Human-Oriented Robotics

Probability Refresher

Kai Arras

Social Robotics Lab, University of Freiburg

Winter term 2014/2015

Probability Refresher

- **Introduction to Probability**
  - Random variables
  - Joint distribution
  - Marginalization
  - Conditional probability
  - Chain rule
  - Bayes' rule
  - Independence
  - Conditional independence
  - Expectation and Variance
- **Common Probability Distributions**
  - Bernoulli distribution
  - Binomial distribution
  - Categorial distribution
  - Multinomial distribution
  - Poisson distribution
  - Gaussian distribution
  - Chi-squared distribution

We assume that you are familiar with the **fundamentals** of probability theory and probability distributions

This is a quick refresher, we aim at **ease of understanding** rather than formal depth

For a more comprehensive treatment, refer, e.g. to A. Papoulis or the references given on the last slide

Why probability theory?

- Consider a **human, animal, or robot** in the **real world** those task involves the solution of a set of problems (e.g. an animal looking for food, a robot serving coffee, ...)
- In order to be successful, it needs to **observe** and **estimate** the state of the world around it and **act** in an appropriate way
- **Uncertainty** is an **inescapable aspect** of the real world
- It is a consequence of several factors, for example,
  - Uncertainty from **partial**, **indirect** and **ambiguous** observations of the world
  - Uncertainty in the **values** of observations (e.g. sensor noise)
  - Uncertainty in the **origin** of observations (e.g. data association)
  - Uncertainty in **action** execution (e.g. from limitations in the control system)
- Probability theory is the **most powerful** (and accepted) **formalism** to deal with uncertainty

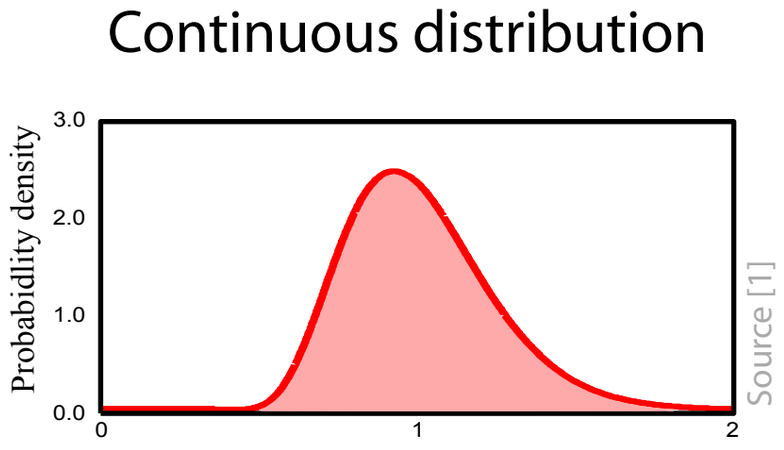
Random Variables

- A **random variable** x denotes an **uncertain quantity**
- x could be the **outcome of an experiment** such as rolling a dice (numbers from 1 to 6), flipping a coin (heads, tails), or measuring a temperature (value in degrees Celcius)
- If we observe several instances $\{x_i\}_{i=1}^{I}$ then it might take a different value each time, some values may occur more often than others. This information is captured by the **probability distribution** $p(x)$ of x
- A random variable may be **continuous** or **discrete**
  - **Continuous** random variables take values that are **real numbers**: **finite** (e.g. time taken to finish 2-hour exam), **infinite** (time until next bus arrives)
  - **Discrete** random variables take values from a **predefined set**: **ordered** (e.g. outcomes 1 to 6), **unordered** (e.g. "sunny", "raining", "cloudy"), **finite** or **infinite**.

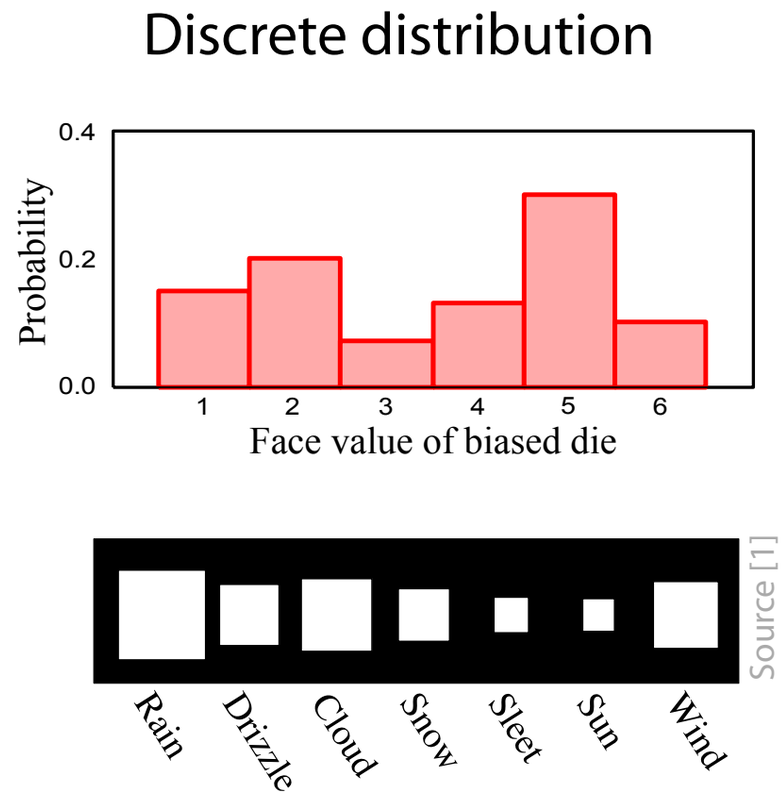
Introduction to Probability

Random Variables

- The probability distribution $p(x)$ of a continuous random variable is called **probability density function** (pdf). This function may take **any** positive value, its integral always sums to **one**



- The probability distribution $p(x)$ of a discrete random variables is called **probability mass function** and can be visualized as a **histogram** (less often: Hinton diagram). Each outcome has a positive probability associated to it whose **sum** is always **one**



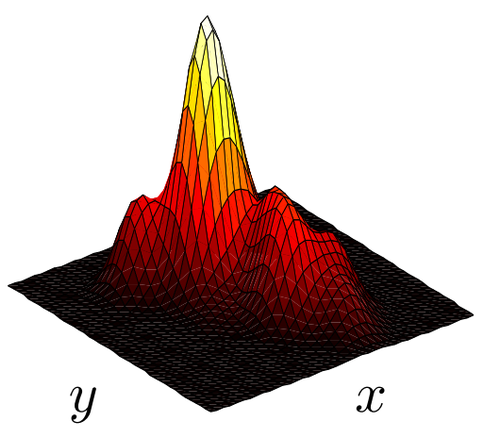
Joint Probability

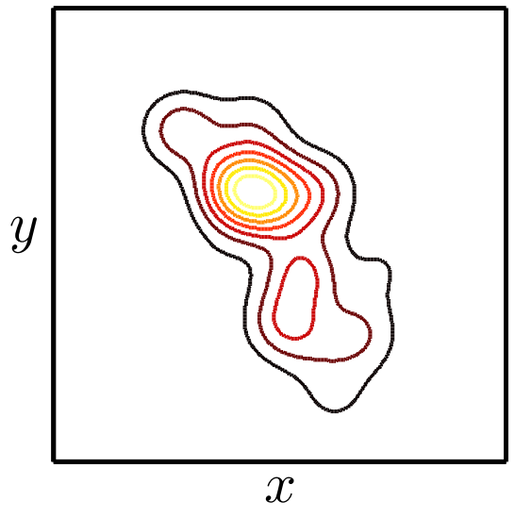
- Consider two random variables x and y
- If we observe multiple paired instances of x and y, then some outcome combinations occur more frequently than others. This is captured in the **joint probability distribution** of x and y, written as $p(x,y)$
- A joint probability distribution may relate variables that are all discrete, all continuous, or mixed discrete-continuous
- Regardless – the **total probability** of all outcomes (obtained by summing or integration) is always **one**
- In general, we can have $p(x,y,z)$. We may also write $p(\mathbf{x})$ to represent the joint probability of all elements in **vector** $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$
- We will write $p(\mathbf{x}, \mathbf{y})$ to represent the joint distribution of all elements from random vectors $\mathbf{x}$ and $\mathbf{y}$

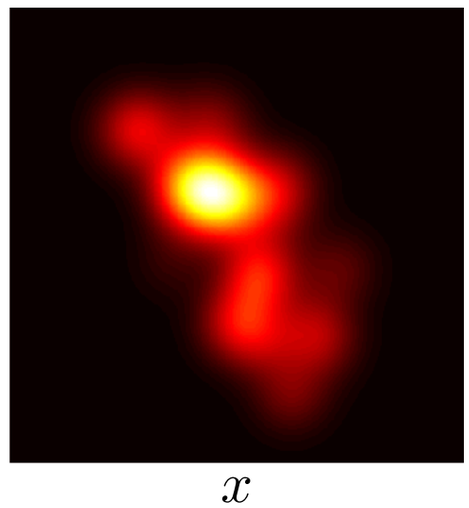
Joint Probability

- Joint probability distribution $p(x,y)$ examples:

Continuous:

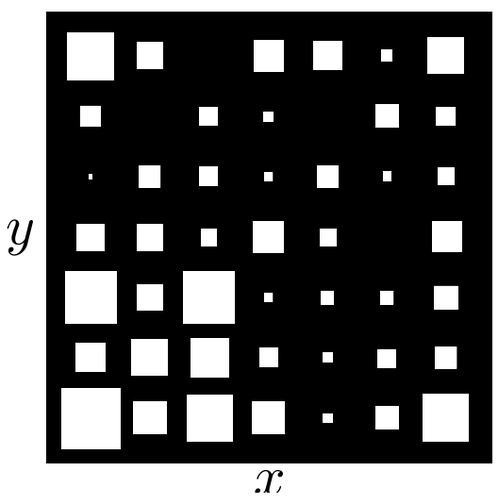




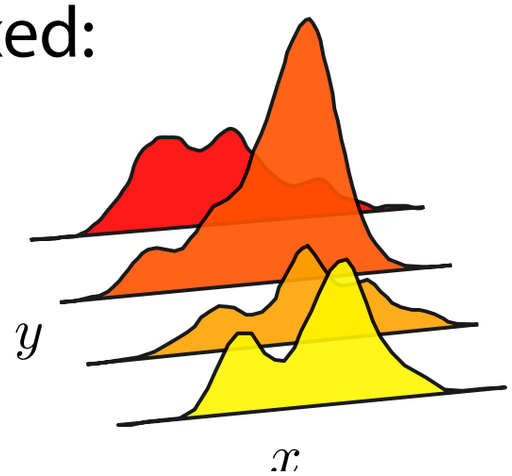


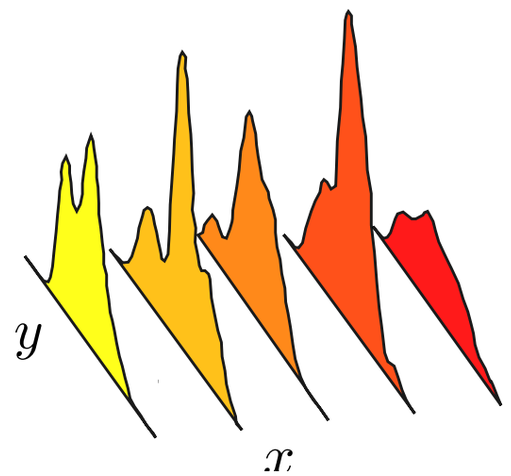
Source [1]

Discrete:



Mixed:





Source [1]

