Probability Theory

Lecture 1

Per Sidén

Division of Statistics and Machine Learning
Dept. of Computer and Information Science
Linköping University

Overview Lecture 1

- Course outline
- Introduction and a recap of some background
- Functions of random variables
- Multivariate random variables

Course Outline

- **Lectures**: theory interleaved with illustrative solved examples.
Responsible: Per.

- **Exercises/Seminars**: problem solving sessions + open discussions.
Responsible: Per and You.

- **Exam**: written exam with formula sheet, but no book or notes.
Responsible: You!

- **Course homepage**: https://www.ida.liu.se/~732A40/ (select english)

Course literature

- Gut, A. *An intermediate course in probability*. 2nd ed. Springer-Verlag, New York, 2009. ISBN 978-1-4419-0161-3
- Chapter 1: Multivariate random variables
- Chapter 2: Conditioning
- Chapter 3: Transforms
- Chapter 4: Order statistics
- Chapter 5: The multivariate normal distribution
- Chapter 6: Convergence

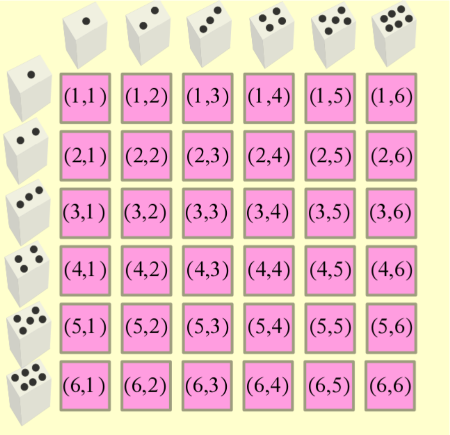
Random variables

- The sample space $\Omega = \{\omega_1, \omega_2, ...\}$ of an experiment is the most basic representation of a problem's randomness (uncertainty).
- More convenient to work with real-valued measurements.
- A **random variable** X is a real-valued function from a sample space: $X = f(\omega)$, where $f : \Omega \to \mathbb{R}$.
- A **multivariate random vector**: $\mathbf{X} = f(\omega)$ such that $f : \Omega \to \mathbb{R}^n$.
- Examples:
  - Roll a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

$$X(\omega) = \begin{cases} 0 & \text{if } \omega = 1, 2 \text{ or } 3 \\ 1 & \text{if } \omega = 4, 5 \text{ or } 6 \end{cases}$$

  - Roll two fair dice. $X(\omega)$=sum of the two dice.

Sample space of two dice example



The distribution of a random variable

- The probabilities of events on the sample space Ω imply a **probability distribution** for a random variable $X(\omega)$ on Ω.
- The probability distribution of X is given by

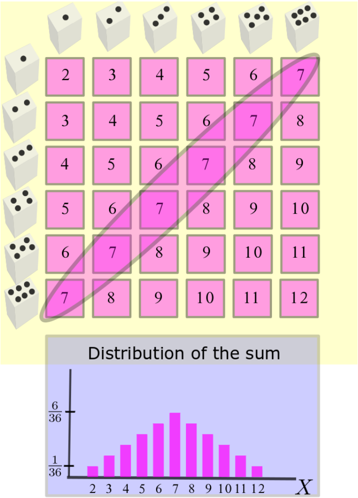
  $$\Pr(X \in C) = \Pr(\{\omega : X(\omega) \in C\}),$$

  where $\{\omega : X(\omega) \in C\}$ is the event (in Ω) consisting of all outcomes ω that gives a value of X in C.
- A random variable is **discrete** if it can take only a finite or a countable number of different values $x_1, x_2, \ldots$.
- **Continuous** random variables can take every value in an interval.

Discrete random variable

- The **probability function** (p.f), is the function

$$p(x) = \Pr(X = x)$$



Uniform, Bernoulli and Poisson

- **Uniform discrete** distribution. $X \in \{a, a+1, ..., b\}$.

$$p(x) = \begin{cases} \frac{1}{b-a+1} & \text{for } x = a, a+1..., b \\ 0 & \text{otherwise} \end{cases}$$

- **Bernoulli distribution**. $X \in \{0, 1\}$. $\Pr(X = 0) = 1 - p$ and $\Pr(X = 1) = p$.

- **Poisson distribution**: $X \in \{0, 1, 2, ...\}$

$$p(x) = \frac{\exp(-\lambda) \cdot \lambda^x}{x!} \quad \text{for } x = 0, 1, 2, ...$$

The Binomial distribution

- **Binomial distribution**. Sum of n independent Bernoulli variables $X_1, X_2, ..., X_n$ with the same success probability p.

$$X = X_1 + X_2 + ... + X_n$$

$$X \sim Bin(n, p)$$

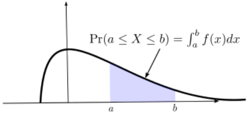
- Probability function for a $\mathrm{Bin}(n, p)$ variable:

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \text{ for } x = 0, 1, ..., n.$$

- The binomial coefficient $\binom{n}{x}$ is the number of binary sequences of length n that sum exactly to x.

