EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

FACULTY OF
SCIENCE
Data Science & Analytics

Data Mining and Probabilistic Reasoning

Basics of Probability Theory

Main sources:

All of statistics

Pattern Recognition and Machine Learning

Outline

- Laws of probability
- Random variables
- Probability distributions
- Expectation, variance, covariance
- Maximum likelihood estimation
- Expectation maximization
- Different views on probabilities
- Bayesian inference

"As far as the propositions of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality." **Albert Einstein, 1921**

Set-theoretic view of probability theory

Probability space

- (Ω, E, P) with
- Ω: **sample space** of elementary **events**
- E: **event space**, i.e. subsets of Ω, closed under $\cap$, $\cup$, and $\neg$, usually $E = 2^{\Omega}$
- P: $E \rightarrow [0,1]$, **probability measure**

Properties of P:

1. $P(\emptyset) = 0$ (impossible event)
2. $P(\Omega) = 1$
3. $P(A) + P(\neg A) = 1$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5. $P(\bigcup_i A_i) = \sum_i P(A_i)$ for pairwise disjoint A_i

Sample space and events: Examples

- Rolling a die
  - Sample space: {1, 2, 3, 4, 5, 6}
  - Probability of even number: Looking for events A = {2}, B = {4}, C = {6} → $P(A \cup B \cup C)$ = 1/6+1/6+1/6 = 0.5
- Tossing two coins
  - Sample space: {HH, HT, TH, TT}
  - Probability of both coins showing tails or heads: A = {HH}, B = {TT} → $P(A \cup B)$ = 1/4+1/4 = 0.5

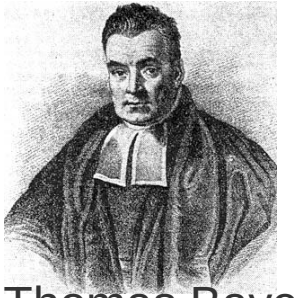
In general, when all outcomes in Ω are equally likely, for an $e \in E$ holds:

$$P(e) = \frac{\text{\# outcomes in } e}{\text{\# outcomes in sample space}}$$

Joint, marginal, and conditional probabilities

Joint and conditional probability

$P(A, B) = P(A \cap B) = P(B|A) \cdot P(A)$ (product rule)


Thomas Bayes

Bayes' theorem

$P(B|A) = \ P(A|B) \cdot P(B)/P(A)$

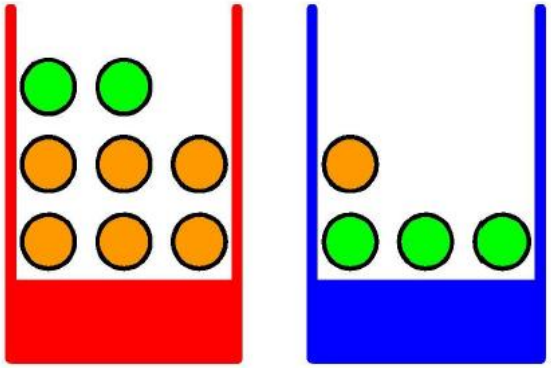
Total/marginal probability

$P(B) = \sum_j P(B \cap A_j)$ for any partitioning of Ω in $A_1, \ldots, A_n$ (sum rule)

Joint, marginal, and conditional probabilities: Example

Suppose: P(B = r) = 2/5

Apples and Oranges



Fruit is orange, what is probability that box was blue?

- $P(B=b|F=o)=\frac{P(F=o|B=b)P(B=b)}{P(F=o)}$
- P(F = o) = P(F = o | B = r) P(B = r) + P(F = o | B = b) P(B = b) = 9/20

Source: *Pattern Recognition and Machine Learning.* C. Bishop

Independent events

Independence

$P(A_1, \dots, A_n) = P(A_1 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2) \cdot \dots \cdot P(A_n)$, for independent events $A_1, \dots, A_n$

Conditional Independence

- A is independent of B given C if and only if $P(A|B, C) = P(A|C)$
- If $A_1, \dots, A_n$ are independent of each other given B then

$$P(A_1 \cap \dots \cap A_n | B) = \prod_i P(A\_i | B)$$

- If A and B are independent, are they also independent given C?

Example of Simpson's Paradox : Which drug works better?

	Women		Men	
	Drug x	Drug y	Drug x	Drug y
Success	100	5	10	500
Failure	900	95	1	500

S = Drug succeeds
X = Drug x is used
Y = Drug y is used
W = Patient is female
M = Patient is male

$P(S|X,W) = 0.10 \quad P(S|X,M) = 0.91$
$P(S|Y,W) = 0.05 \quad P(S|Y,M) = 0.5$

Drug x works better

$P(S|X) \approx 0.11$
$P(S|Y) \approx 0.46$

Drug y works better

Observation: In above table being a male is a strong cause for both drug usage and recovery
Solution: In such cases, one should evaluate the probabilities on the subgroups separately and report weighted averages

Discrete and continuous random variables

Random variable on probability space (Ω, E, P)