Marginalization

- We can recover the probability distribution of a **single variable** from a joint distribution by **summing over all the other variables**
- Given a continuous $p(x,y)$

$$p(x) = \int p(x, y)\, dy$$

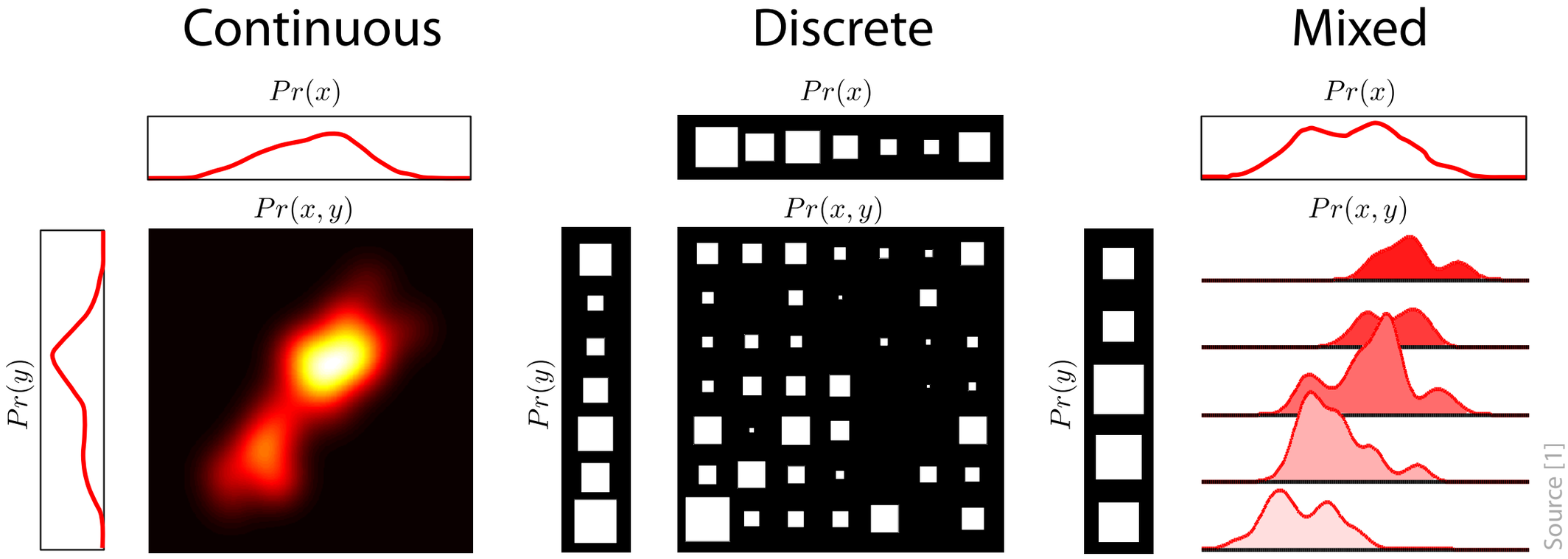
$$p(y) = \int p(x, y)\, dx$$

- The integral becomes a **sum** in the **discrete** case
- Recovered distributions are referred to as **marginal distributions**. The process of integrating/summing is called **marginalization**
- We can recover **any subset of variables**. E.g., given w, x, y, z where w is discrete

$$p(x, y) = \sum_{w} \int p(w, x, y, z)\, dz$$

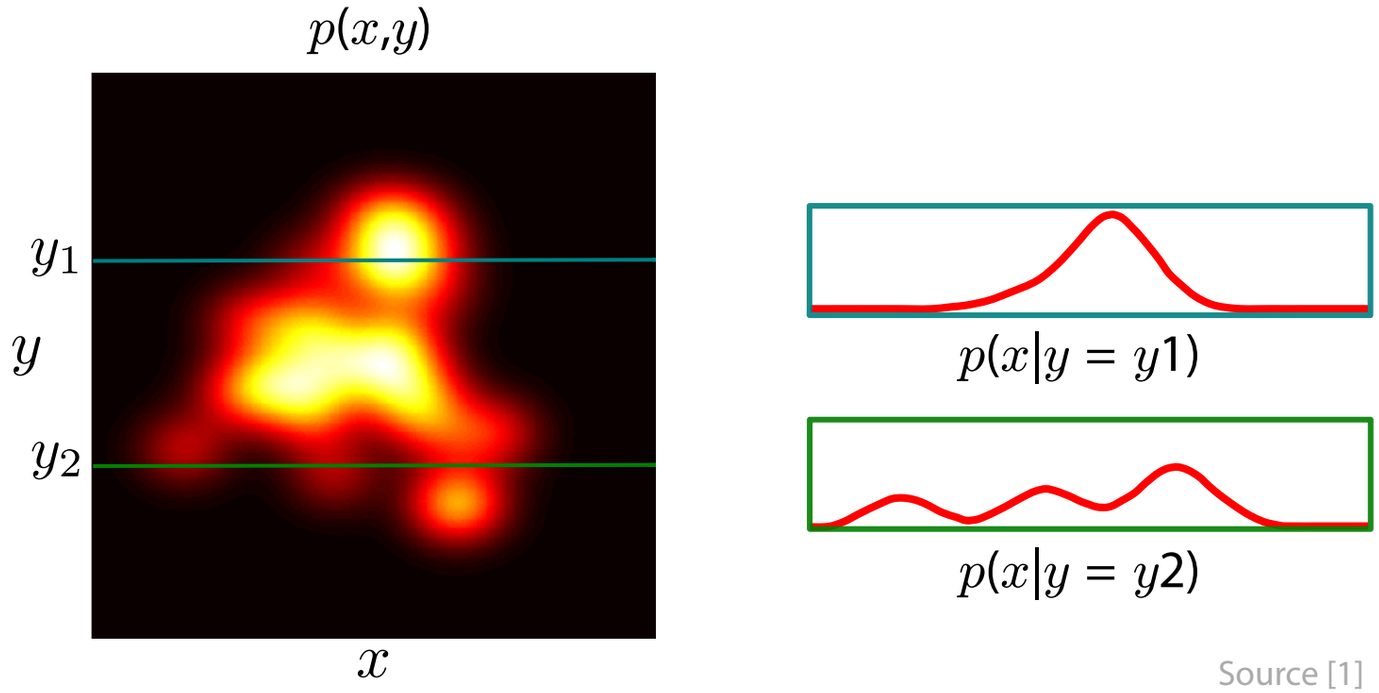
Marginalization

- Calculating the marginal distribution $p(x)$ from $p(x,y)$ has a **simple interpretation**: we are finding the probability distribution of x **regardless of** y (in absence of information about y)
- Marginalization is also known as **sum rule** of **law of total probability**



Conditional Probability

- The **probability of** x **given that** y **takes a fixed value** y^* tells us the relative frequency of x to take different outcomes given the conditioning event that y equal y^*
- This is written $p(x|y = y^*)$ and is called the **conditional probability** of x given y equals y^*
- The conditional probability $p(x|y)$ can be **recovered** from the **joint distribution** $p(x,y)$
- This can be visualized by a **slice** $p(x,y = y^*)$ **through the joint distribution**



Source [1]

Conditional Probability

- The values in the slice tell us about the **relative probability** of x given $y = y^*$, but they do not themselves form a valid probability distribution
- They **cannot sum to one** as they constitute only a small part of $p(x,y)$ which itself sums to one
- To calculate a **proper** conditional probability distribution, we hence **normalize by the total probability** in the slice

$$p(x|y=y^*) = \frac{p(x,y=y^*)}{\int p(x,y=y^*)\,dx} = \frac{p(x,y=y^*)}{p(y=y^*)}$$

where we use marginalization to simplify the denominator

Conditional Probability

- Instead of writing

$$p(x|y=y^*) = \frac{p(x, y=y^*)}{p(y=y^*)}$$

it is common to use a more compact notation and write the conditional probability relation **without explicitly defining the value** $y = y^*$

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

- This can be **rearranged** to give $p(x,y) = p(x|y)\,p(y)$
- By **symmetry** we also have $p(x,y) = p(y|x)\,p(x)$

Bayes' Rule

- In the last two equations, we expressed the joint probability in two ways. When **combining them** we get a **relationship between** $p(x|y)$ **and** $p(y|x)$

$$p(x|y)\,p(y) = p(y|x)\,p(x)$$

- Rearranging gives

$$p(x|y) = \frac{p(y|x)\,p(x)}{p(y)}$$

$$= \frac{p(y|x)\,p(x)}{\int p(x,y)\,dx} = \frac{p(y|x)\,p(x)}{\int p(y|x)\,p(x)\,dx}$$

where we have expanded the denominator using the definition of marginal and conditional probability, respectively

Bayes' Rule

- In the last two equations, we expressed the joint probability in two ways. When **combining them** we get a **relationship between** $p(x|y)$ **and** $p(y|x)$

$$p(x|y)\,p(y) = p(y|x)\,p(x)$$

- Rearranging gives

Bayes' rule

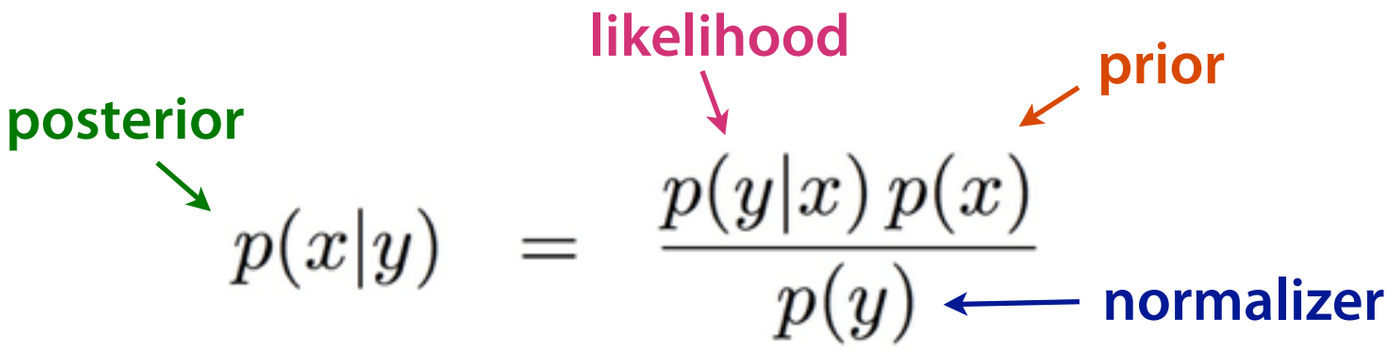
$$p(x|y) = \frac{p(y|x)\,p(x)}{p(y)}$$

$$= \frac{p(y|x)\,p(x)}{\int p(x,y)\,dx} = \frac{p(y|x)\,p(x)}{\int p(y|x)\,p(x)\,dx}$$

where we have expanded the denominator using the definition of marginal and conditional probability, respectively

Bayes' Rule

- Each term in Bayes' rule has a name



(a.k.a. marginal likelihood, ~~evidence~~)

- The **posterior** represents what we know about x given y
- Conversely, the **prior** is what is known about x **before** considering y
- Bayes' rule provides a way to **change your existing beliefs** in the **light of new evidence**. It allows us to combine new data with the existing knowledge or expertise
- Bayes' rule is important in that it allows us to **compute the conditional probability** $p(x|y)$ from the **"inverse"** conditional probability $p(y|x)$

Bayes' Rule Example

Suppose that a tuberculosis (TB) skin test is 95% accurate. That is, if the patient is TB-infected, then the test will be positive with probability 0.95, and if the patient is healthy, then the test will be positive with probability 0.05.

A person gets a positive test result. What is the probability that he is infected?

- **Wanted:** $p(\text{TB}|\text{Positive})$ given $p(\text{Positive}|\text{TB}) = 0.95$, $p(\text{Positive}|\neg\text{TB}) = 0.05$
- **Naive reasoning:** given that the test result is wrong 5% of the time, then the probability that the subject is infected is **0.95**

- **Bayes' rule:** we need to consider the prior probability of TB infection $p(\text{TB})$, and the probability of getting positive test result $p(\text{Positive})$

$$p(\text{TB}|\text{Positive}) = \frac{p(\text{Positive}|\text{TB})\,p(\text{TB})}{p(\text{Positive})}$$

Example from [2]

Human-Oriented Robotics
Prof. Kai Arras
Social Robotics Lab

Bayes' Rule Example (cont.)

- What is the probability of getting a positive test result, $p(\text{Positive})$?
- Let's expand the denominator

$$p(\text{TB}|\text{Positive}) = \frac{p(\text{Positive}|\text{TB})\,p(\text{TB})}{p(\text{Positive})}$$

$$= \frac{p(\text{Positive}|\text{TB})\,p(\text{TB})}{p(\text{Positive}|\text{TB})\,p(\text{TB}) + p(\text{Positive}|\neg\text{TB})\,p(\neg\text{TB})}$$

- Suppose that 1 in 1000 of subjects who get tested is infected: $p(\text{TB})$
- We see that $0.95 \cdot 0.001 = 0.00095$ infected subjects get a positive result, and $0.05 \cdot 0.999 = 0.04995$ uninfected subjects get a positive result. Thus, $p(\text{Positive}) = 0.00095 + 0.04995 = 0.0509$
- Applying Bayes' rule, we obtain $p(\text{TB}|\text{Positive}) = 0.95 \cdot 0.001 \,/\, 0.0509 \approx \mathbf{0.0187}$

Bayes' Rule Example (cont.)