Probability density functions

- Continuous random variables can assume **every** value in an interval.
- **Probability density function (pdf)** $f(x)$
  - $\Pr(a \leq X \leq b) = \int_a^b f(x)dx$



  - $f(x) \geq 0$ for all x
  - $\int_{-\infty}^{\infty} f(x)dx = 1$
- A pdf is like a histogram with tiny bin widths. Integral replaces sums.
- Continuous distributions assign probability zero to individual values, but

$$\Pr\left(a - \frac{\epsilon}{2} \leq X \leq a + \frac{\epsilon}{2}\right) \approx \epsilon \cdot f(a).$$

Densities - Some examples

- The **uniform** distribution

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise.} \end{cases}$$

- The **triangle** or linear pdf

$$f(x) = \begin{cases} \frac{2}{a^2}x & \text{for } 0 < x < a \\ 0 & \text{otherwise} \end{cases}$$

- The **normal**, or **Gaussian**, distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Expected values, moments

- The **expected value** of X is

$$E(X) = \begin{cases} \sum_{k=i}^{\infty} x_k \cdot p(x_k) & , X \textit{ discrete} \\ \int_{-\infty}^{\infty} x \cdot f(x) & , X \textit{ continuous} \end{cases}$$

- Example: $E(X)$ when $X \sim Uniform(a, b)$
- The nth **moment** is defined as $E(X^n)$
- The **variance** of X is $Var(X) = E(X - EX)^2 = E(X^2) - (EX)^2$

The Cumulative distribution function

- The (cumulative) **distribution function (cdf)** $F(\cdot)$ of a random variable X is the function

$$F(x) = \Pr(X \leq x) \text{ for } -\infty \leq x \leq \infty$$

- Same definition for discrete and continuous variables.
- The cdf is **non-decreasing**

$$\text{If } x_1 \leq x_2 \text{ then } F(x_1) \leq F(x_2)$$

- Limits at $\pm\infty$: $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.
- For continuous variables: **relation between pdf and cdf**

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

and conversely

$$\frac{dF(x)}{dx} = f(x)$$

Functions of random variables

- Quite common situation: You know the distribution of X, but need the distribution of $Y = g(X)$, where $g(\cdot)$ is some function.
- Example 1: $Y = a + b \cdot X$, where a and b are constants.
- Example 2: $Y = 1/X$
- Example 3: $Y = \ln(X)$.
- Example 4: $Y = \log \frac{X}{1-X}$
- $Y = g(X)$, where X is discrete.
- $p_X(x)$ is p.f. for X. $p_Y(y)$ is p.f. for Y:

$$p_{\mathrm{Y}}(\mathrm{y}) = \Pr(Y = y) = \Pr[g(X) = y] = \sum_{x:g(x)=y} p_X(x)$$

Function of a continuous random variable

- Suppose that X is continuous with support (a, b). Then

$$F_Y(y) = \Pr(Y \leq y) = \Pr\left[g(X) \leq y\right] = \int_{x:g(x)\leq y} f_X(x)dx$$

- Let $g(X)$ be monotonically *increasing* with inverse $X = h(Y)$. Then

$$F_Y(y) = Pr(Y \leq y) = Pr(g(X) \leq y) = Pr(X \leq h(y)) = F_X(h(y))$$

and

$$f_Y(y) = f_X(h(y)) \cdot \frac{\partial h(y)}{\partial y}$$

- For general monotonic transformation $Y = g(X)$ we have

$$f_Y(y) = f_X\left[h(y)\right] \left|\frac{\partial h(y)}{y}\right| \text{ for } \alpha\text{<y<}\beta$$

where (α, β) is the mapped interval from (a, b).

Examples: functions of a random variable

- Example 1. $Y = a{\cdot}X + b$.

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$

- Example 2: **log-normal**. $X \sim N(\mu, \sigma^2)$. $Y = g(X) = \exp(X)$. $X = h(Y) = \ln Y$.

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(\ln y - \mu\right)^2\right) \cdot \frac{1}{y} \text{ for } y > 0.$$

- Example 3. $X \sim LogN(\mu, \sigma^2)$. $Y = a \cdot X$, where $a > 0$. $X = h(Y) = Y/a$.

$$\begin{aligned} f_Y(y) &= \frac{1}{y/a} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(\ln \frac{y}{a} - \mu\right)^2\right) \frac{1}{a}. \\ &= \frac{1}{y} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(\ln y - \mu - \ln a\right)^2\right) \end{aligned}$$

which means that $Y \sim LogN(\mu + \ln a, \sigma^2)$.

