Discrete Probability Distributions

Sargur N. Srihari

Binary Variables

Bernoulli, Binomial and Beta

Bernoulli Distribution

- Expresses distribution of Single binary-valued random variable $x \in \{0,1\}$
- Probability of $x=1$ is denoted by parameter μ, *i.e.,*

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

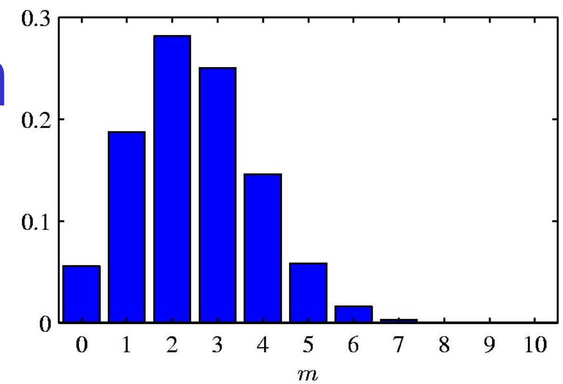
  $$p(x=0|\mu)=1-\mu$$
- Probability distribution has the form $\mathrm{Bern}(x|\mu)=\mu^x(1-\mu)^{1-x}$
- Mean is shown to be $E[x]=\mu$
- Variance is $Var[x]=\mu(1-\mu)$
- Likelihood of n observations independently drawn from $p(x|\mu)$ is

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

  $$\ln p(D|\mu)=\sum_{n=1}^{N} \ln p(x_n|\mu)=\sum_{n=1}^{N}\{x_n \ln\mu+(1-x_n)\ln(1-\mu)\}$$
- Maximum likelihood estimator
  - obtained by setting derivative of $\ln p(D|\mu)$ wrt m equal to zero is $\mu_{ML}=\frac{1}{N}\sum_{n=1}^{N} x_n$
- If no of observations of $x=1$ is m then $\mu_{ML}=m/N$

Jacob Bernoulli
1654-1705

Binomial Distribution



Histogram of Binomial for $N=10$ and $m=0.25$

- Related to Bernoulli distribution
- Expresses Distribution of m
  - No of observations for which $x=1$
- It is proportional to $Bern(x|\mu)$
- Add up all ways of obtaining heads

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

Binomial Coefficients:

$$\binom{N}{m} = \frac{N!}{m!(N-m)!}$$

- Mean and Variance are

$$E[m] = \sum_{m=0}^{N} m Bin(m \mid N, \mu) = N\mu$$

$$Var[m] = N\mu(1-\mu)$$

Beta Distribution

- Beta distribution

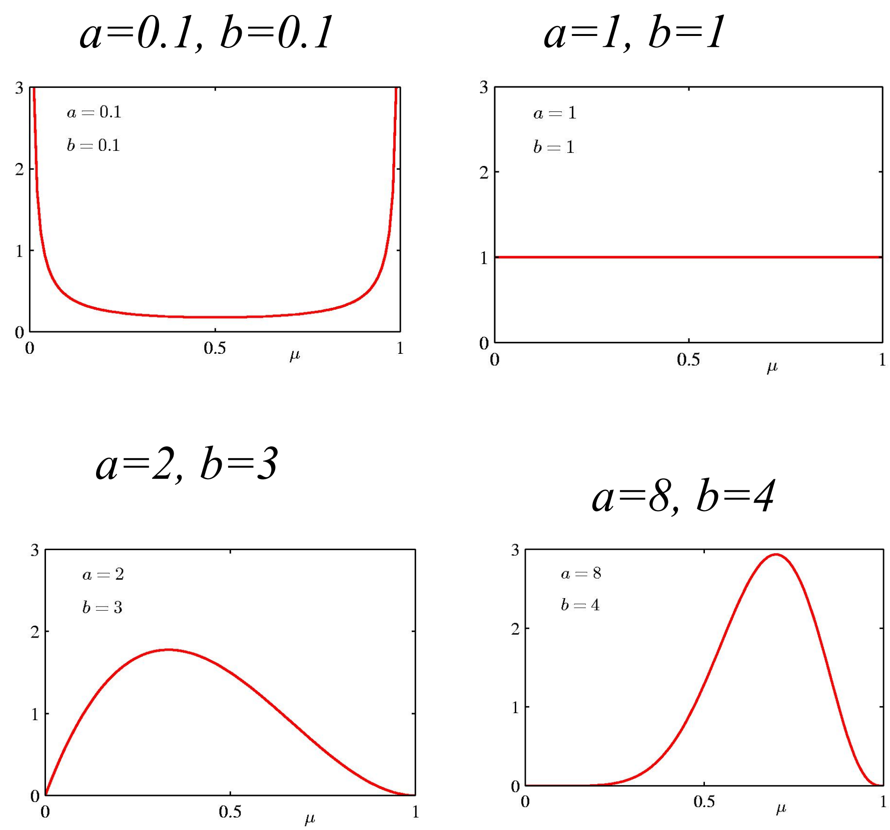
$$Beta(\mu \mid a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1}$$