- Wait, **only 2%**?
- This is much more than the prior infection probability of 0.001 – which shows the usefulness of our test – but still…
- **Insights**
  - Our subject was a **random person** for which $p(\mathrm{TB}) = 0.001$ is indeed low
  - Our clinical test is very **inaccurate**, in particular $p(\mathrm{Positive}|\neg\mathrm{TB})$ is high
  - If we set $p(\mathrm{Positive}|\neg\mathrm{TB}) = 0.0001$ (0.1 ‰) leaving all other values the same, we obtain a posterior probability of **0.90**
  - If we set $p(\mathrm{Positive}|\mathrm{TB}) = 0.9999$ leaving all other values the same, we obtain a posterior of **0.0196**
  - If we needed a more accurate result, the **false positive rate** is important

Chain Rule

- Another immediate result of the definition of conditional probability is the **chain rule**

$$p(x, y) = p(x|y)\, p(y)$$

- In **general**,

$$p(x_1, x_2, \ldots, x_K) = p(x_1)\, p(x_2|x_1)\, p(x_3|x_1, x_2) \cdots p(x_k|x_1, x_2, \ldots, x_{K-1})$$

can be compactly expressed as

$$p(x_1, x_2, \ldots, x_K) = \prod_{i=1}^{K} p(x_i|x_1, \ldots, x_{i-1})$$

Chain Rule

- In other words, we can express the **joint probability** of random variables in terms of the probability of the **first**, the probability of the **second given the first**, and so on
- Note that we can expand this expression using **any order of variables**, the result will be the same
- The chain rule is also known as the **product rule**

Independence

- Assume that the value of variable x **tells us nothing about variable** y and vice versa. Formally,

$$p(x|y) = p(x) \qquad p(y|x) = p(y)$$

- Then, we say x and y are **independent**
- When substituting this into the conditional probability relation

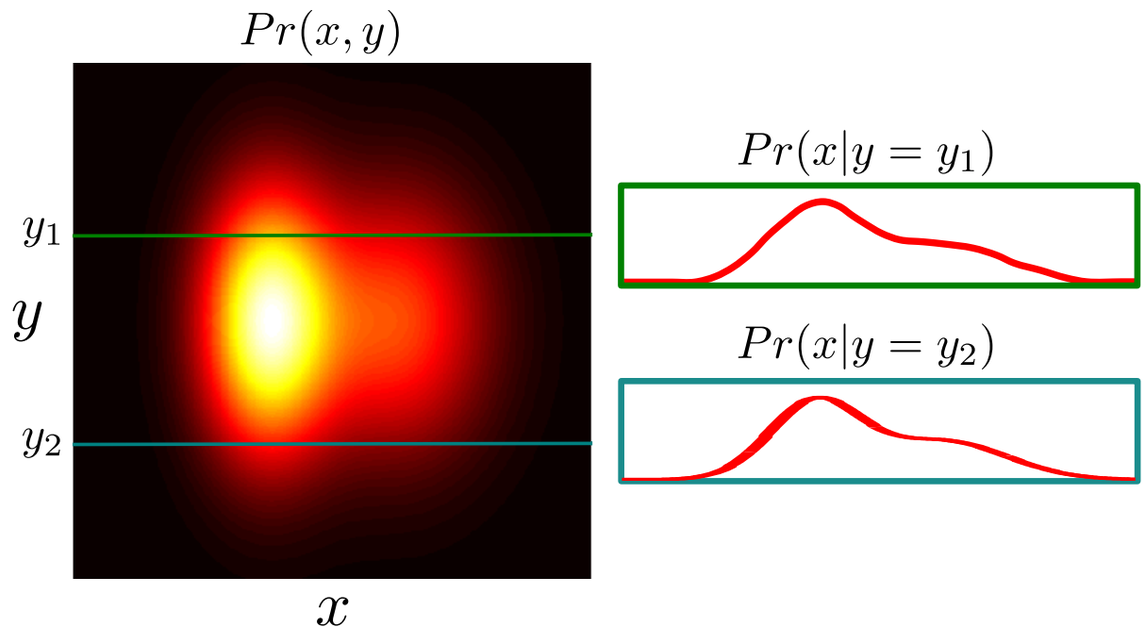
$$p(x,y) = p(x|y)\,p(y)$$

we see that for **independent variables** the joint probability is the **product of the marginal probabilities**

$$\boxed{p(x,y) = p(x)\,p(y)}$$

Independence

- Let us **visualize** this for the joint distribution of two independent variables x and y
- **Independence** of x and y means that **every conditional distribution is the same** (recall that the conditional distribution is the "normalized version of the slice")
- The value of y **tells us nothing** about x and vice versa



Source [1]

Conditional Independence

- While independence is a useful property, it is **not often** that we encounter two independent events. A more common situation is when **two variables are independent given a third one**
- Consider three variables x_1, x_2, x_3. Conditional independence is written as

$$p(x_1|x_2, x_3) = p(x_1|x_2)$$
$$p(x_3|x_1, x_2) = p(x_3|x_2)$$

  and implies that if we know x_2, then x_1 provides **no further information** about x_3 (and vice versa)
- Note that when x_1 and x_3 are conditionally independent given x_2, this does **not** mean that x_1 and x_3 are **themselves independent**.
- Typically occurs in **chain of events:** if x_1 causes x_2 and x_2 causes x_3, then the dependence of x_3 on x_1 is entirely "contained" in x_2

Conditional Independence

- Example: entering a hip nightclub
  - Suppose we want to reason about the chance that a student enters the two hottest nightclubs in town. Denote *A* the event "student passes bouncer of club A", and *B* the event "student passes bouncer of club B"
  - Usually, these two events are **not independent** because if we learn that the student could enter club B, then our estimate of his/her probability of entering club A is higher since it is a sign that the student is hip, properly dressed and not too drunk
  - Now suppose that the doormen base their decisions **only** on the looks of the student's company, and we know their preferences. Thus, learning that event *B* has occurred should not change the probability of event *A:* the looks of the company contains **all relevant information** to his/her chances of passing. Finding out whether he/she could enter club *B* **does not change** that
  - Formally, $p(\mathrm{A}|\mathrm{B},\mathrm{Looks}) = p(\mathrm{A}|\mathrm{Looks})$
- In this case, we say A **is conditionally independent of** B **given** Looks

Conditional Independence

- Example: rolling a blue and red die
  - The two results are **independent** of each other

  - Now someone tells you **"the blue result isn't a 6 and the red result isn't a 1"**
  - From this information, you cannot gain any knowledge about the red die by looking at the blue die. The probability for each number except 1 on the red one is still 1/5
  - The information does not affect the independence of the results

  - Now someone tells you **"the sum of the two results is even"**
  - This allows you to learn a lot about the red die by looking at the blue die
  - For instance, if you see a 3 on the blue die, the red die can only be 1, 3 or 5
  - The result probabilities are **not conditionally independent** given this information

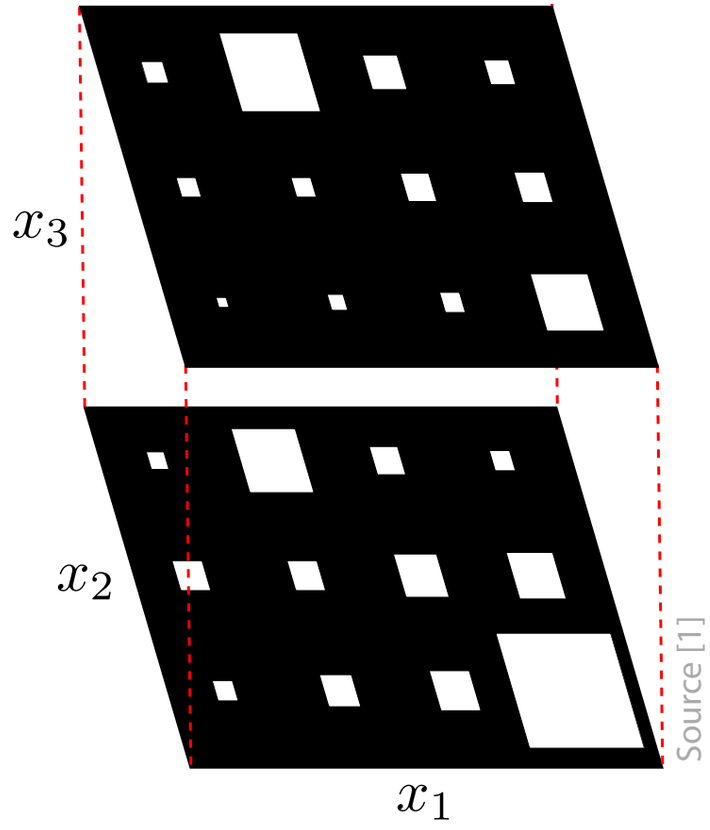
- Conditional independence is always **relative to the given condition**

Conditional Independence

- Variable x_1 is said to be **conditional independent of variable** x_3 **given variable** x_2 if – given **any** value of x_2 – the probability distribution of x_1 is the same for all values of x_3 and the probability distribution of x_3 is the same for all values of x_1

$$Pr(x_1, x_2, x_3)$$

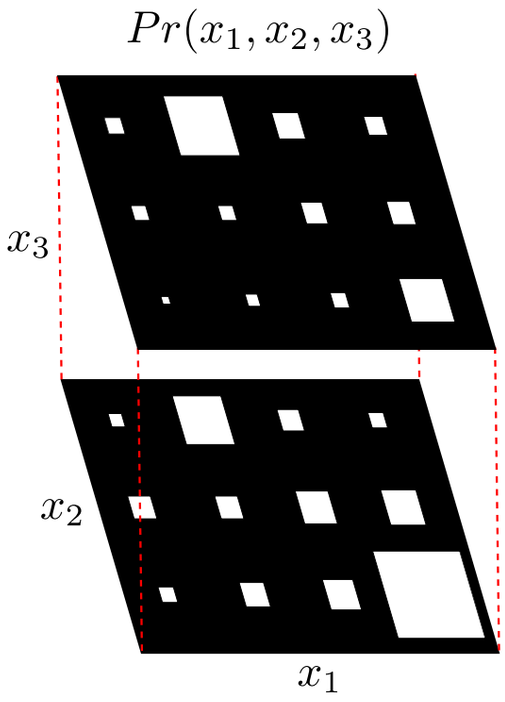
- Let us look at a **graphical example**
- Consider the joint density function of three discrete random variables x_1, x_2, x_3 which take 4, 3, and 2 possible values, respectively
- All 24 probabilities sum to **one**

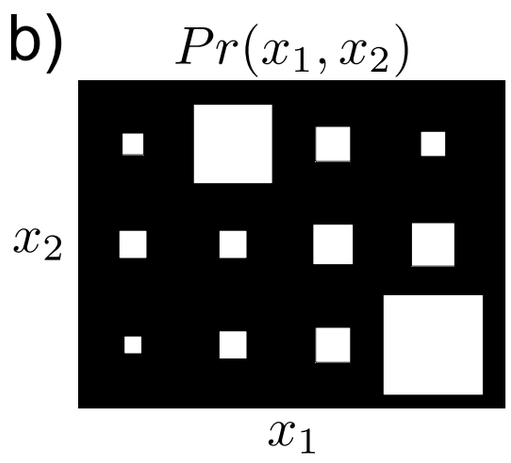


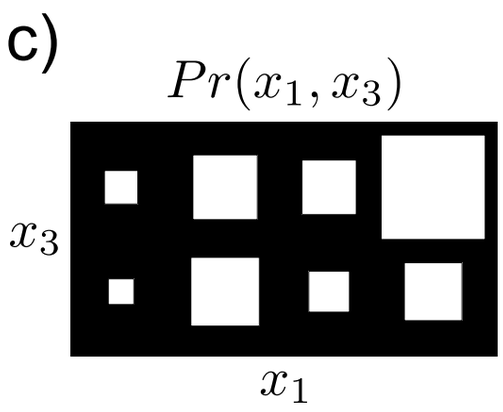
Conditional Independence

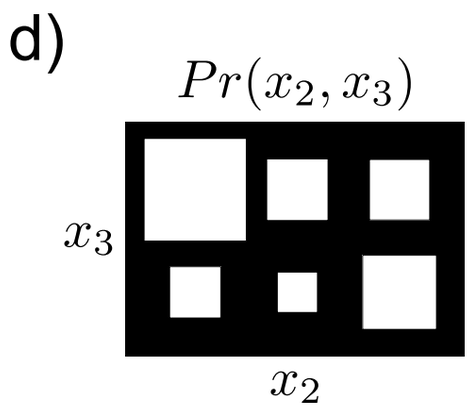
First, let's consider independence:

- Figure b, marginalization of x_3 : no independence between x_1 and x_2
- Figure c, marginalization of x_2 : no independence between x_1 and x_3
- Figure d, marginalization of x_1 : no independence between x_2 and x_3







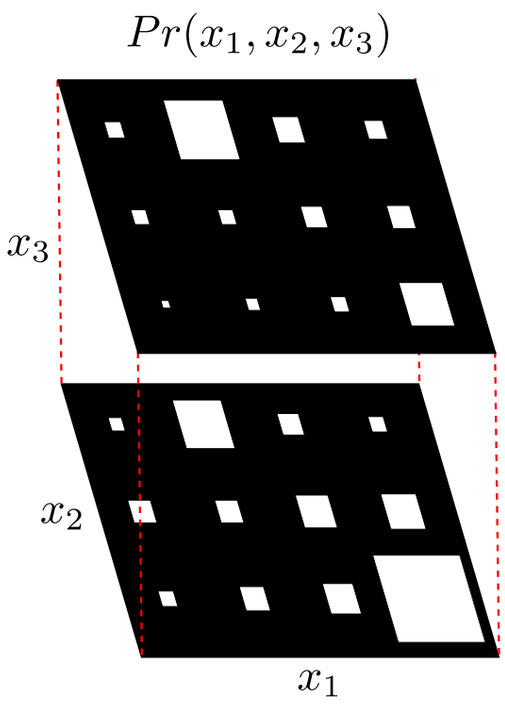


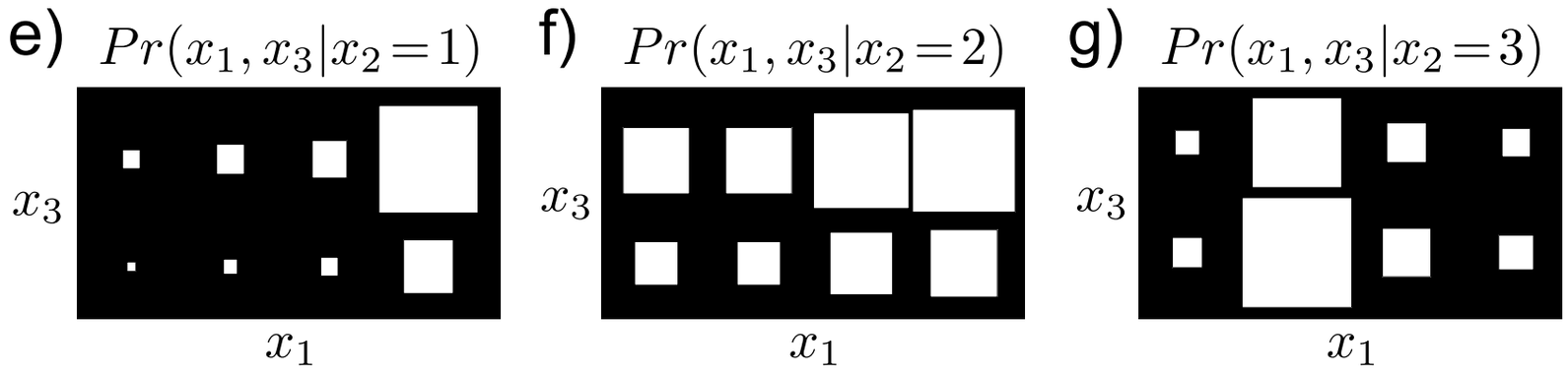
Source [1]

Conditional Independence

Now let's consider **conditional** independence given x_2

- Figures e, f, g: value of x_2 is fixed at 1, 2, 3 respectively





Source [1]

- For fixed x_2, variable x_1 **tells us nothing more** about x_3 and vice versa
- Thus, x_1 and x_3 are **conditionally independent given** x_2

Expectation

- Intuitively, the expected value of a random variable is the value one would "expect" to find if one could **repeat the random variable process an infinite number of times** and take the average of the values obtained
- Let x be a discrete random variable, then the **expectation of x under the distribution** p is

$$\mathrm{E}[x] = \sum_x x \cdot p(x)$$

- In the continuous case, we use density functions and integrals

$$\mathrm{E}[x] = \int x \cdot p(x)\, dx$$

- It is a **weighted average** of all possible values where the weights are the corresponding values of the probability mass/density function

Expectation

- For example, if x models the outcome of rolling a fair die, then

$$\mathrm{E}[x] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = 3.5$$

- With a biased die where $p(x = 6) = 0.5$ and $p(x = x^*) = 0.1$ for $x^* < 6$, then

$$\mathrm{E}[x] = 1 \cdot 0.1 + \cdots + 5 \cdot 0.1 + 6 \cdot 0.5 = 4.5$$

- Often, we are interested in expectations of a **function of random variables**. Thus, we extend the definition to

$$\mathrm{E}[f(x)] = \sum_x f(x) \cdot p(x)$$

$$\mathrm{E}[f(x)] = \int f(x) \cdot p(x)\, dx$$

Expectation

- This idea also generalizes to functions of **more than one** variable

$$\mathrm{E}[f(x,y)] = \iint f(x,y) \cdot p(x,y)\, dx\, dy$$

- Note however, that any function g of a set of a random variable x, or a set of variables $(x_1, x_2, \ldots, x_k)$ is essentially a **new** random variable y
- For some choices of function f, the expectation is given a **special name**

Function $f(x)$, $f(x,y)$	Expectation
x	mean μ_x
x^k	k-th moment about zero
$(x-\mu_x)^k$	k-th central moment
$(x-\mu_x)^2$	variance
$(x-\mu_x)^3$	skew
$(x-\mu_x)^4$	kurtosis
$(x-\mu_x)(y-\mu_y)$	covariance of x and y

Skew and kurtosis are also defined as **standardized moments**

$$\frac{(x-\mu_x)^k}{\sigma^k}$$

Expectation

- The expected value of a specified integer power of the deviation of the random variable from the mean is called **central moment** or **moment about the mean** of a probability distribution

$$\mu_x^k = \mathrm{E}[(x - \mathrm{E}[x])^k] = \int_{-\infty}^{+\infty} (x - \mu_x)^k \cdot p(x)\, dx$$

- Ordinary moments (or raw moments) are defined **about zero**
- Moments are used to **characterize the shape** of a distribution
  - The mean is the first raw moment. It's actually a location measure
  - The variance describes the distribution's **width** or **spread**
  - The skew describes – loosely speaking – the extent to which a probability distribution **"leans" to one side** of the mean. A measure of asymmetry
  - The kurtosis is a measure of the **"peakedness"** of the probability distribution

Expectation

- There are four **rules for manipulating expectations**, which can be easily proved from the original definition
- Expected value of a **constant**

$$\mathrm{E}[a] = a$$

- Expected value of a **constant times** a random variable

$$\mathrm{E}[a \cdot x] = a\,\mathrm{E}[x] \qquad \text{thus} \quad \mathrm{E}[a \cdot x + b] = a\,\mathrm{E}[x] + b$$

- Expected value of the **sum** of two random variables

$$\mathrm{E}[x + y] = \mathrm{E}[x] + \mathrm{E}[y]$$

- Expected value of the **product** of two random variables

$$\mathrm{E}[x \cdot y] = \mathrm{E}[x] \cdot \mathrm{E}[y] \qquad \text{if } x,y \text{ are independent}$$

Human-Oriented Robotics
Prof. Kai Arras
Social Robotics Lab

Expectation

- These properties also apply to **functions of random variables**
- Expected value of a **constant**

$$\mathrm{E}[a] = a$$

- Expected value of a **constant times** a function

$$\mathrm{E}[a \cdot f(x)] = a\,\mathrm{E}[f(x)] \qquad \text{thus } \mathrm{E}[af(x)+b] = a\,\mathrm{E}[f(x)]+b$$

- Expected value of the **sum** of two functions

$$\mathrm{E}[f(x) + g(x)] = \mathrm{E}[f(x)] + \mathrm{E}[g(x)]$$

- Expected value of the **product** of two functions

$$\mathrm{E}[f(x) \cdot g(x)] = \mathrm{E}[f(x)] \cdot \mathrm{E}[g(x)] \qquad \text{if } x,y \text{ are independent}$$

Variance

- The variance is the **second central moment**, defined as

$$\mathrm{Var}[x] = \mathrm{E}[(x - \mathrm{E}[x])^2] = \int_{-\infty}^{+\infty} (x - \mu_x)^2 \cdot p(x)\, dx$$

- Alternative formulation $\mathrm{Var}[x] = \mathrm{E}[x^2] - (\mathrm{E}[x])^2$

- Its square root is called the **standard deviation** $\sigma_x = \sqrt{\mathrm{Var}[x]}$

- The **rules for manipulating** variances are as follows

  Variance of a **linear function**

$$\mathrm{Var}[a \cdot x + b] = a^2 \cdot \mathrm{Var}[x]$$

  Variance of a **sum** of random variables

$$\mathrm{Var}[x + y] = \mathrm{Var}[x] + \mathrm{Var}[y] \qquad \text{if } x,y \text{ are independent}$$

Probability Refresher

- Introduction to Probability
  - Random variables
  - Joint distribution
  - Marginalization
  - Conditional probability
  - Chain rule
  - Bayes' rule
  - Independence
  - Conditional independence
  - Expectation and Variance
- **Common Probability Distributions**
  - Bernoulli distribution
  - Binomial distribution
  - Categorial distribution
  - Multinomial distribution
  - Poisson distribution
  - Gaussian distribution
  - Chi-squared distribution

We assume that you are familiar with the **fundamentals** of probability theory and probability distributions

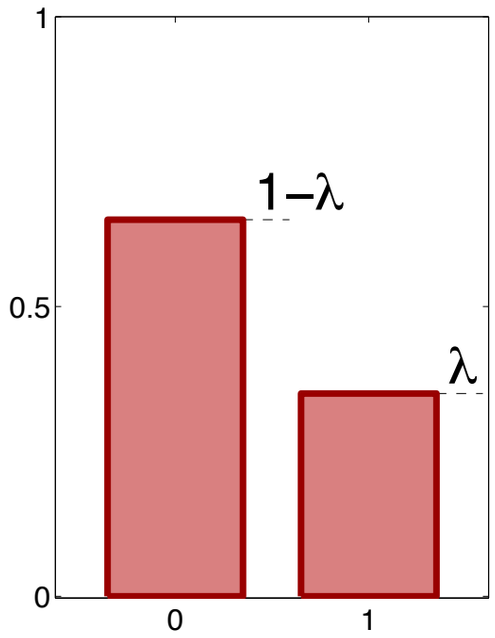
This is a quick refresher, we aim at **ease of understanding** rather than formal depth

For a more comprehensive treatment, refer, e.g. to A. Papoulis or the references on the last slide

Common Probability Distributions

Bernoulli Distribution

- Given a **Bernoulli experiment**, that is, a **yes/no experiment** with outcomes **0** (“failure”) or **1** (“success”)
- The Bernoulli distribution is a **discrete** probability distribution, which takes value 1 with success probability λ and value 0 with failure probability 1 – λ



- **Probability mass function**

$$\left.\begin{aligned} p(x=0) &= 1-\lambda \\ p(x=1) &= \lambda \end{aligned}\right\} \quad p(x) = \lambda^x (1-\lambda)^{1-x}$$

