Pattern Recognition

Chapter 02 – Probability Distributions

Content

Content
Introduction https://statisticsbyjim.com/basics/probability-distributions/ , https://www.kdnuggets.com/2020/02/probability-distributions-data-science.html
Bernoulli Distribution
Binomial Distribution
Beta Distribution
Multinomial Distribution
Dirichlet Distribution

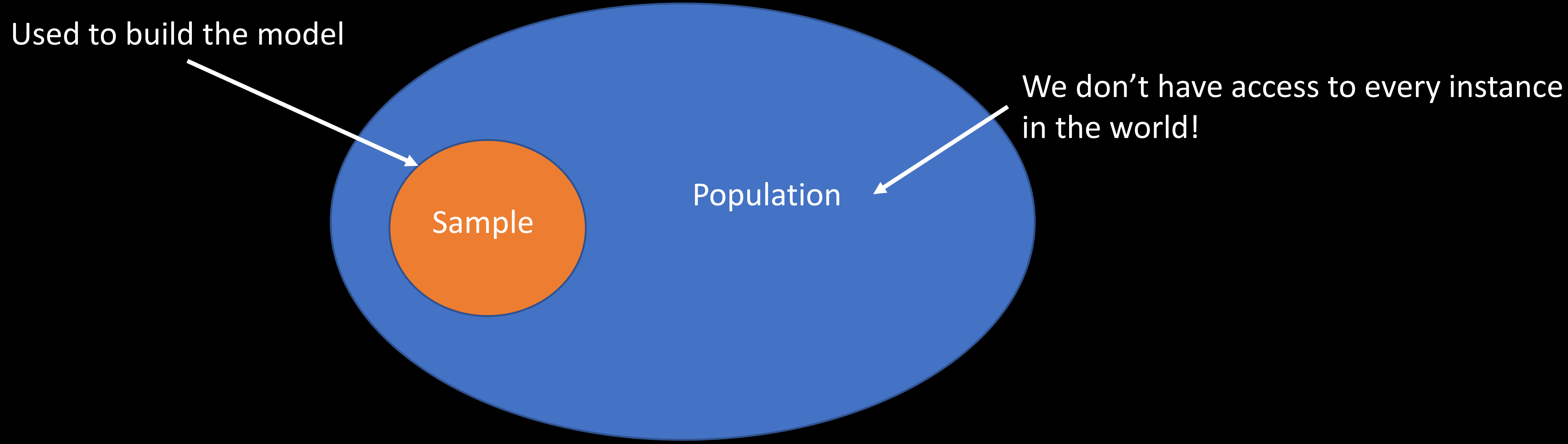
Introduction

- Probability is used for predicting the likelihood of future events.
    - Given a random variable x, what is the probability that $x = 5$?
- Statistics involves the analysis of the frequency of past events.
    - What is the mean of patients with diabetes? How many males with age 35?

- Probability and Statistics allows us to build complex models:
    - Measure the variability of the data
    - Measure the variability of the noise within the data
    - Measure the uncertainty in our model

Introduction

- Our dataset is a **sample** from a **population**.
  - The data samples are used to build models that can be deployed to predict new, unknown instances from the **population**.