- Where the Gamma function is defined as

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\, du$$

- a and b are hyperparameters that control distribution of parameter μ
- Mean and Variance

$$E[\mu] = \frac{a}{a+b} \qquad \mathrm{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$



Beta distribution as function of μ
For values of hyperparameters a and b

Bayesian Inference with Beta

- MLE of μ in Bernoulli is fraction of observations with $x=1$
  - Severely over-fitted for small data sets
- Likelihood function takes products of factors of the form $\mu^x(1-\mu)^{(1-x)}$
- If prior distribution of μ is chosen to be proportional to powers of μ and $1-\mu$, posterior will have same functional form as the prior
  - Called conjugacy
- Beta has form suitable for a prior distribution of $p(\mu)$

Bayesian Inference with Beta

- Posterior obtained by multiplying beta prior with binomial likelihood yields

$$p(\mu \mid m,l,a,b) \,\alpha\, \mu^{m+a-1}(1-\mu)^{l+b-1}$$

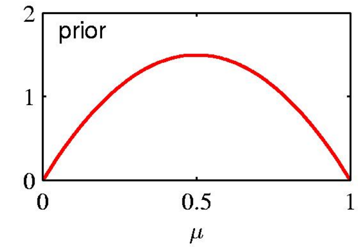
  - where $l=N-m$, which is no of tails
  - m is no of heads
- It is another beta distribution

$$p(\mu \mid m,l,a,b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}$$

  - Effectively increase value of a by m and b by l
  - As number of observations increases distribution becomes more peaked

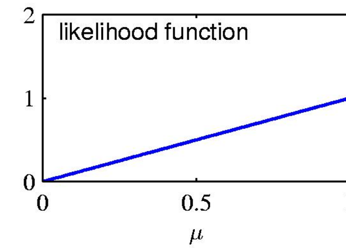
Illustration of one step in process

$a=2,\ b=2$



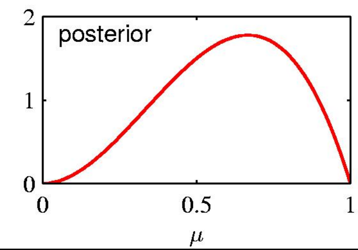
$p(\mu)$

$N=m=1$, with $x=1$



$p(x=1/\mu)=$ $\mu^1(1-\mu)^0$

$a=3,\ b=2$



$p(\mu/x=1)$

Predicting next trial outcome

- Need predictive distribution of x given observed D
  - From sum and products rule

$$p(x=1\,|\,D)=\int_0^1 p(x=1,\mu\,|\,D)d\mu=\int_0^1 p(x=1\,|\,\mu)p(\mu\,|\,D)d\mu=$$

$$=\int_0^1 \mu p(\mu\,|\,D)d\mu=E[\mu\,|\,D]$$

- Expected value of the posterior distribution can be shown to be

$$p(x=1\,|\,D)=\frac{m+a}{m+a+l+b}$$

  - Which is fraction of observations (both fictitious and real) that correspond to $x=1$
- Maximum likelihood and Bayesian results agree in the limit of infinite observations
  - On average uncertainty (variance) decreases with observed data

Summary of Binary Distributions

- Single Binary variable distribution is represented by Bernoulli
- Binomial is related to Bernoulli
  - Expresses distribution of number of occurrences of either 1 or 0 in N trials
- Beta distribution is a conjugate prior for Bernoulli
  - Both have the same functional form

Sample Matlab Code

Probability Distributions

- Binomial Distribution:
  - Probability Density Function : Y = binopdf (X,N,P)

  returns the binomial probability density function with parameters N and P at the values in X.