- Notation

$$\mathrm{Bern}_x(\lambda) = \lambda^x (1-\lambda)^{1-x}$$

Parameters

- λ : probability of observing a success

Expectation

- $\mathrm{E}[x] = \lambda$

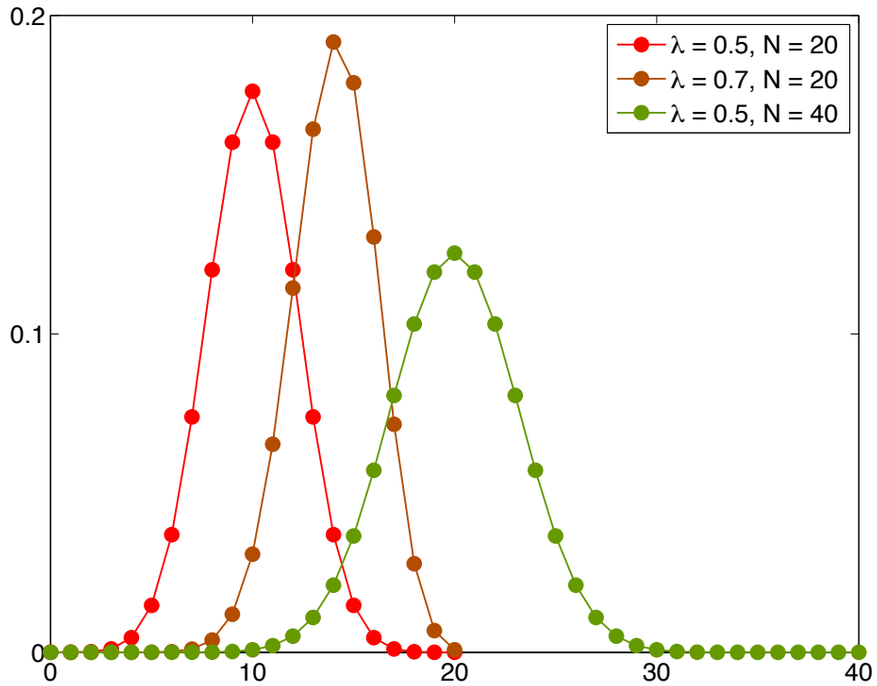
Variance

- $\mathrm{Var}[x] = \lambda(1-\lambda)$

Common Probability Distributions

Binomial Distribution

- Given a **sequence** of Bernoulli experiments
- The binomial distribution is the **discrete** probability distribution of the **number of successes** m in a **sequence** of N independent yes/no experiments, each with a success probability of λ



- **Probability mass function**

$$p(m) = \binom{N}{m} \lambda^m (1-\lambda)^{N-m}$$

- Notation

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

Parameters

- N : number of trials
- λ : success probability

Expectation

- $\mathrm{E}[m] = N\,\lambda$

Variance

- $\mathrm{Var}[m] = N\,\lambda(1-\lambda)$

Common Probability Distributions

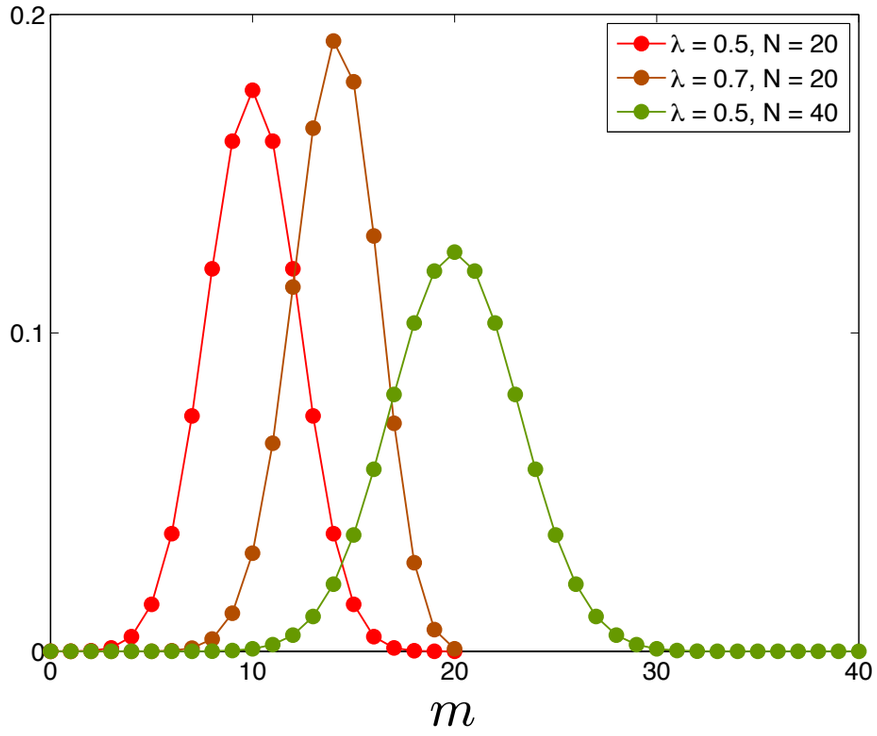
Binomial Distribution

- The quantity

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

is the binomial coefficient ("N choose m") and denotes the **number of ways of choosing** m objects out of a total of N identical objects

- For $N = 1$, the binomial distribution is the **Bernoulli distribution**
- For fixed expectation $N\,\lambda$, the Binomial converges towards the **Poisson distribution** as N goes to infinity



Parameters

- N : number of trials
- λ : success probability

Expectation

- $\mathrm{E}[m] = N\,\lambda$

Variance

- $\mathrm{Var}[m] = N\,\lambda(1-\lambda)$

Common Probability Distributions

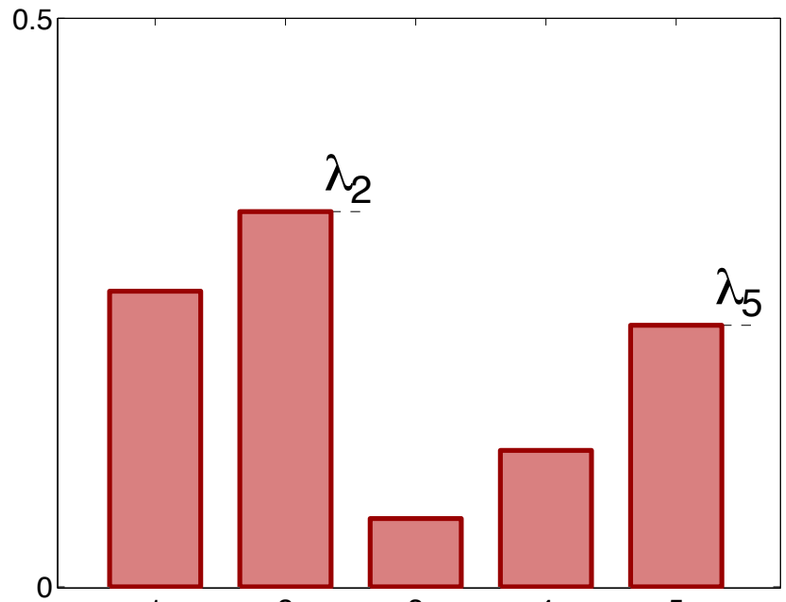
Categorial Distribution

- Considering a single experiment with K **possible outcomes**
- The categorial distribution is a **discrete** distribution that describes the probability of observing **one of K possible outcomes**
- **Generalizes** the Bernoulli distribution
- The probability of each outcome is specified as $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_K]$ with $\sum_{k=1}^{K} \lambda_k = 1$
- **Probability mass function**

$$p(x = k) = \lambda_k$$

- Notation

$$\mathrm{Cat}_x(\boldsymbol{\lambda}) = p(x)$$



Parameters

- $\boldsymbol{\lambda}$: vector of outcome probabilities

Expectation

- $\mathrm{E}[x = k] = \lambda_k$

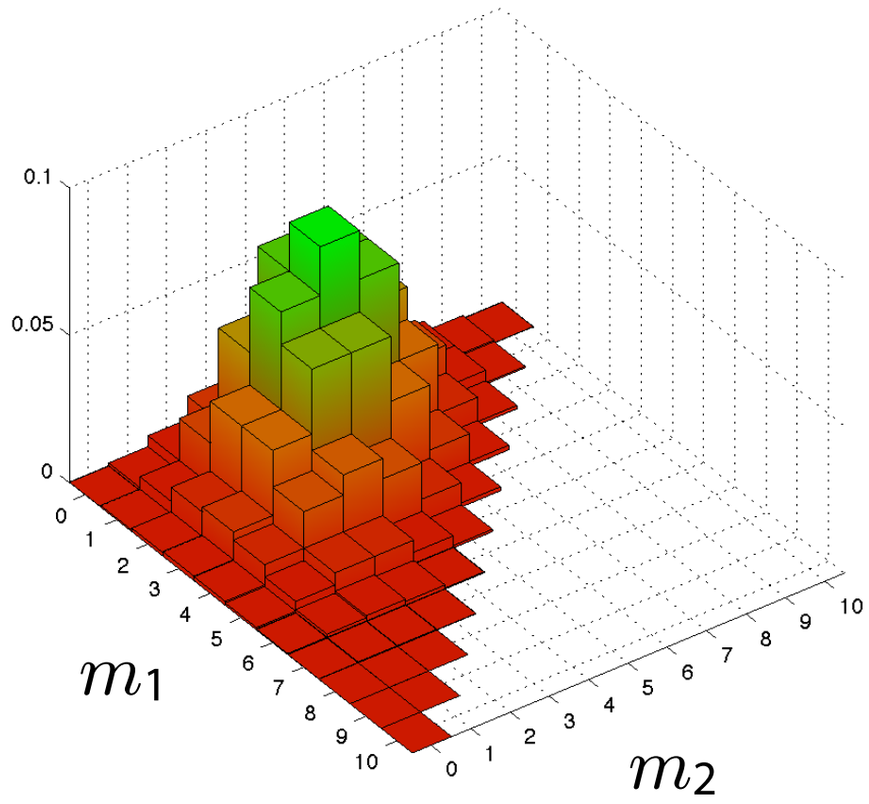
Variance

- $\mathrm{Var}[x = k] = \lambda_k(1 - \lambda_k)$

Common Probability Distributions

Multinomial Distribution

- Given a **sequence** of experiments, each with K **possible outcomes**
- The multinomial distribution is the **discrete** probability distribution of the **number of observations of values** {1,2,...,K} with **counts** $\{m_1, m_2, \ldots, m_K\}$ in a **sequence** of N independent trials
- In other words:
  For N independent trials each of which leads to a success for exactly one of K categories, the multinomial distribution gives the **probability of a combination of numbers of successes** for the various categories



Parameters

- N : number of trials
- $\boldsymbol{\lambda}$: success probabilities

Expectation

- $\mathrm{E}[m_k] = N\lambda_k$

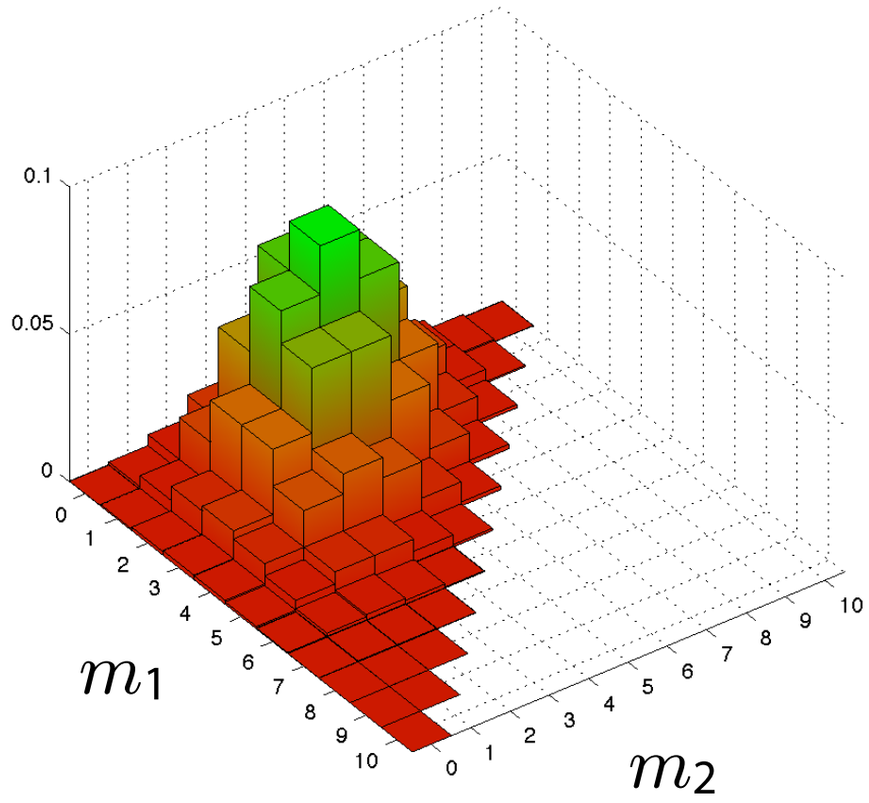
Variance

- $\mathrm{Var}[m_k] = N\lambda_k(1-\lambda_k)$

Common Probability Distributions

Multinomial Distribution

- Each category has a given fixed **success probability** $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_K]$ subject to $\lambda_1 + \lambda_2 + \cdots + \lambda_K = 1$