Introduction

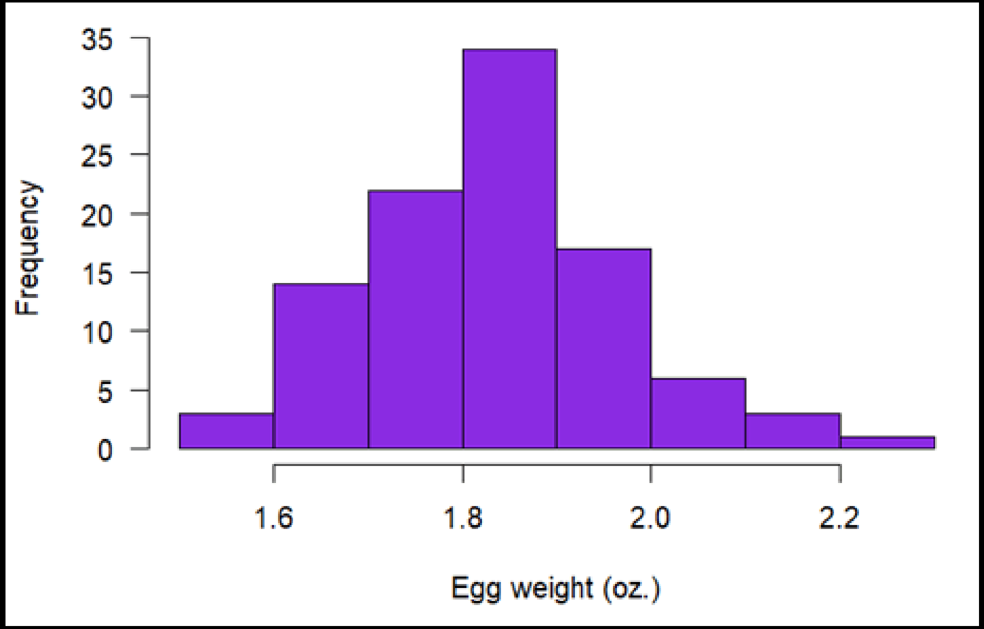
- Datasets can have two types of data:
    - **Numerical**
    - **Categorical** (e.g., {spam, not spam} , {red, green, blue}, names}
- Numerical data can be:
    - **Discrete**: take specific numeric values (number of children, number of courses)
    - **Continuous**: real numbers in any interval (distance, speed, weight, time)

Introduction

- What is a probability distribution?

Introduction

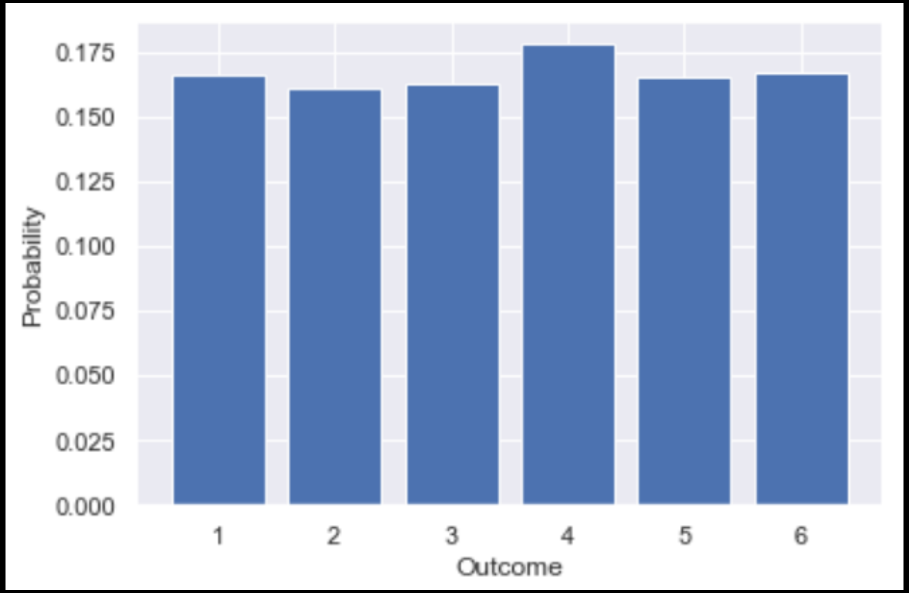
- What is a probability distribution?
  A function that describes the likelihood of obtaining all possible values that a random variable can take.
  - For example, measuring the weights of students in the class.
  - As you measure the weights, we create a distribution.
  - If we need to calculate the probability that a random student's weight is between 90KG and 100KG, we have to calculate the likelihood based on the created distribution.