X: $\Omega \rightarrow M \subseteq \mathbb{R}$ (numerical representations of outcomes)

with $\{e|X(e) \leq x\} \in E$ for all $x \in M$

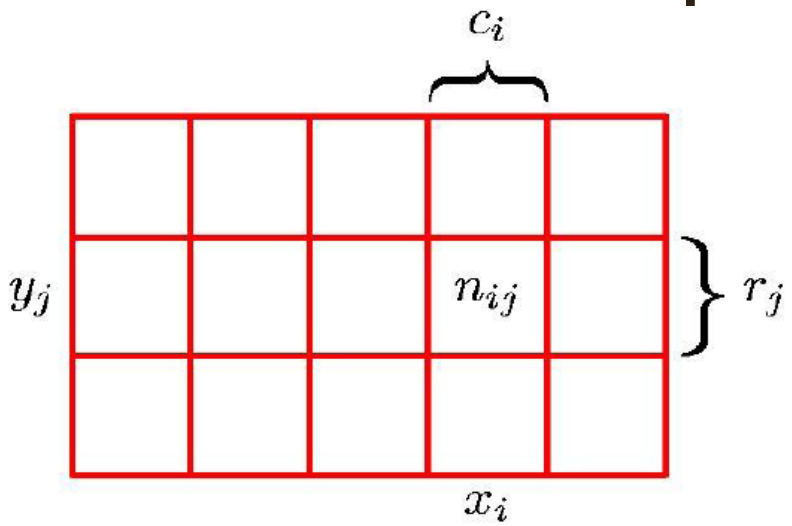
If M is countable, X is called discrete, otherwise continuous

Examples

- Rolling a die: $X(i) = i$
- The exact pair of faces when rolling two dice: $X(a, b) = 6(a - 1) + b$
- The sum of faces for two dice: $X(\{a, b\}) = a + b$

Random variables $X_1, \ldots, X_2$ are called **independent and identically distributed (i.i.d.)** if each random variable has the same probability distribution as the others and all are mutually independent

Random variables and probabilities



Marginal probability

$$P(X = x_i) = \frac{c_i}{N}$$

Sum rule

$$P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$$
$$= \frac{1}{N} \sum_j n_{ij} = \frac{c_i}{N}$$

Joint probability

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Product rule

$$P(X = x_i, Y = y_j) = P(Y = y_j | X = x_i) P(X = x_i)$$
$$= \frac{n_{ij}}{c_i} \frac{c_i}{N} = \frac{n_{ij}}{N}$$

Source: *Pattern Recognition and Machine Learning.* C. Bishop

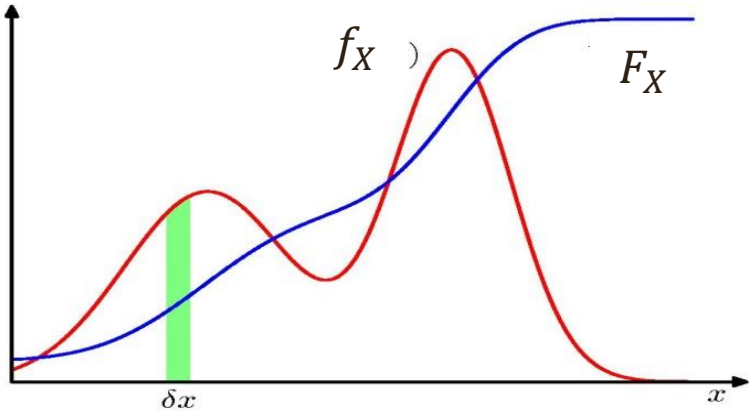
Probability functions

Cumulative distribution function (cdf)

$F_X: M \to [0,1]$ with $F_X(x) = P(X \leq x)$

Probability density function (pdf) (or prob. mass function for discrete case)

$f_X: M \to [0,1]$ with $f_X(x) = P(X = x) := P(x \leq X \leq x + \delta x), \delta x \to 0$



Quantile function

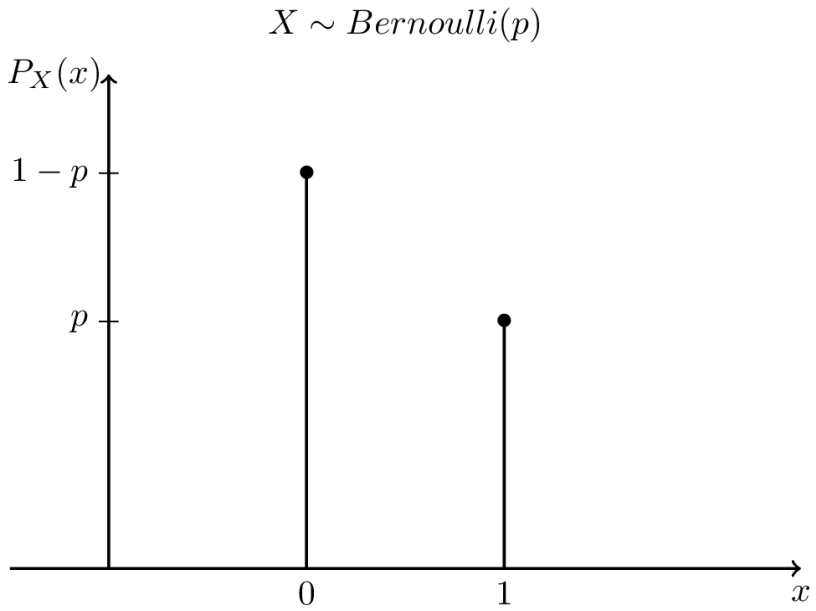
$F^{-1}(q) = \inf\{x | F_X(x) > q\}, \ q \in [0,1]$ (for $q = 0.5$, $F^{-1}(q)$ is called **median**)

Source: *Pattern Recognition and Machine Learning.* C. Bishop

Useful discrete distributions (1)