- **Probability mass function**

$$p(m) = \binom{N}{m_1\, m_2 \cdots m_K} \lambda_1^{m_1} \lambda_2^{m_2} \cdots \lambda_K^{m_K}$$

- Notation

$$\mathrm{Mult}_{\mathbf{m}}(N, \boldsymbol{\lambda}) = p(m)$$

with $\mathbf{m} = \{m_1, m_2, \ldots, m_K\}$

Parameters

- N : number of trials
- $\boldsymbol{\lambda}$: success probabilities

Expectation

- $\mathrm{E}[m_k] = N\lambda_k$

Variance

- $\mathrm{Var}[m_k] = N\lambda_k(1 - \lambda_k)$

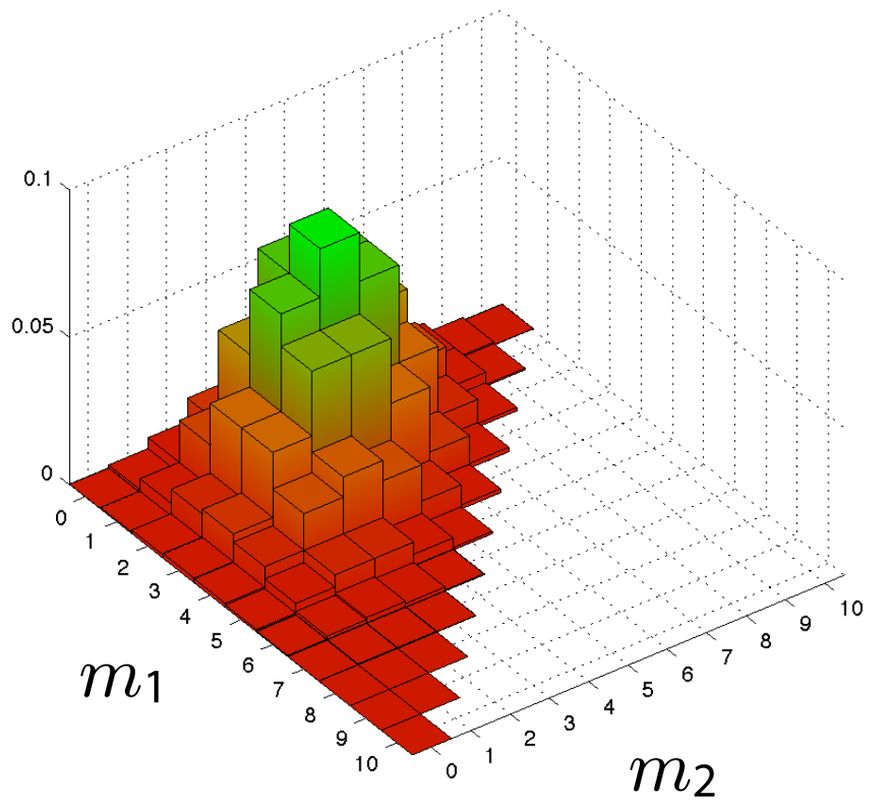
Common Probability Distributions

Multinomial Distribution

- The quantity

$$\binom{N}{m_1\, m_2 \cdots m_K} = \frac{N!}{m_1! m_2! \cdots m_K!}$$

is the multinomial coefficient and denotes the **number of ways of taking** N identical objects and **assigning** m_k of them to bin k



- **Generalizes** the **binomial** distribution to K outcomes
- **Generalizes** the **categorial** distribution to sequences of N trials

Parameters

- N : number of trials
- $\boldsymbol{\lambda}$: success probabilities

Expectation

- $\mathrm{E}[m_k] = N\lambda_k$

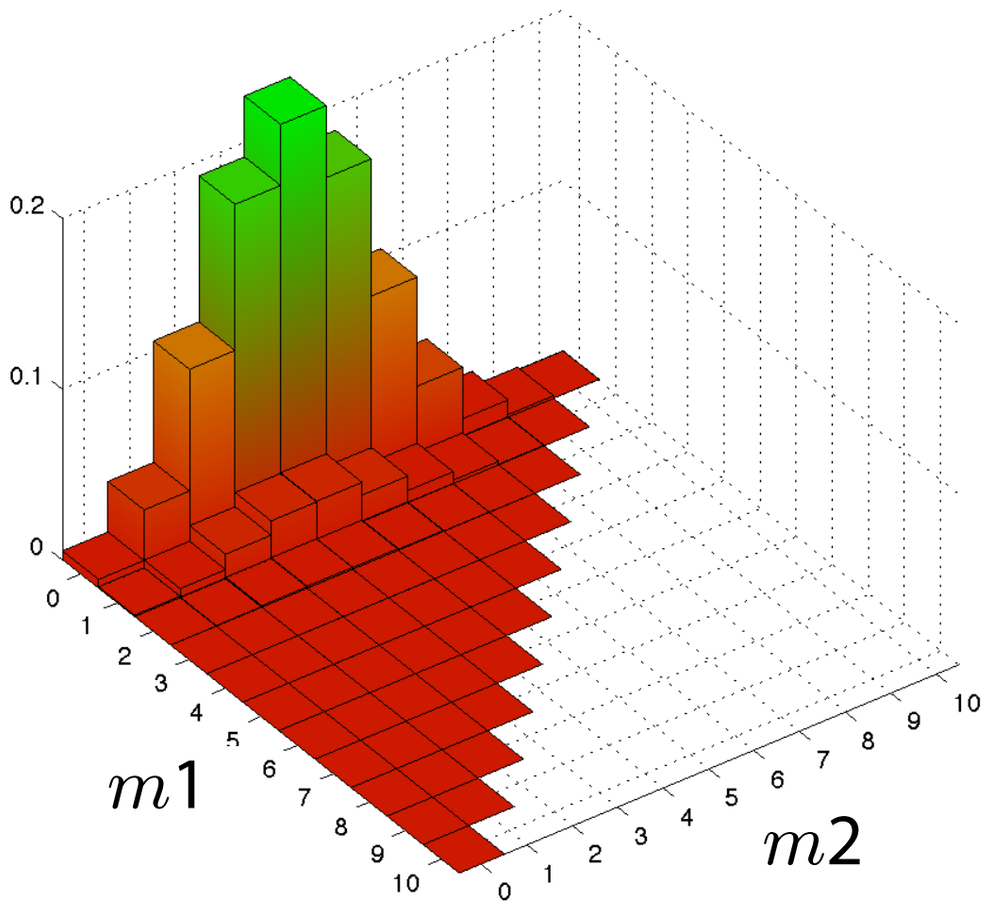
Variance

- $\mathrm{Var}[m_k] = N\lambda_k(1-\lambda_k)$

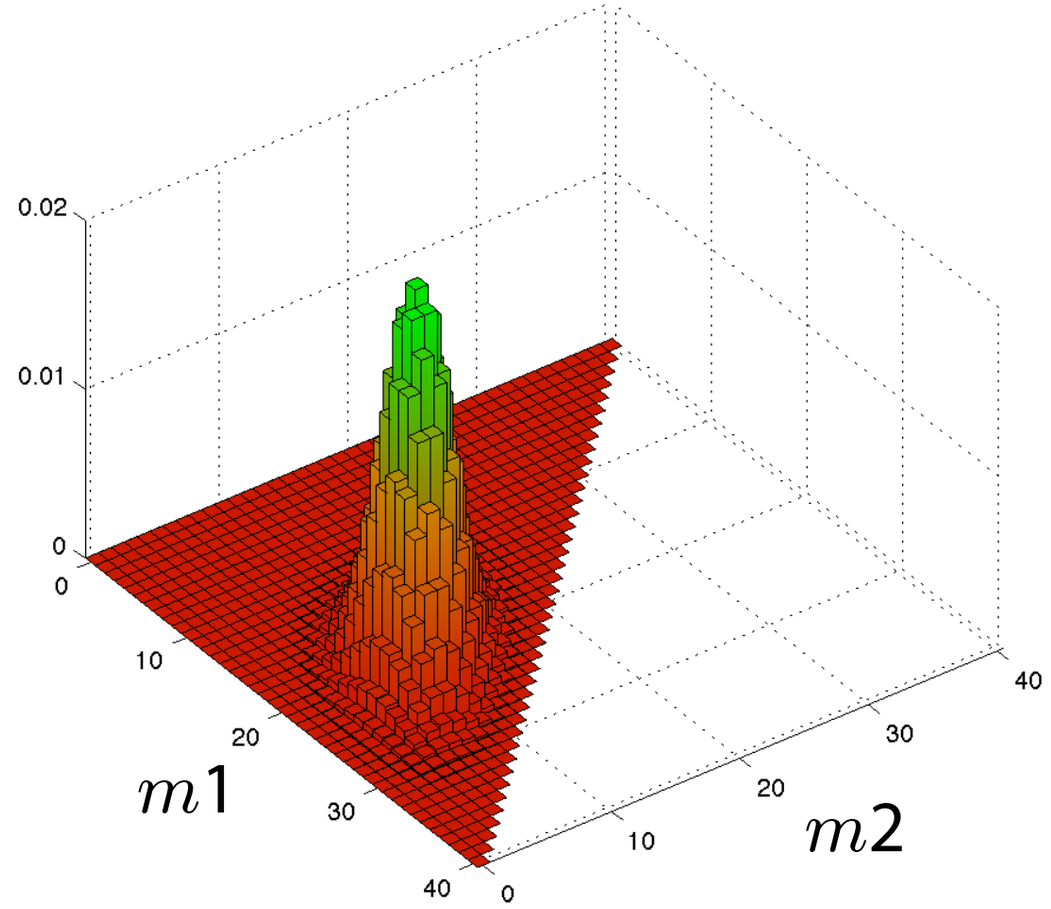
Common Probability Distributions

Multinomial Distribution

- $N = 10$, $\lambda_1 = 0.01$, $\lambda_2 = 0.4$, $\lambda_3 = 0.49$
- Maximum at $m_1 = 1, m_2 = 4$
- Showing successes for m_1, m_2



- $N = 40$, $\lambda_1 = 0.5$, $\lambda_2 = 0.25$, $\lambda_3 = 0.25$
- Maximum at $m_1 = 20, m_2 = 10$
- Showing successes for m_1, m_2