  - Random Number Generator: R = binornd (N,P,MM,NN)

  returns n MM-by-NN matrix of random numbers chosen from a binomial distribution with parameters N and P

- Beta Distribution
  - Probability Density Function : Y = betapdf (X,A,B)

  returns the beta probability density function with parameters A and B at the values in X.

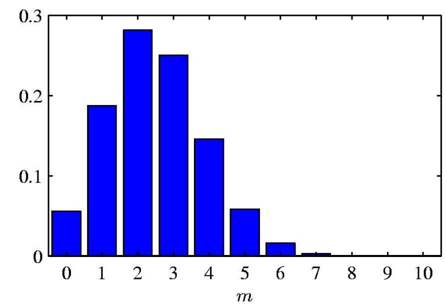
  - Random Number Generator: R = betarnd (A,B)

  returns a matrix of random numbers chosen from the beta distribution with parameters A and B.

Multinomial Variables

Generalized Bernoulli and Dirichlet

Generalization of Binomial



Histogram of Binomial for $N=10$ and $\mu=0.25$

- Binomial
  - Tossing a coin
  - Expresses probability of no of successes in N trials
    - Probability of 3 rainy days in 10 days
- Multinomial
  - Throwing a Die
  - Probability of a given frequency for each value
    - Probability of 3 specific letters in a string of N
- [Probability Calculator]
  - http://stattrek.com/Tables/Multinomial.aspx

Generalization of Bernoulli

- Bernoulli distribution x is 0 or 1

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

- Discrete variable that takes one of K values (instead of 2)
- Represent as 1 of K scheme
  - Represent x as a K-dimensional vector
  - If $x=3$ then we represent it as $\mathrm{x}=(0,0,1,0,0,0)^T$
  - Such vectors satisfy $\sum_{k=1}^{K} x_k = 1$
- If probability of $x_k=1$ is denoted μ_k then distribution of x is given by

$$p(\mathrm{x} \mid \mu) = \prod_{k=1}^{K} \mu_k^{x_k} \text{ where } \mu = (\mu_1, .., \mu_K)^T$$

Generalized Bernoulli

MLE of Generalized Bernoulli Parameters

- Data set D of N ind. observations $x_1,..x_N$
  - where the n^{th} observation is written as $[x_{n1},.., x_{nK}]$
- Likelihood function has the form

$$p(D\mid\mu)=\prod_{n=1}^{N}\prod_{k=1}^{K}\mu_k^{x_{nk}}=\prod_{k=1}^{K}\mu_k^{\left(\sum_n x_{nk}\right)}=\prod_{k=1}^{K}\mu_k^{m_k}$$

  - where $m_k=\Sigma_n x_{nk}$ is the no. of observations of $x_k=1$
- Maximum likelihood solution (obtained by setting derivative wrt μ of log-likelihood to zero) is

$$\mu_k^{ML}=\frac{m_k}{N}$$

which is fraction of N observations for which $x_k=1$

Generalized Binomial Distribution

- Multinomial distribution (with K-state variable)

$$Mult(m_1 m_2 .. m_K \mid \mu, N) = \binom{N}{m_1 m_2 .. m_k} \prod_{k=1}^{K} \mu_k^{m_k} \qquad \sum_k \mu_k = 1$$

  - Where the normalization coefficient is the no of ways of partitioning N objects into K groups of size $m_1, m_2 .. m_k$
    - Given by $\binom{N}{m_1 m_2 .. m_k} = \frac{N!}{m_1! m_2! .. m_k!}$