https://www.scribbr.com/statistics/probability-distributions/

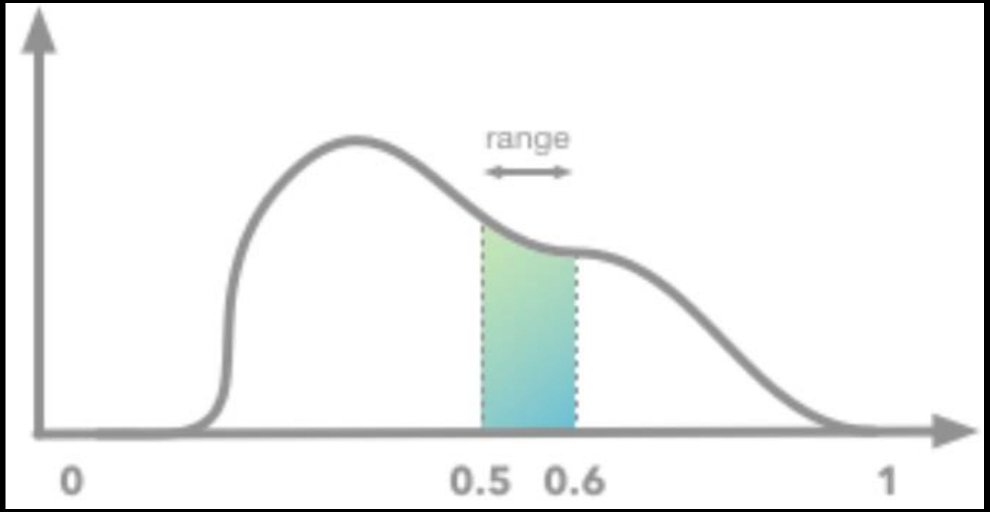
Introduction

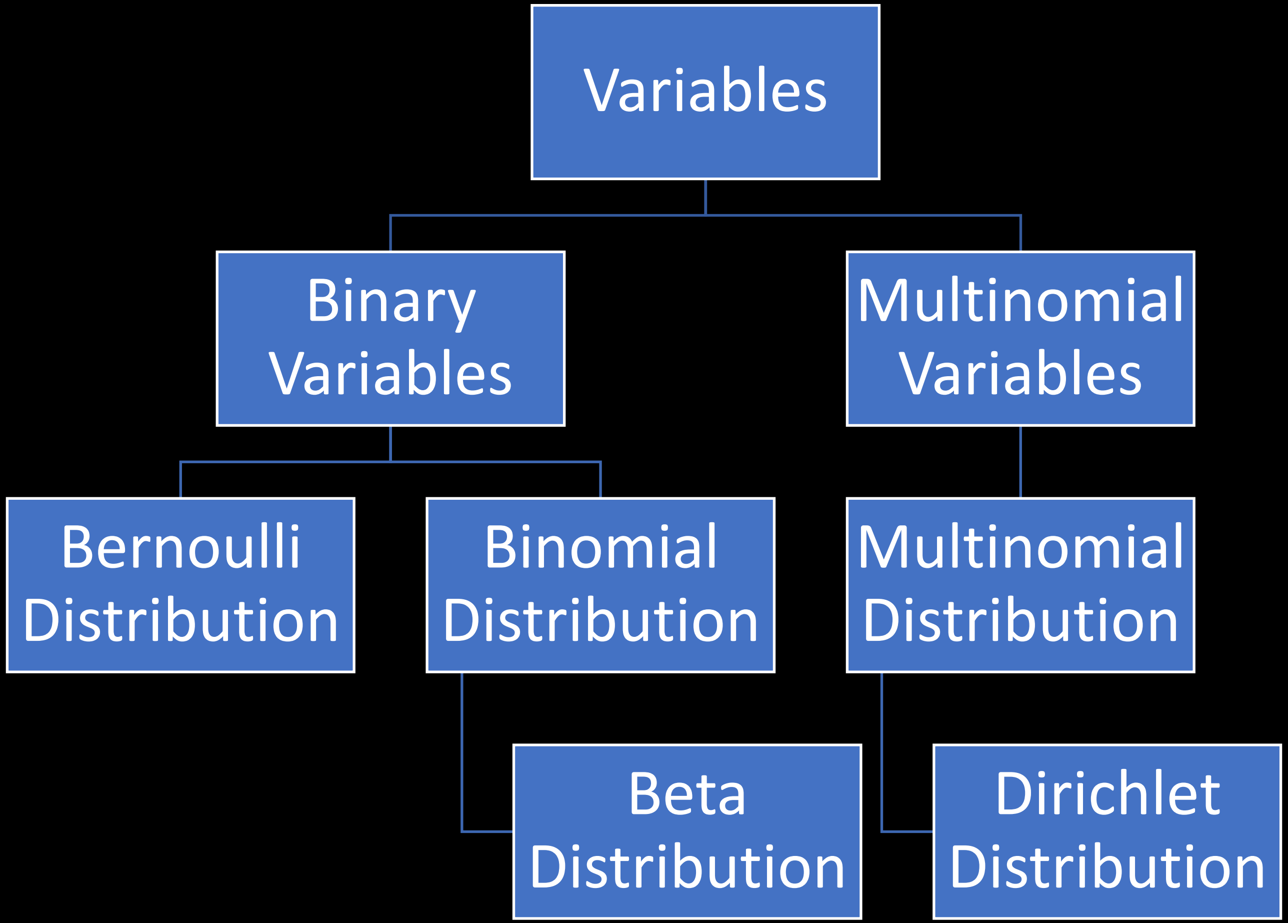
- **Probability Mass Function (PMF)**
  - Describes the probability distribution of a discrete variable (the probability that a discrete random variable can take a specific value).