Common Probability Distributions

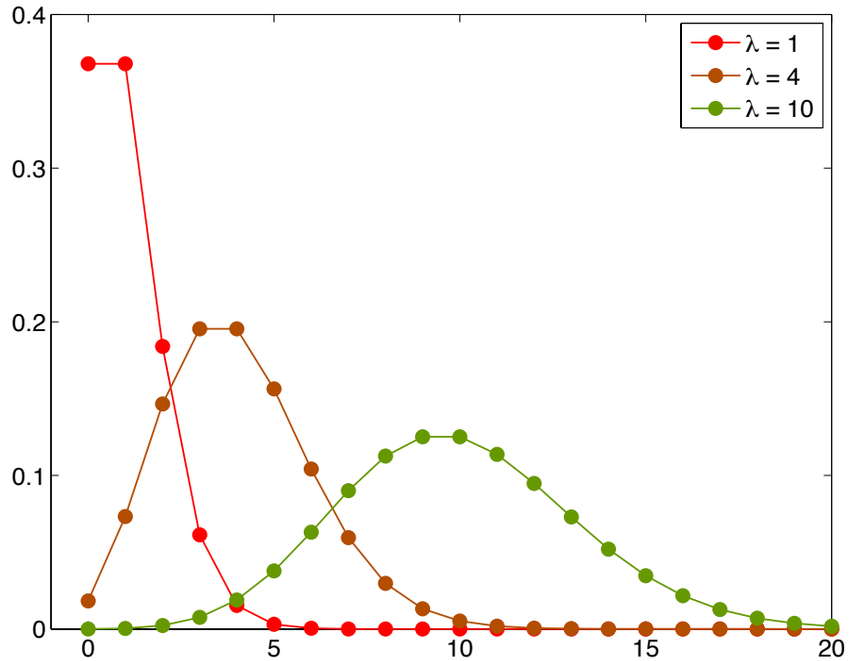
Poisson Distribution

- Consider independent **events** that **happen with an average rate** of λ over time
- The Poisson distribution is a **discrete** distribution that describes the **probability** of a given **number of events** occurring in a fixed interval of time
- Can also be defined over other intervals such as **distance**, **area** or **volume**
- **Probability mass function**

$$p(x) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- Notation

$$\mathrm{Pois}_x(\lambda) = p(x)$$



Parameters

- λ : average rate of events over time or space

Expectation

- $\mathrm{E}[x] = \lambda$

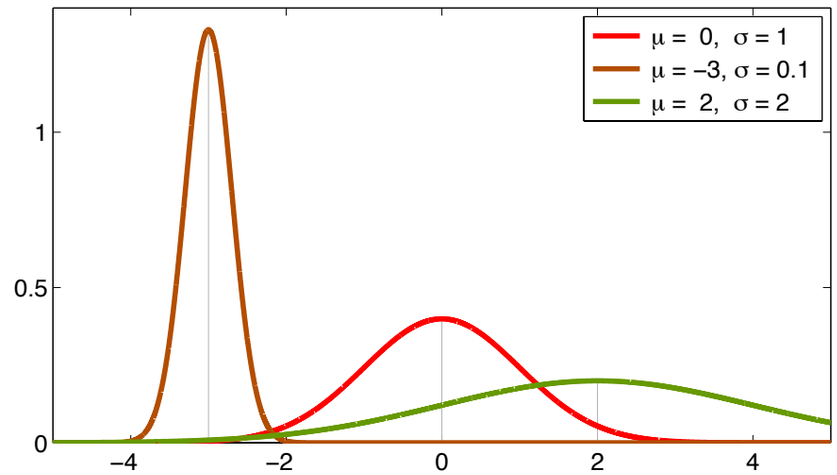
Variance

- $\mathrm{Var}[x] = \lambda$

Common Probability Distributions

Gaussian Distribution

- **Most widely** used distribution for **continuous** variables
- Reasons: (i) **simplicity** (fully represented by only two moments, mean and variance) and (ii) the **central limit theorem** (CLT)



- The CLT states that, under mild conditions, the **mean** (or sum) of many independently drawn random variables is distributed approximately **normally**, irrespective of the form of the original distribution
- **Probability density function**

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Parameters

- μ : mean
- σ^2 : variance

Expectation

- $\mathrm{E}[x] = \mu$

Variance

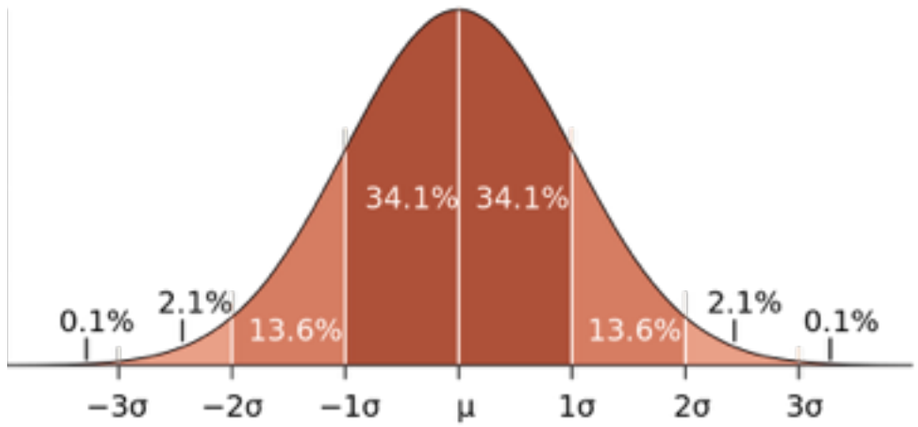
- $\mathrm{Var}[x] = \sigma^2$

Common Probability Distributions

Gaussian Distribution

- Notation

$$\mathcal{N}_x(\mu, \sigma^2) = p(x)$$



- Called **standard normal distribution** for $\mu = 0$ and $\sigma = 1$
- **About 68%** (~two third) of values drawn from a normal distribution are within a **range of ±1 standard deviations** around the mean
- **About 95%** of the values lie within a **range of ±2 standard deviations** around the mean
- Important e.g. for **hypothesis testing**

Parameters

- μ : mean
- σ^2 : variance

Expectation

- $\mathrm{E}[x] = \mu$

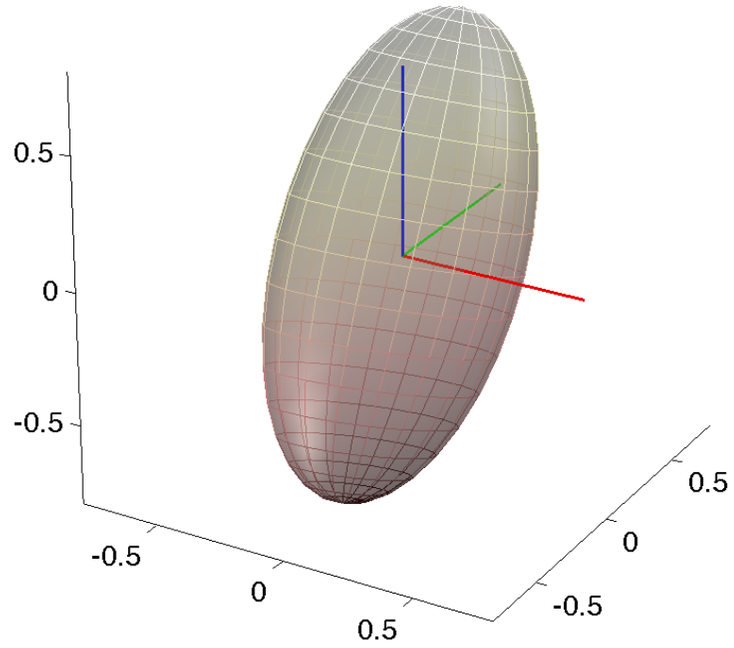
Variance

- $\mathrm{Var}[x] = \sigma^2$

Common Probability Distributions

Multivariate Gaussian Distribution

- For d-dimensional random vectors, the **multivariate Gaussian distribution** is governed by a d-dimensional **mean vector** $\boldsymbol{\mu}$ and a D x D **covariance matrix** Σ that must be symmetric and positive semi-definite



- **Probability density function**

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

- Notation

$$\mathcal{N}_x(\boldsymbol{\mu}, \Sigma) = p(\mathbf{x})$$

Parameters

- $\boldsymbol{\mu}$: mean vector
- Σ: covariance matrix

Expectation

- $\mathrm{E}[\mathbf{x}] = \boldsymbol{\mu}$

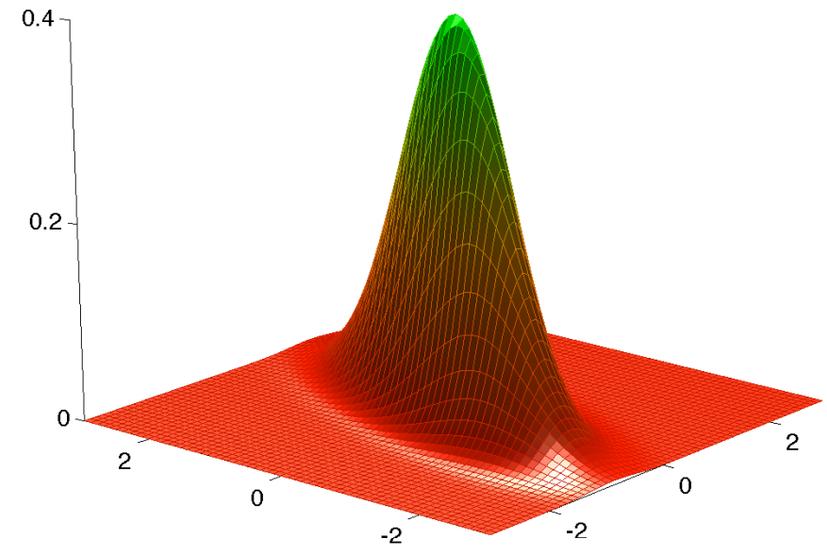
Variance

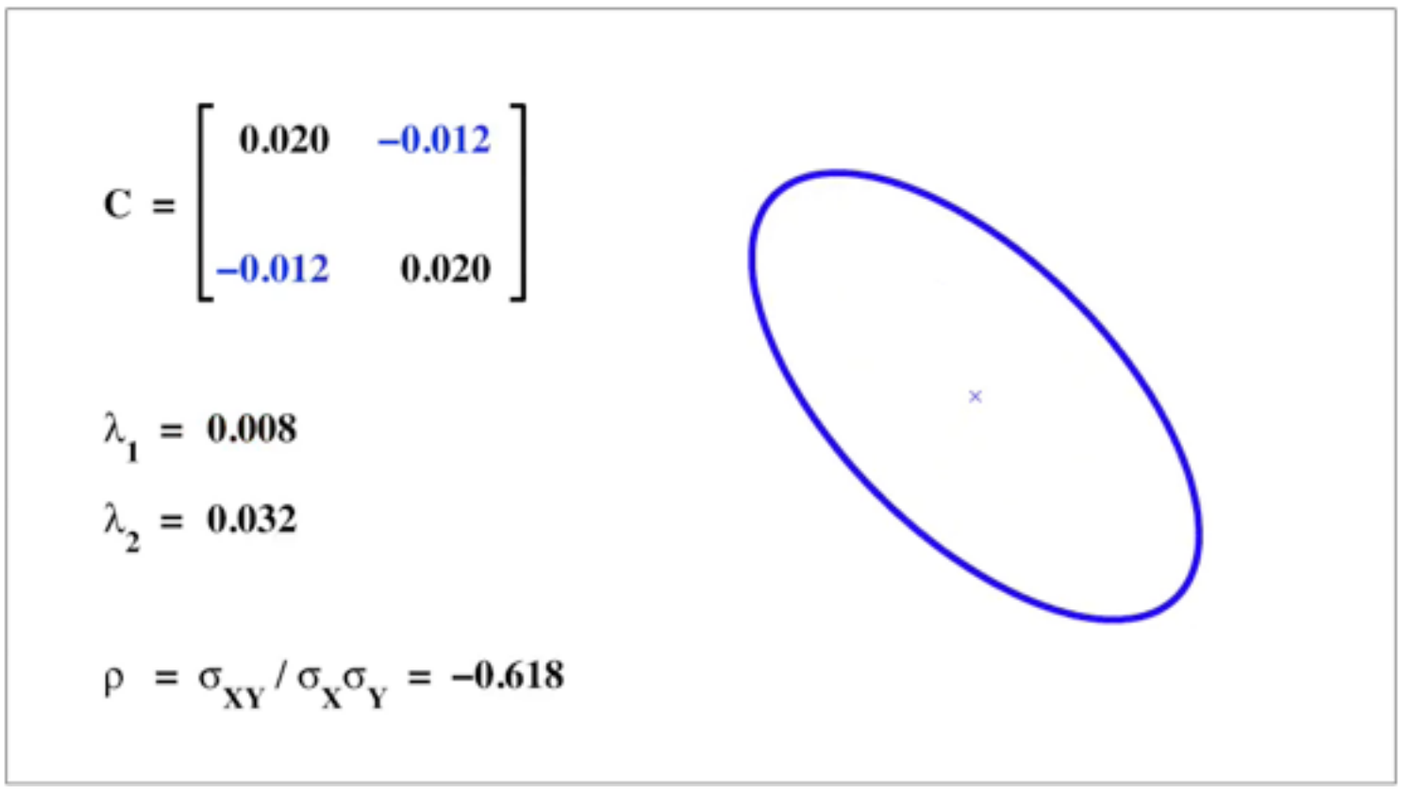
- $\mathrm{Var}[\mathbf{x}] = \Sigma$

Common Probability Distributions

Multivariate Gaussian Distribution

- For $d = 2$, we have the **bivariate** Gaussian distribution
- The covariance matrix Σ (often C) determines the **shape of the distribution** (video)