Dirichlet Distribution

Lejeune Dirichlet
1805-1859

- Family of prior distributions for parameters μ_k of multinomial distribution
- By inspection of multinomial, form of conjugate prior is

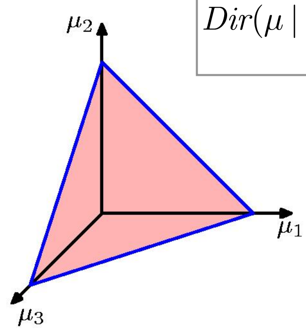
$$p(\mu \mid \alpha) \, \alpha \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \text{ where } 0 \le \mu_k \le 1 \text{ and } \sum_k \mu_k = 1$$

- Normalized form of Dirichlet distribution

$$Dir(\mu \mid \alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \text{ where } \alpha_0 = \sum_{k=1}^{K} \alpha_k$$

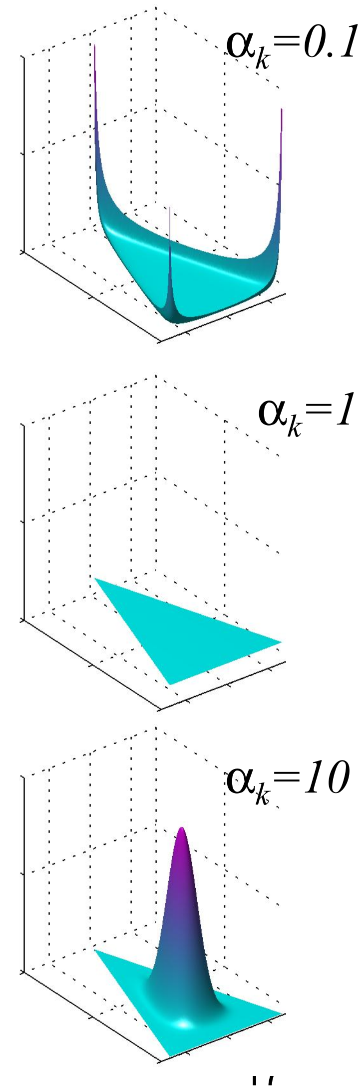
Dirichlet over 3 variables

- Due to summation constraint $\sum_k \mu_k = 1$
  - Distribution over space of $\{\mu_k\}$ is confined to the simplex of dimensionality K-1
  - For $K=3$



$$Dir(\mu \mid \alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_3)} \prod_{k=1}^{3} \mu_k^{\alpha_k - 1} \text{ where } \alpha_0 = \sum_{k=1}^{3} \alpha_k$$

Plots of Dirichlet distribution over the simplex for various settings of parameters α_k



Dirichlet Posterior Distribution

- Multiplying prior by likelihood

$$p(\mu \mid D,\alpha) \;\alpha\; p(D \mid \mu)p(\mu \mid \alpha) \;\alpha \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$$

- Which has the form of the Dirichlet distribution

$$p(\mu \mid D,\alpha) = Dir(\mu \mid \alpha + m)$$

$$= \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1)..\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k + m_k - 1}$$

Summary of Discrete Distributions

- Bernoulli (2 states) : $Bern(x|\mu)=\mu^x (1-\mu)^{1-x}$
  - Binomial: $$Bin(m\,|\,N,\mu)=\binom{N}{m}\mu^m(1-\mu)^{N-m}$$ $$\binom{N}{m}=\frac{N!}{m!(N-m)!}$$
- Generalized Bernoulli (K states):

$$p(\mathrm{x}\,|\,\mu)=\prod_{k=1}^{K}\mu_k^{x_k} \text{ where } \mu=(\mu_1,..,\mu_K)^T$$