Uniform distribution over {1, 2, …, m}: $P(X = k) = f_X(k) = \frac{1}{m}$

Bernoulli distribution with parameter p: $P(X = x) = f_X(x) = p^x(1-p)^{1-x}$

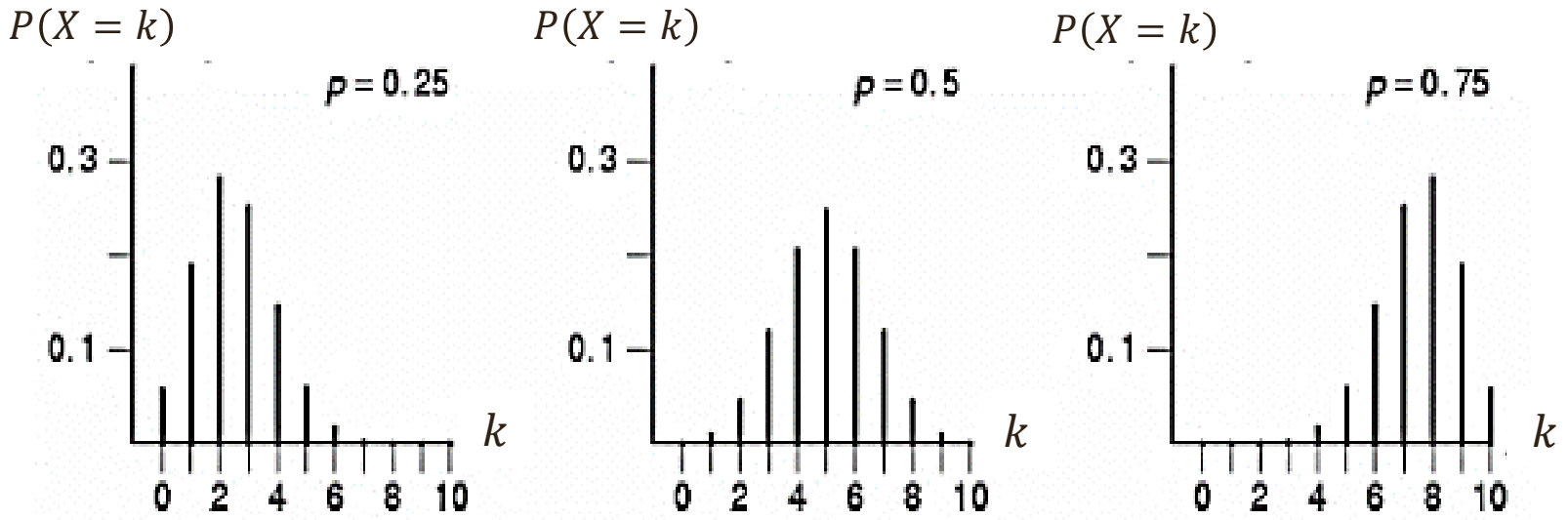


Useful discrete distributions (2)

Binomial distribution with parameter p, m:

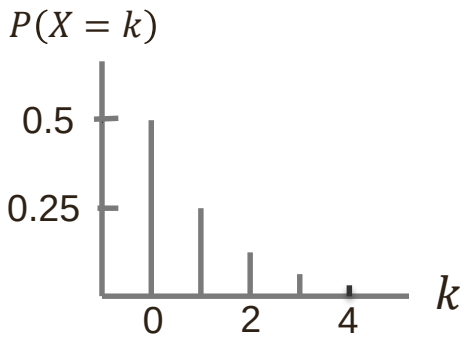
$$P(X = k) = f_X(k) = \binom{m}{k} p^k (1-p)^{m-k}$$

$$X \sim Binomial(p, 10)$$



Useful discrete distributions (3)

Geometric distribution with parameter p: $P(X = k) = f_X(k) = (1-p)^k p$



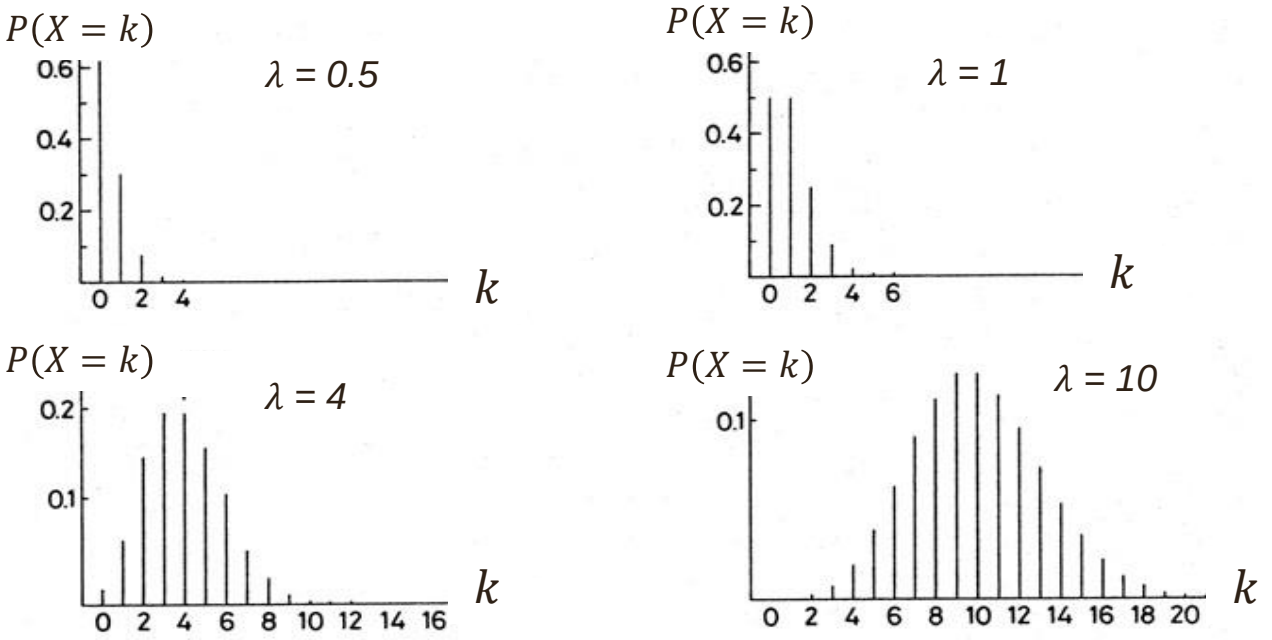
Poisson distribution with parameter λ: $P(X = k) = f_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$