Parameters

- $\boldsymbol{\mu}$: mean vector
- Σ: covariance matrix

Expectation

- $\mathrm{E}[\mathbf{x}] = \boldsymbol{\mu}$

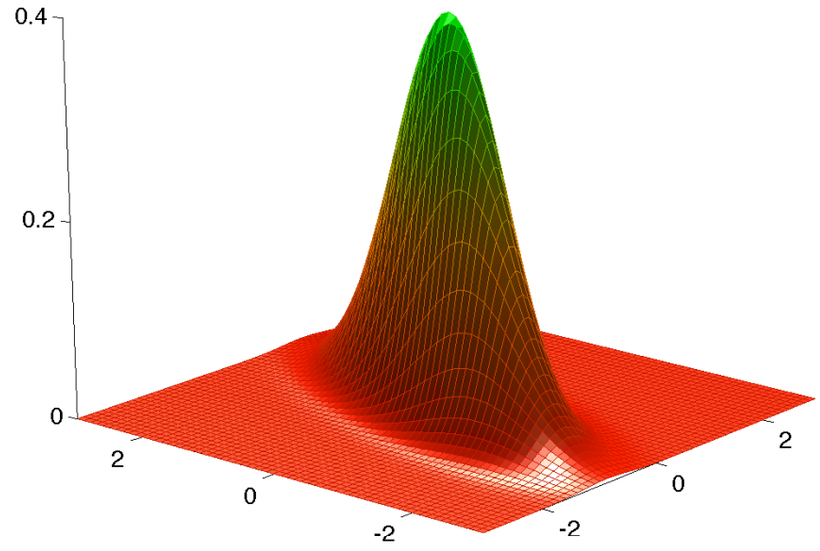
Variance

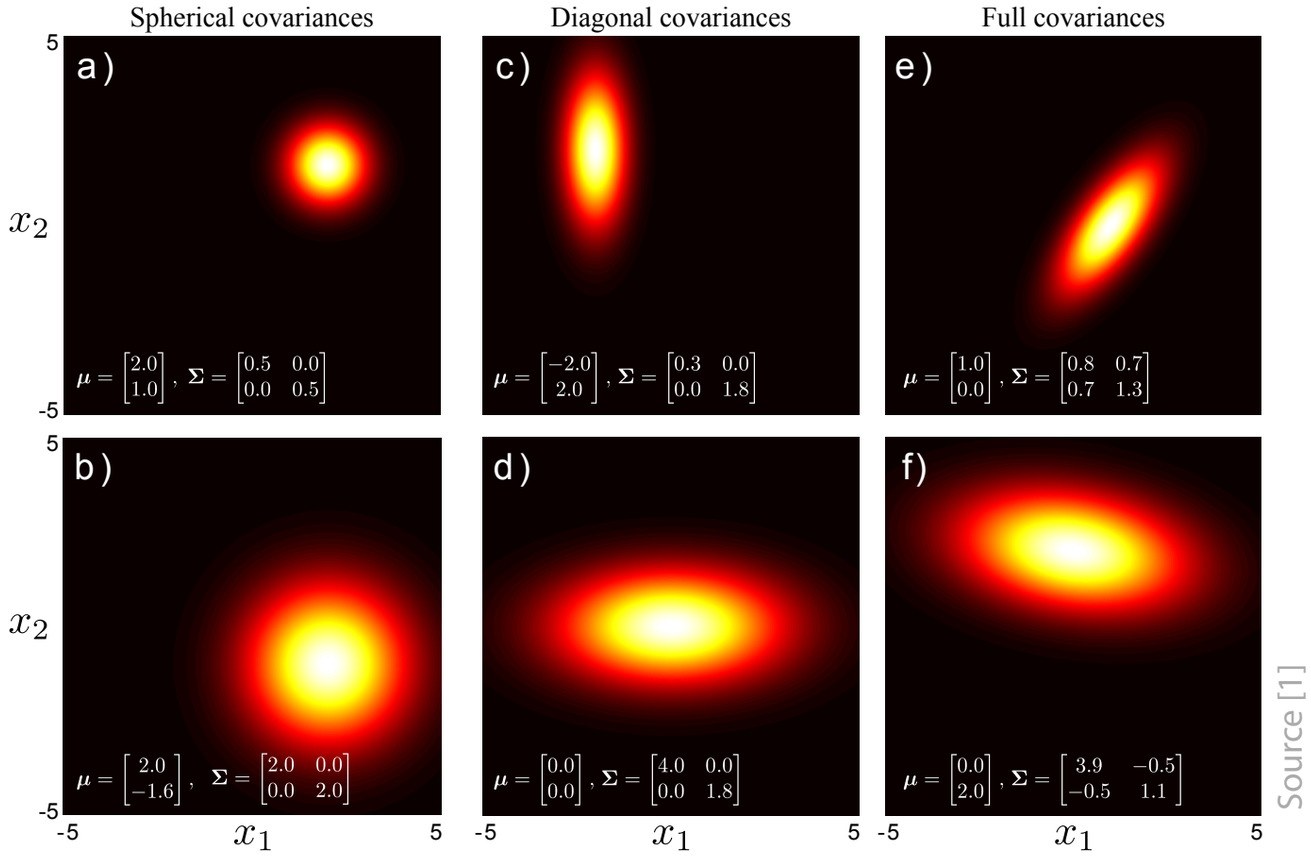
- $\mathrm{Var}[\mathbf{x}] = \Sigma$

Common Probability Distributions

Multivariate Gaussian Distribution

- For $d = 2$, we have the **bivariate** Gaussian distribution
- The covariance matrix Σ (often C) determines the **shape of the distribution** (video)





Source [1]

Parameters

- $\boldsymbol{\mu}$: mean vector
- Σ: covariance matrix

Expectation

- $\mathrm{E}[\mathbf{x}] = \boldsymbol{\mu}$

Variance

- $\mathrm{Var}[\mathbf{x}] = \Sigma$

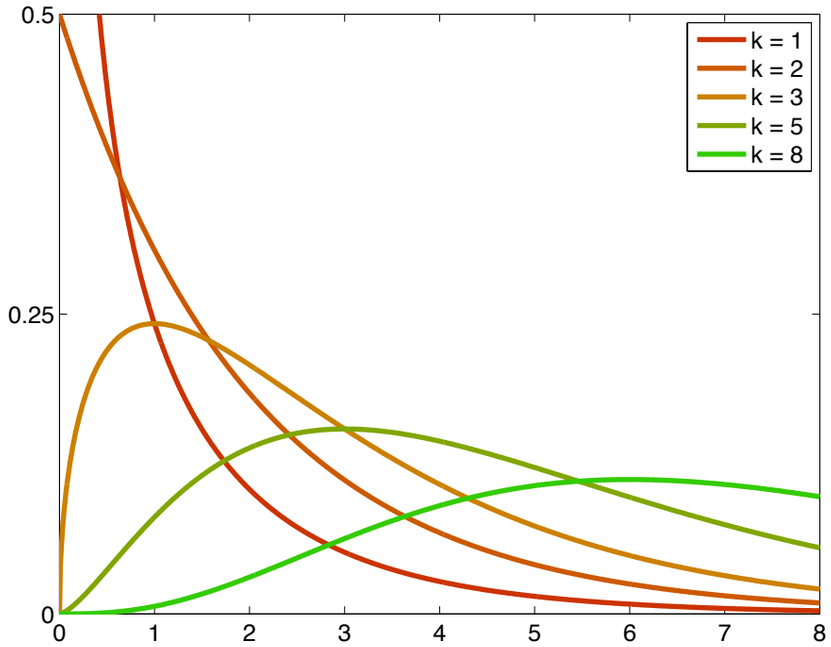
Common Probability Distributions

Chi-squared Distribution

- Consider k independent **standard normally distributed** random variables x_i
- The chi-squared distribution is the **continuous** distribution of a **sum of the squares** of k independent standard normal random variables

$$q = \sum_{i=1}^{k} x_i^2 \qquad q \sim \chi_k^2$$

- **Parameter** k is called the number of "degrees of freedom"
- It is one of the **most widely** used probability distributions in statistical inference, e.g., in hypothesis testing



Parameters

- k : degrees of freedom

Expectation

- $\mathrm{E}[x] = k$

Variance

- $\mathrm{Var}[x] = 2k$

Common Probability Distributions

Chi-squared Distribution

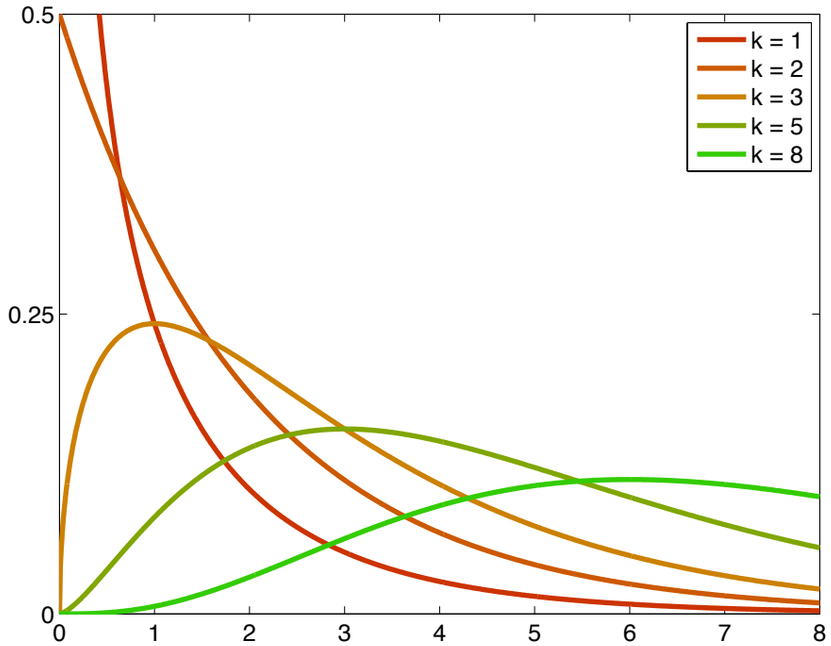
- **Probability density function** (for $x \geq 0$)

$$p(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}} \Gamma\left(\frac{k}{2}\right)}$$

- Notation

$$\chi^2_x(k) = \chi^2_k = p(x)$$

- For **hypothesis testing**, values of the **cumulative distribution function** are taken, typically from tables in statistics text books or online sources



Parameters

- k : degrees of freedom

Expectation

- $\mathrm{E}[x] = k$

Variance

- $\mathrm{Var}[x] = 2k$

Summary

- **Uncertainty** is an **inescapable aspect** of every system in the real world
- Probability theory is a **very powerful framework** to represent, propagate, reduce and reason about uncertainty
- The rules of probability are **remarkably compact and simple**
- The concepts of **marginalization**, **joint** and **conditional probability**, **independence** and **conditional independence** underpin many today algorithms in robotics, machine learning, computer vision and AI
- Two immediate results of the definition of conditional probability are **Bayes' rule** and the **chain rule**
- Together with the **sum rule** (marginalization) they form the foundation of even the most advanced inference and learning methods. Memorize them!
- There are also **alternative approaches** to **uncertainty representation**
  - Fuzzy logic, possibility theory, set theory, belief functions, qualitative uncertainty representations

Sources and Further Readings

The first section, Introduction to Probability, follows to large parts chapter 2 of Prince et al. [1] and the nice figures are taken from his book. The section also contains material from chapters 1 and 2 in Koller and Friedman [2].

Another good compact summary of probability theory can be found in the book by Bischop [3]. A comprehensive treatment of probability theory is, for instance, the book by Papoulis and Pillai [4].

[1] S.J.D. Prince, "Computer vision: models, learning and inference", Cambridge University Press, 2012. See www.computervisionmodels.com

[2] D. Koller, N. Friedman, "Probabilistic graphical models: principles and techniques", MIT Press, 2009. See http://pgm.stanford.edu

[3] C.M. Bischop, "Pattern Recognition and Machine Learning", Springer, 2nd ed., 2007. See http://research.microsoft.com/en-us/um/people/cmbishop/prml

[4] A. Papoulis, S.U. Pillai, "Probability, Random Variables and Stochastic Processes", McGraw-Hill, 4th edition, 2002. See http://www.mhhe.com/engcs/electrical/papoulis