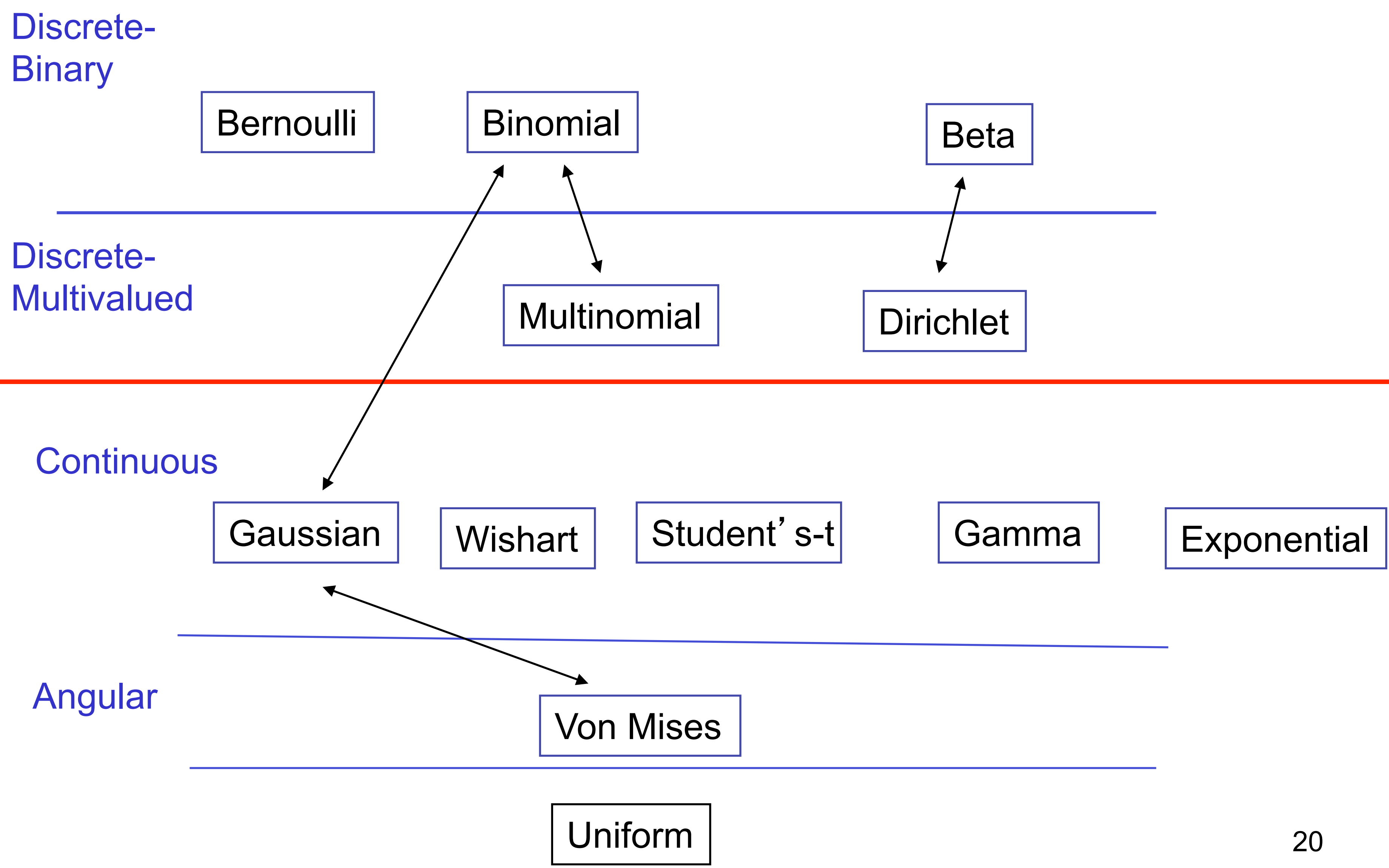
  - Multinomial $$Mult(m_1m_2..m_K\,|\,\mu,N)=\binom{N}{m_1m_2..m_k}\prod_{k=1}^{K}\mu_k^{m_k}$$
- Conjugate priors:
  - Binomial is Beta

$$Beta(\mu\,|\,a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$$

  - Multinomial is Dirichlet

$$Dir(\mu\,|\,\alpha)=\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)...\Gamma(\alpha_k)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1} \text{ where } \alpha_0=\sum_{k=1}^{K}\alpha_k$$

Distributions: Landscape



Distributions: Relationships