Poisson process

- Counting process
- $P(X{=}k)$: probability that there will be k increments per time unit
- Parameter λ: expected number of increments per time unit

Useful discrete distributions (4)

Examples for the Poisson distribution: $P(X = k) = f_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$

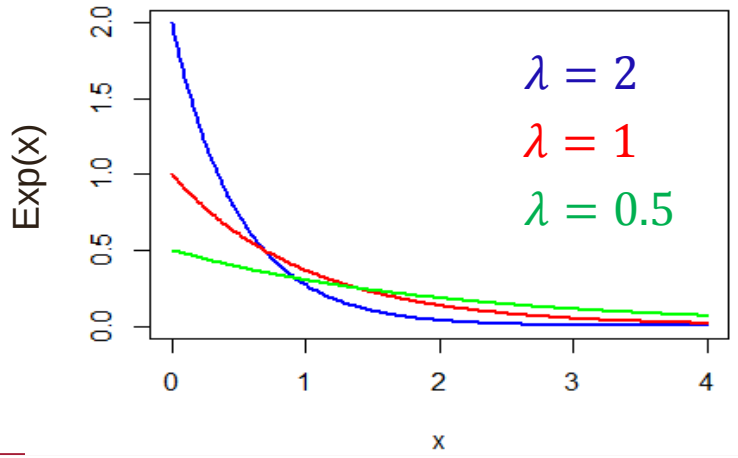


Useful continuous distributions (1)

Uniform distribution over $[a, b]$: $P(X = x) = f_X(x) = \frac{1}{b-a}$ for $a \leq x \leq b$

Exponential distribution with parameter λ: $P(X = x) = f_X(x) = \lambda e^{-\lambda x}$ for $x > 0$

- Describes a process in which events occur continuously and independently at constant average rate λ
- Can be used to model
  - Time between two phone calls
  - Modeling of radioactive decay
  - Durability of electronic devices



Useful continuous distributions (2)

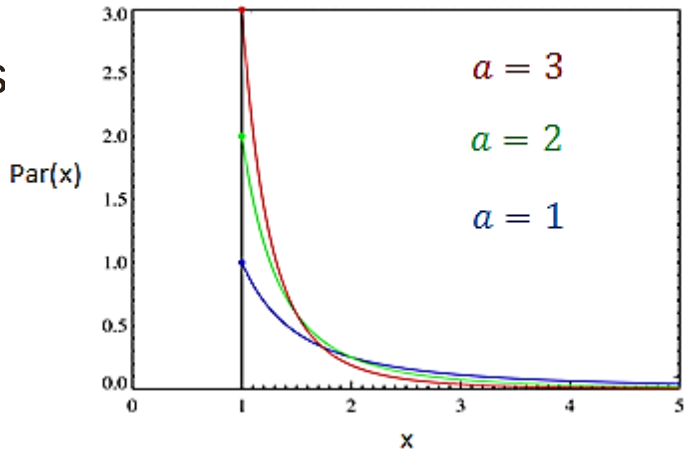
Pareto distribution with parameters a, b:

$$P(X = x) = f_X(x) = \frac{a}{b}\left(\frac{b}{x}\right)^{a+1}, \qquad x > b$$

Pareto principle: 80% of the effects come from 20% of the causes

Examples of such power-law distributions

- Distribution of populations over cities
- Distribution of wealth
- Citations distribution over research papers
- Distribution over the number of followers in social networks

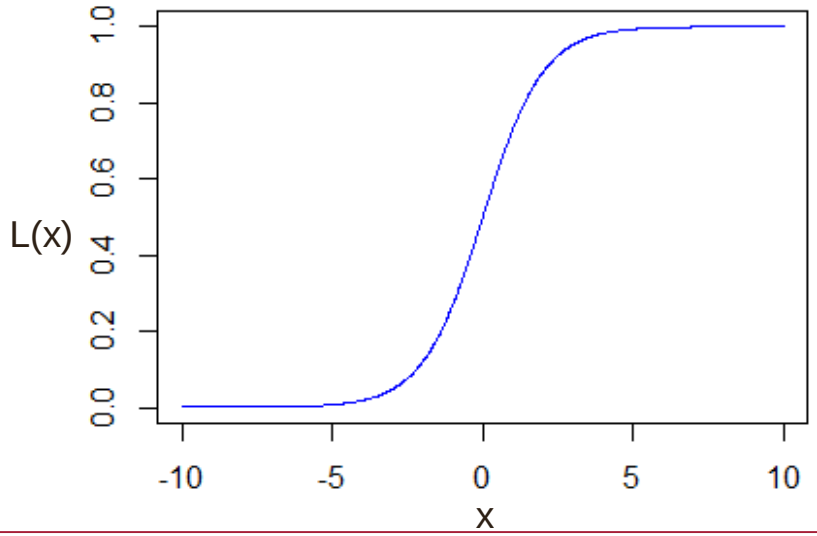


Useful continuous distributions (3)