kdnuggets.com/2019/05/probability-mass-density-functions.html

- **Probability Density Function (PDF)**
  - Describes the probability of a continuous variable.
  - Probabilities need to be integrated over the given range.




Variables
Binary Variables
Multinomial Variables
Bernoulli Distribution
Binomial Distribution
Multinomial Distribution
Beta Distribution
Dirichlet Distribution

Content

Content
Introduction https://statisticsbyjim.com/basics/probability-distributions/ , https://www.kdnuggets.com/2020/02/probability-distributions-data-science.html
Bernoulli Distribution
Binomial Distribution
Beta Distribution
Multinomial Distribution
Dirichlet Distribution

Bernoulli Distribution

- Bernoulli distribution is a discrete probability distribution where variables can have value 0 or 1.

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

  - x is the random variable
  - μ is the probability that $x = 1$

$$p(x = 1|\mu) = \mu$$
$$p(x = 0|\mu) = 1 - \mu$$

- The mean (expected value) of the distribution: $E(x) = \mu$
- The variance of the distribution: $var[x] = \mu(1-\mu)$

Bernoulli Distribution

- Given a dataset $D = \{x_1, x_2, \ldots, x_N\}$, the likelihood function is

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

- We notice that the function depends on the parameter μ.
  - What is the best value of μ that maximizes the likelihood?

Bernoulli Distribution

- Given a dataset $D = \{x_1, x_2, \ldots, x_N\}$, the likelihood function is

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

- We notice that the function depends on the parameter μ.
  - What is the best value of μ that maximizes the likelihood?

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

  - μ_{ML} is called a **maximum likelihood estimator**.

Bernoulli Distribution

The idea of Maximum Likelihood Estimation
is to select the parameters (e.g., μ) that make the
observed data is most likely to happen.

Bernoulli Distribution

Go to code

Content

Content
Introduction https://statisticsbyjim.com/basics/probability-distributions/ , https://www.kdnuggets.com/2020/02/probability-distributions-data-science.html
Bernoulli Distribution
Binomial Distribution
Beta Distribution
Multinomial Distribution
Dirichlet Distribution

Binomial Distribution

- The **Bernoulli distribution** is a special case of the **Binomial distribution**.
  - The Bernoulli distribution represents the success or failure of a single Bernoulli trial.
  - The Binomial Distribution represents the number of successes and failures in n independent Bernoulli trials for some given value of n.
- Binomial distribution works for the case when we get samples of the data each time.
  - For example, if we have 1,000,000 object in a dataset, we cannot train them all at once, we can train 100 samples each time.

Binomial Distribution

- Binomial distribution

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

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

  - m is the number of objects of x=1.
  - N is the total number of the objects in the dataset.

Remember!

$\binom{N}{m} = \frac{N!}{(N-m)!m!}$ is a combination; choose m things from a set of N things where the order is not important.

Content

Content
Introduction https://statisticsbyjim.com/basics/probability-distributions/ , https://www.kdnuggets.com/2020/02/probability-distributions-data-science.html
Bernoulli Distribution
Binomial Distribution
Beta Distribution
Multinomial Distribution
Dirichlet Distribution

Beta Distribution

- As seen in the Bernoulli distribution, and binomial distribution, the parameter μ is estimated on a fraction set of the whole dataset.
  - So, if this fraction contains only data with x=1, this causes overfitting.
- To solve this problem, we introduce a **prior distribution** $p(\mu)$ over the parameter μ.
  - A **prior distribution** represents the information about an uncertain parameter μ.
  - The prior distribution is combined with the probability distribution of new data to yield the **posterior distribution**.
  - This posterior distribution is used for future predictions involving μ.

