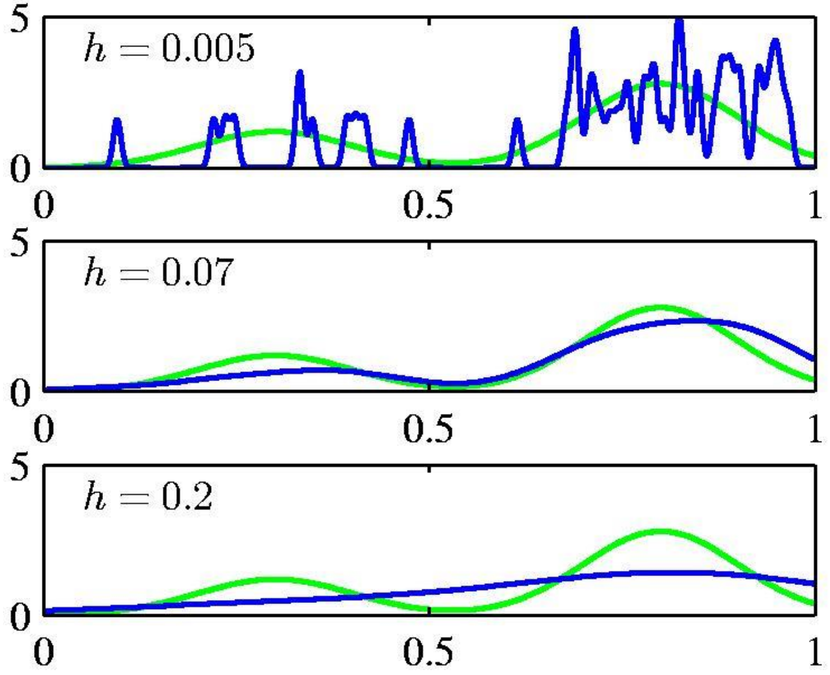
To avoid discontinuities in p(x), use a smooth kernel, e.g. a Gaussian

$$p(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{(2\pi h^2)^{D/2}} \exp\left\{-\frac{\|\mathbf{x}-\mathbf{x}_n\|^2}{2h^2}\right\}$$

Any kernel such that

$$\begin{aligned} k(\mathbf{u}) &\geqslant 0, \\ \int k(\mathbf{u})\,\mathrm{d}\mathbf{u} &= 1 \end{aligned}$$

will work.



h acts as a smoother.