Logistic distribution: $P(X \leq x) = F_X(x) = \frac{1}{1+e^{-x}}$

Applications, e.g.:

- Classification, e.g., with logistic regression
- Inference in neural networks
- In Psychometrics, e.g. Item-Response Theory



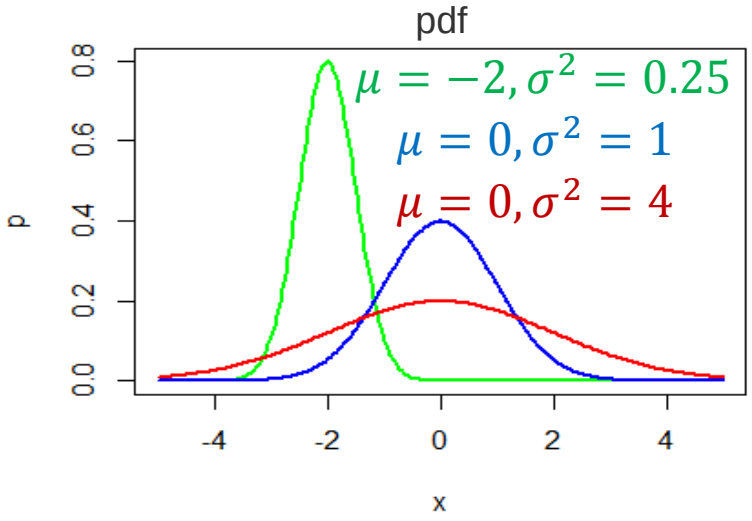
Useful continuous distributions (4)

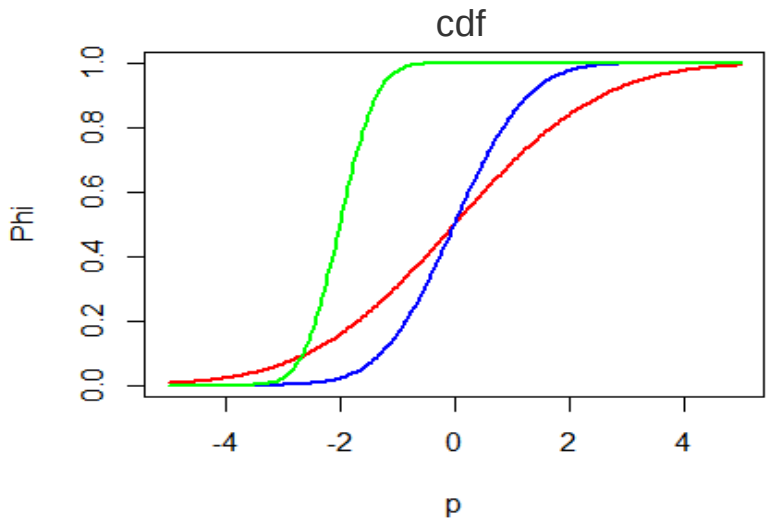
Normal distribution (Gaussian)

$$X \sim N(\mu, \sigma^2) \Leftrightarrow f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

μ: mean, σ: standard deviation

Cumulative distribution of $N(0,1)$: $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx$





Multivariate distributions

Let $X_1, \dots, X_m$ be random variables over the same probability space with domains $D(X_1), \dots, D(X_m)$

The ***joint distribution*** of $X_1, \dots, X_m$ has a pdf $f_{X_1,\dots,X_m}(x_1, \dots, x_m)$ with

$$\sum_{x_1 \in D(X_1)} \cdots \sum_{x_m \in D(X_m)} f_{X_1,\dots,X_m}(x_1, \dots, x_m) = 1$$

$$\int_{x_1 \in D(X_1)} \cdots \int_{x_m \in D(X_m)} f_{X_1,\dots,X_m}(x_1, \dots, x_m)\ dx_1 \dots\ dx_m = 1$$

The ***marginal distribution*** of X_i is $F_{X_1,\dots,X_m}(x_i) =$

$$\sum_{x_1 \in D(X_1)} \cdots \sum_{x_{i-1} \in D(X_{i-1})} \sum_{x_{i+1} \in D(X_{i+1})} \cdots \sum_{x_m \in D(X_m)} f_{X_1,\dots,X_m}(x_1, \dots, x_m)$$

$$\int_{x_1 \in D(X_1)} \cdots \int_{x_{i-1} \in D(X_{i-1})} \int_{x_{i+1} \in D(X_{i+1})} \cdots \int_{x_m \in D(X_m)} f_{X_1,\dots,X_m}(x_1, \dots, x_m)\ dx_1 \dots\ dx_m$$