Beta Distribution

- We choose a prior distribution for the parameter μ called a Beta distribution.
- Beta distribution

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

  - $\Gamma(x)$ is a gamma function.
  - a and b are hyperparameters that control the distribution of μ.
  - Consider a as the number of objects with x=1.
  - Consider b as the number of objects with x=0.

Beta Distribution

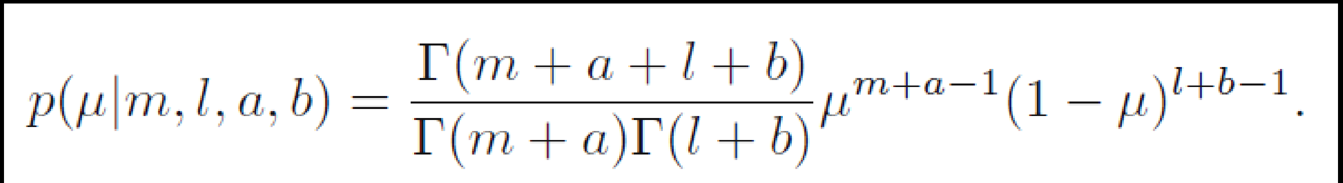
- Now, we have
  - The prior distribution for μ (beta distribution)

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

  - The likelihood function (binomial distribution)

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

- We obtain the posterior distribution by multiplying the prior distribution by the likelihood function
  - Where $l = N - m$,
    the objects with x=0.



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

Beta Distribution

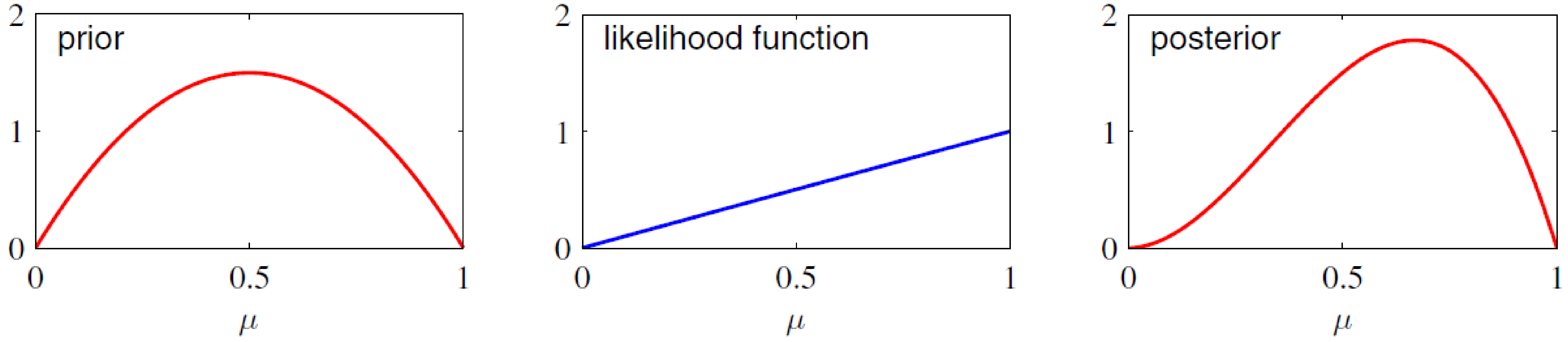
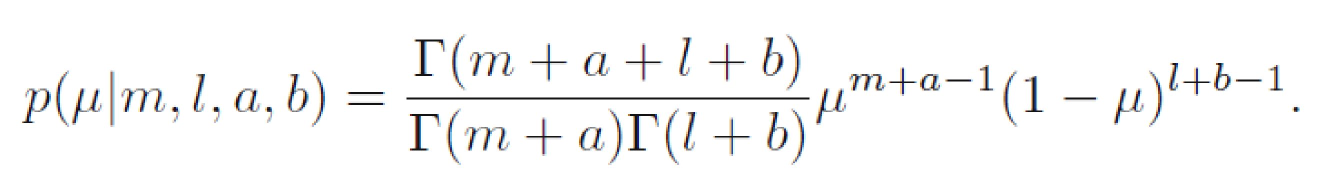


Figure 2.3 Illustration of one step of sequential Bayesian inference. The prior is given by a beta distribution with parameters $a = 2$, $b = 2$, and the likelihood function, given by (2.9) with $N = m = 1$, corresponds to a single observation of $x = 1$, so that the posterior is given by a beta distribution with parameters $a = 3$, $b = 2$.