Useful multivariate distributions

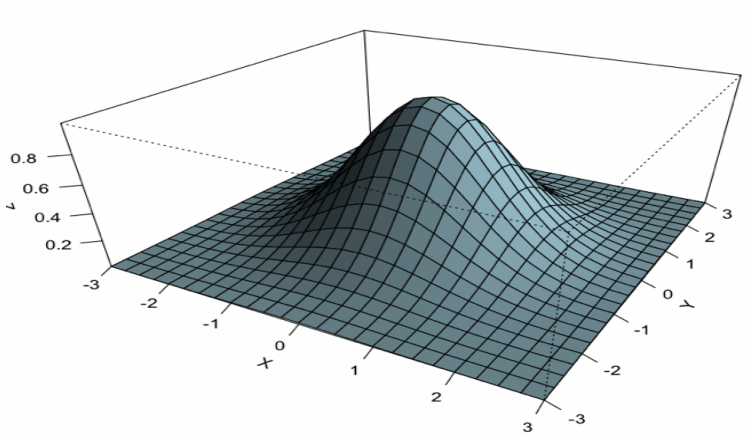
Multinomial distribution with parameters n, m (rolling n m-sided dice)

$$P(X_1 = k_1 \dots X_m = k_m) = f_{X_1,\dots,X_m}(k_1, \dots, k_m) = \frac{n!}{k_1! \dots k_m!} {p_1}^{k_1} \dots {p_m}^{k_m}$$

with $k_1 + \dots + k_m = n$ and $p_1 + \dots + p_m = 1$

Multivariate Gaussian with parameters $\vec{\mu}, \Sigma$ where $\Sigma_{ij} := Cov(X_i, X_j)$

$$f_{X_1,\dots,X_m}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^m |\Sigma|}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \Sigma^{-1} (\vec{x}-\vec{\mu})}$$



Expectation of random variables

For discrete variable X: $E(X) = \sum_x x\, f_X(x)$ is the expectation of X

For continuous variable X: $E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$

Properties

- $E(X_i + X_j) = E(X_i) + E(X_j)$
- $E(X_i\, X_j) = E(X_i)E(X_j)$ for independent, identically distributed (i.i.d.) variables X_i, X_j
- $E(aX + b) = aE(X) + b$ for constants a, b

Variance, standard deviation, and covariance

Variance

$$Var(X) = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

Properties

$Var(X_i + X_j) = Var(X_i) + Var(X_j)$ for i.i.d. variables X_i, X_j

$Var(aX + b) = a^2 Var(x)$ for constants a, b

Standard deviation

$$StDev(X) = \sqrt{Var(X)}$$

Covariance

$$Cov(X_i, X_j) = E[(X_i - E[X_i])\ (X_j - E[X_j])]$$

$Var(X) = Cov(X, X)$

In general: $Var(X + Y) = Var(X) + Var(Y) + Cov(X, Y)$

Maximum Likelihood Estimation (MLE)

Suppose that after tossing a coin n times, we have seen k times head.
Let p be the unknown probability of the coin showing head
Is it possible to estimate p?
We know that observation corresponds to Binomial distribution, hence:

$$L(p; k, n) = P(k, n|p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Maximizing $L(p; k, n)$ is equivalent to maximizing $\log L(p; k, n)$
$\log L(p; k, n)$ is called ***log-likelihood function***

$$\log L(p; k, n) = \log \binom{n}{k} + k \log p + (n-k) \log (1-p)$$

$$\frac{\partial \log L}{\partial p} = \frac{k}{p} - \frac{(n-k)}{(1-p)} = 0 \Rightarrow p = \frac{k}{n}$$

MLE example

Assume $x_1, \dots, x_n$ originate from a Gaussian with unknown μ and σ^2

$$L(\mu, \sigma; x_1, \dots, x_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$\log L(\mu, \sigma; x_1, \dots, x_n) = n \cdot \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right) + \sum_i -\frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial \log L}{\partial \mu} = \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i - \mu) = 0; \quad \frac{\partial \log L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0$$

$$\Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \; \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

MLE generalization

Let $x_1, \dots, x_n$ be a random sample from a distribution $f(\boldsymbol{\theta}, x)$
$x_1, \dots, x_n$ can be viewed as the values of i.i.d. random variables $X_1, \dots, X_n$:

$$L(\boldsymbol{\theta}; x_1, \dots, x_n) = P[x_1, \dots, x_n \text{ originate from } f(\boldsymbol{\theta}, x)]$$

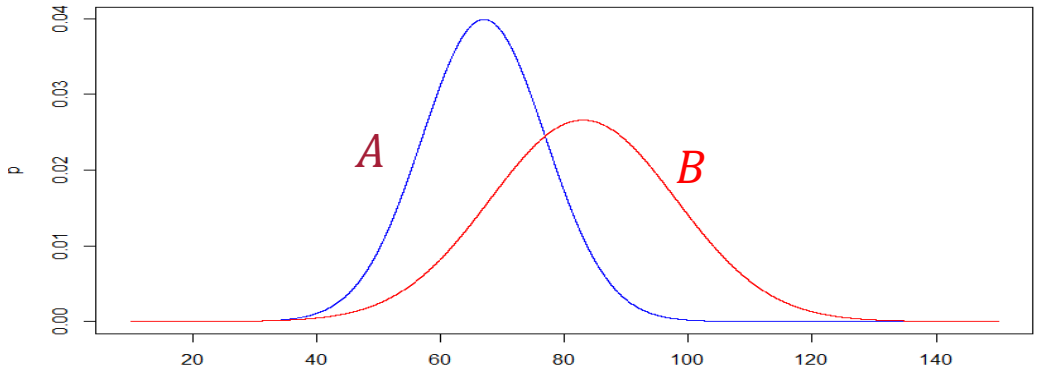
Maximizing $L(\boldsymbol{\theta}; x_1, \dots, x_n)$ is equivalent to maximizing $\log L(\boldsymbol{\theta}; x_1, \dots, x_n)$, i.e., the **log-likelihood function**: $\log P(x_1, \dots, x_n|\boldsymbol{\theta})$.

If $\frac{\partial \log L}{\partial\, p}$ is analytically intractable, use iterative numerical methods, e.g. **Expectation Maximization (EM)**

Mixture models

Example: Gaussian Mixture Model (GMM)

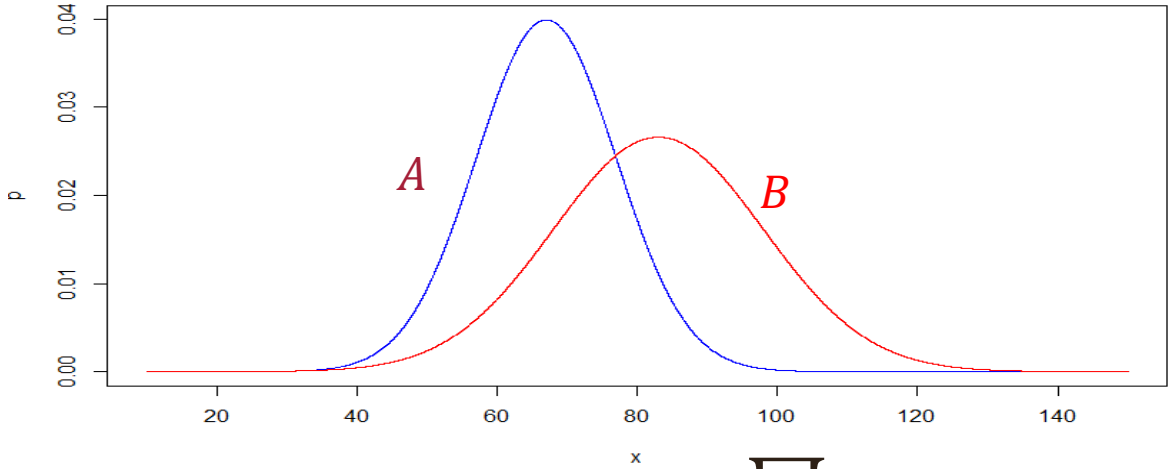
Suppose $x_1, \dots, x_n$ are random samples from a mixture of Gaussians $M(A, B)$ with $A(\mu_A, {\sigma_A}^2)$ and $B(\mu_B, {\sigma_B}^2)$, with unknown means and variances (e.g., weights of women and men)



$$L(\mu_A, \sigma_A, \mu_B, \sigma_B, p_A, p_B; x_1, \dots, x_n) = \prod_i (p_A \, P(x_i|A) + p_B \, P(x_i|B))$$

with $p_A + p_B = 1$ and $P(x_i|A) = A(\mu_A, {\sigma_A}^2, x_i) = \frac{1}{\sqrt{2\pi {\sigma_A}^2}} e^{-\frac{(x_i - \mu_A)^2}{2{\sigma_A}^2}}$

Expectation maximization (EM)



$$L(\mu_A, \sigma_A, \mu_B, \sigma_B, p_A, p_B; x_1, \dots, x_n) = \prod_i (p_A\, P(x_i|A) + p_B\, P(x_i|B))$$

1. **Expectation step:** Estimate the expected membership value of each point x_i given the current estimations of $\mu_A, \sigma_A, \mu_B, \sigma_B, p_A, p_B$
2. **Maximization step:** Use the expected membership values to re-estimate the parameters, and continue with Step 1 until convergence of $\log L(\mu_A, \sigma_A, \mu_B, \sigma_B, p_A, p_B; x_1, \dots, x_n)$

EM algorithm for mixture models

$$L(\mu_A, \sigma_A, \mu_B, \sigma_B, p_A, p_B; x_1, \dots, x_n) = \prod_i (p_A\, P(x_i|A) + p_B\, P(x_i|B))$$

Initialize the parameters $\mu_A, \sigma_A, \mu_B, \sigma_B, p_A, p_B$ to some random values (constraint: $p_A + p_B = 1$)

1. **E-step:** For each x_i compute expected membership values $P(A|x_i), P(B|x_i)$
2. **M-step:** Re-estimate the parameters
3. **Iterate** steps 2 and 3 until convergence (i.e., until changes of log likelihood are negligible)

Exact EM calculations for the GMM example

$$L(\mu_A, \sigma_A, \mu_B, \sigma_B, p_A, p_B; x_1, \ldots, x_n) = \prod_i (p_A\, P(x_i|A) + p_B\, P(x_i|B))$$

Start with random parameters.