Beta Distribution

Given the dataset D.

And and the posterior for $p(\mu|D)$,



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

And the mean of the beta distribution: $\frac{a}{a+b}$

- We get the model

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

Beta Distribution

Go to code

Content

Content
Introduction https://statisticsbyjim.com/basics/probability-distributions/ , https://www.kdnuggets.com/2020/02/probability-distributions-data-science.html
Bernoulli Distribution
Binomial Distribution
Beta Distribution
➡ Multinomial Distribution
Dirichlet Distribution

Multinomial Distribution

- Binary variables describe quantities that can take one of two possible values.
  - E.g., 0 or 1, True or False
- Multinomial variables describes variables that can take on one of K possible values.
  - E.g., {0, 1, 2, 3, 4, 5, 6}, {apple, orange, banana}
- Multinomial variables are represented using **1-of-k vector**.
  - One of the elements x_k equals 1, and all remaining elements equal 0.
  - Example: a variable can take K = 6 states and a particular observation of the variable happens to correspond to the state where $x_3 = 1$, then x will be represented by

$$\boldsymbol{x} = (0,0,1,0,0,0)^T$$

Multinomial Distribution

- Given the vector $\boldsymbol{x}$ and the vector $\boldsymbol{\mu}$, the distribution of $\boldsymbol{x}$ is given by

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

  - ○ $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots \mu_k)^T$ and the parameter μ_k is the probability that x=1.
- Given a dataset D of N observations $\boldsymbol{x_1}, \dots, \boldsymbol{x_N}$, the likelihood function is

$$p(D|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{m_k}$$

- The maximum likelihood for μ_k

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

Multinomial Distribution

- The **Multinomial Distribution** is the joint distribution of the quantities $m_1, \dots, m_k$ conditioned on the parameters $\boldsymbol{\mu}$ and the total number of observations N.

$$mult(m_1, \dots, m_k | \mu, N) = \binom{N}{m_1 m_2 \dots m_k} \prod_{k=1}^{K} \mu_k^{m_k}$$

And

$$\binom{N}{m_1 m_2 \dots m_k} = \frac{N!}{m_1! \ \dots m_k!}$$

$$\sum_{k=1}^{K} m_k = N$$

Multinomial Distribution

<u>Remember</u>!

- A joint probability distribution shows a probability distribution for two (or more) random variables. (https://www.statisticshowto.com/joint-probability-distribution/)
- $\binom{N}{m_1 m_2 \ldots m_k} = \frac{N!}{m_1! \ldots m_k!}$ means a permutation; choose m things from a set of N things where the order matters.

Content

	Content
	Introduction https://statisticsbyjim.com/basics/probability-distributions/ , https://www.kdnuggets.com/2020/02/probability-distributions-data-science.html
	Bernoulli Distribution
	Binomial Distribution
	Beta Distribution
	Multinomial Distribution
➡	Dirichlet Distribution

Dirichlet Distribution

- The Dirichlet distribution is used as a prior for calculating the parameters $\{\mu_k\}$ for the multinomial distribution.

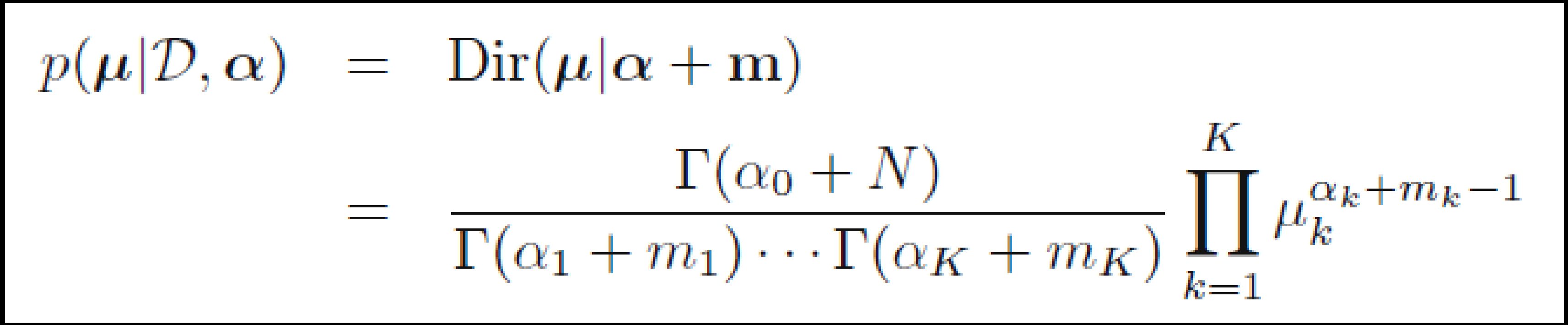
$$Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1), \dots \Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