Maximize log-likelihood (i.e., target function) by iterating following steps

- Compute **membership weights**: $w_{Ai} = P(A|x_i) = \frac{P(x_i|A)\, p_A}{P(x_i|A)p_A + P(x_i|B)p_B}$
- Compute parameters

$$p_A = \frac{1}{n}\sum_i w_{Ai} \;;\; p_B = \frac{1}{n}\sum_i w_{Bi}$$

$$\mu_A = \frac{w_{A1}x_1 + \ldots + w_{An}x_n}{w_{A1} + \ldots + w_{An}} \;;\; \mu_B = \frac{w_{B1}x_1 + \ldots + w_{Bn}x_n}{w_{B1} + \ldots + w_{Bn}}$$

$${\sigma_A}^2 = \frac{w_{A1}(x_1 - \mu_A)^2 + \ldots + w_{An}(x_n - \mu_A)^2}{w_{A1} + \ldots + w_{An}} \;;\; {\sigma_B}^2 = \frac{w_{B1}(x_1 - \mu_B)^2 + \ldots + w_{Bn}(x_n - \mu_B)^2}{w_{B1} + \ldots + w_{Bn}}$$

EM generalization

For **observed data points** $x_1, \dots, x_n$ and **hidden values** $z_1, \dots, z_m$ and model parameters $\boldsymbol{\theta}$, estimate the **maximum likelihood** of

$$L(\boldsymbol{\theta}; \mathbf{x}) = \sum_{\mathbf{z}} P(\mathbf{x}, \mathbf{z} | \boldsymbol{\theta})$$

Expectation step

- Estimate the expected value of $\mathbf{z}$ under the current parameters $\boldsymbol{\theta}^{(t)}$ and the observed data points $\mathbf{x}$
- Estimate the expected value of $\log P(\mathbf{x}, \mathbf{z} | \boldsymbol{\theta}^{(t)})$ with the current value of $\mathbf{z}$

Maximization step: Use the just computed estimation of $\mathbf{z}$ to find $\boldsymbol{\theta}^{(t+1)}$ that maximizes $\log P(\mathbf{x}, \mathbf{z} | \boldsymbol{\theta}^{(t+1)})$

Note: EM monotonically approaches local maximum

Different views: The frequentists' view

Probability of an event should be assessed objectively, i.e., measure the probability of the event as the relative occurrence frequency of that event based on a large number of trials

Examples

- Fraction of heads when tossing a coin n times
- Relative frequency with which the face 6 shows up when rolling a die n times
- Relative frequency with which a drug shows certain adverse reaction when tested on n subjects

Shortcomings

- Can be only applied to frequently repeatable events
- The higher the frequency of an event, the more "meaningful" the probability estimate

Different views: The Bayesian view

Prior beliefs / probabilities are used to quantify the uncertainty about the occurrence of events, i.e., prior beliefs are used to quantify the uncertainty of parameters of a statistical model

Prior beliefs are updated based on new observations and allow the adaptation of the parameters to the new data

With increasing number of observations, prior beliefs become less and less relevant (i.e., uncertainty is reduced)

Drawback: Reasoning and inference has to include the prior beliefs

Bayesian inference

By applying Bayes' theorem: $P(\boldsymbol{\theta}|x_1, \ldots, x_n) = \frac{P(x_1, \ldots, x_n|\boldsymbol{\theta})\, P(\boldsymbol{\theta})}{P(x_1, \ldots, x_n)}$

$$P(\boldsymbol{\theta}|x_1, \ldots, x_n) \propto P(x_1, \ldots, x_n|\boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

Evidence

Posterior distribution of parameters

Likelihood of parameters in light of data

Prior probability distribution for parameters

Iterative substitution as more and more data points are observed

Typically with **exponential family distributions** with pdfs of the form:

$$P(\boldsymbol{x}; \boldsymbol{\theta}) = h(\boldsymbol{x}) g(\boldsymbol{\theta}) \exp\{\boldsymbol{\theta}^T \boldsymbol{u}(\boldsymbol{x})\}$$

Important property: closure under multiplication (→algebraic convenience)

Bayesian inference: Example

$P(\boldsymbol{\theta}|x_1, \dots, x_n) \propto P(x_1, \dots, x_n|\boldsymbol{\theta})\, P(\boldsymbol{\theta})$

Example

- Suppose $P(k_1, k_2|\theta) = \binom{k_1+k_2}{k_1} \theta^{k_1}(1-\theta)^{k_2}$ (binomially distributed data)
- Assume $P(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$ (θ is Beta distributed with hyper-parameters a, b: counts reflecting belief formation)

$$P(\theta|k_1, k_2) = \frac{\binom{k_1+k_2}{k_1} \theta^{k_1+a-1}(1-\theta)^{k_2+b-1} \frac{1}{B(a,b)}}{\int_{\theta=0}^{1} \left(\binom{k_1+x_2}{k_1} \theta^{k_1+a-1}(1-\theta)^{k_2+b-1} \frac{1}{B(a,b)} \right) d\theta}$$

$$= \frac{\theta^{k_1+a-1}(1-\theta)^{k_2+b-1}}{B(k_1+a, k_2+b)}$$

Posterior of parameters has same form as the prior

Bayesian inference: Conjugate priors

$P(\boldsymbol{\theta})$ is called a **conjugate prior** of $P(x_1, \dots, x_n|\boldsymbol{\theta})$ if the posterior, $P(\boldsymbol{\theta}|x_1, \dots, x_n)$, is in the same pdf family as the prior.

Examples

Likelihood function	Conjugate prior
Bernoulli	Beta
Binomial	Beta
Poisson	Gamma
Multinomial	Dirichlet
Gaussian	Gaussian

Cox's theorem

Any belief system satisfying the following conditions can be described by the laws of probability

- *The belief in the occurrence of an event is dependent on information about the event (**dependency**)*
- *The belief in the occurrence of an event can be represented by a real number (**numerical comparability**)*
- *The belief in the occurrence of an event changes sensibly with observations (**common sense**)*
- *If the belief in the occurrence of an event can be derived in many ways, all the results must be equal (**consistency**)*