  - o $\Gamma(x)$ is the gamma function.
  - o $\boldsymbol{\alpha}$ are the parameters that control the distributions.

Dirichlet Distribution

- The posterior distribution is calculated by multiplying the prior by the likelihood function

$$\left[Dir(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1),\ldots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha_k-1}\right] * \left[p(D|\boldsymbol{\mu}) = \prod_{k=1}^{K}\mu_k^{m_k}\right] =$$



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

Dirichlet Distribution

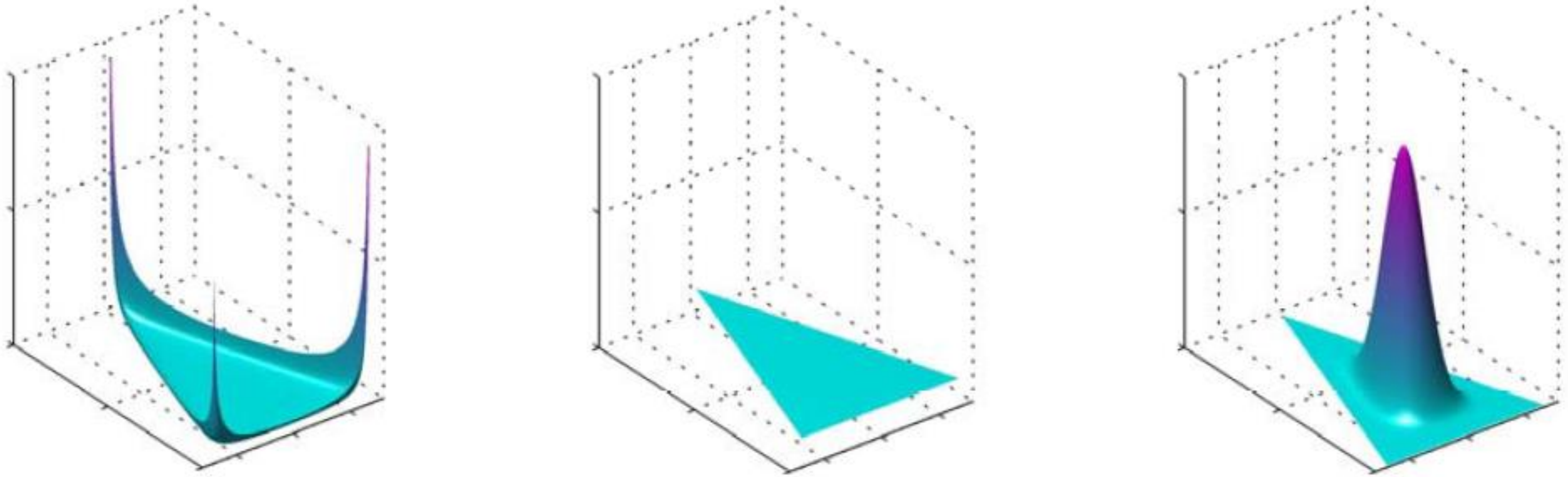


Figure 2.5 Plots of the Dirichlet distribution over three variables, where the two horizontal axes are coordinates in the plane of the simplex and the vertical axis corresponds to the value of the density. Here $\{\alpha_k\} = 0.1$ on the left plot, $\{\alpha_k\} = 1$ in the centre plot, and $\{\alpha_k\} = 10$ in the right plot.

Dirichlet Distribution

Go to code

Tasks

- What is Exploratory Data Analysis (EDA).