Examples: functions of a random variable

- Example 4. $X \sim LogN(\mu, \sigma^2)$. $Y = X^a$, where $a \neq 0$.
  $X = h(Y) = Y^{1/a}$.

$$
\begin{aligned}
f_Y(y) &= \frac{1}{y^{1/a}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(\ln y^{1/a} - \mu\right)^2\right) \frac{1}{a} y^{1/a-1}. \\
&= \frac{1}{y} \frac{1}{\sqrt{2\pi}a\sigma} \exp\left(-\frac{1}{2a^2\sigma^2}(\ln y - a\mu)^2\right)
\end{aligned}
$$

which means that $Y \sim LogN(a\mu, a^2\sigma^2)$.

Bivariate distributions

- The **joint** (or **bivariate**) **distribution** of the two random variables X and Y is the collection of all probabilities of the form

$$\Pr\left[(X, Y) \in C\right]$$

- Example 1:
  - $X =$# of visits to doctor.
  - $Y =$#visits to emergency.
  - C may be $\{(x, y) : x = 0 \text{ and } y \geq 1\}$.
- Example 2:
  - $X =$monthly percentual return to SP500 index
  - $Y =$monthly return to Stockholm index.
  - C may be $\{(x, y) : x < -10 \text{ and } y < -10\}$.
- **Discrete** random variables: **joint probability function** (joint p.f.)

$$f_{X,Y}(x, y) = \Pr(X = x, Y = y)$$

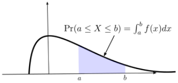
such that $\Pr\left[(X, Y) \in C\right] = \sum_{(x,y)\in C} f_{X,Y}(x, y)$ and $\sum_{All\ (x,y)} f_{X,Y}(x, y) = 1$.

Continuous joint distributions

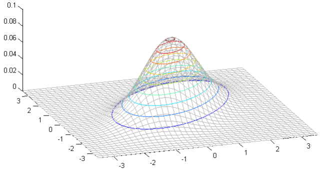
- **Continuous joint distribution** (joint p.d.f.)

$$\Pr[(X, Y) \in C] = \iint_C f_{X,Y}(x, y)dxdy,$$

  where $f_{X,Y}(x, y) \geq 0$ is the **joint density**.
- Univariate distributions: probability is area under density.



- Bivariate distributions: probability is volume under density.

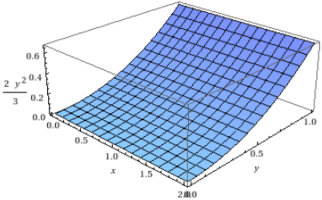


- Be careful about the regions of integration. Example: $C = \{(x, y) : x^2 \leq y \leq 1\}$

Example

- Example

$$f_{X,Y}(x, y) = \frac{3}{2}y^2 \text{ for } 0 \leq x \leq 2 \text{ and } 0 \leq y \leq 1.$$



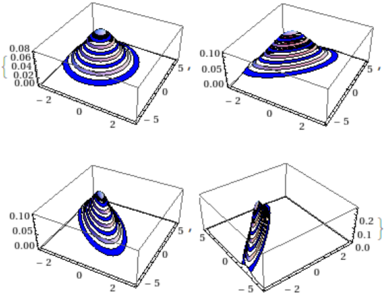
Bivariate normal distribution

- The most famous of them all: the **bivariate normal distribution**, with pdf

$$f_{X,Y}(x,y) = \frac{1}{2\pi(1-\rho^2)^{1/2}\sigma_x\sigma_y} \times$$

$$\exp\left(-\frac{1}{2\left(1-\rho^2\right)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right)$$

- Five parameters: $\mu_x, \mu_y, \sigma_x, \sigma_y$ and ρ.



Bivariate c.d.f.

- Joint cumulative distribution function (joint c.d.f.):

$$F_{X,Y}(x,y) = \Pr(X \leq x, Y \leq y)$$

- Calculating probabilities of rectangles $\Pr(a < X \leq b \text{ and } c < Y \leq d)$:

$$F_{X,Y}(b,d) - F_{X,Y}(a,d) - F_{X,Y}(b,c) + F_{X,Y}(a,c)$$

- Properties of the joint c.d.f.
    - Marginal of X: $F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y)$
    - $F_{X,Y}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(r,s)drds$
    - $f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$

Marginal distributions

- Marginal p.f. of a bivariate distribution is

$$f_X(x) = \sum_{All\ y} f_{X,Y}(x,y) \text{ [Discrete case]}$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \text{ [Continuous case]}$$

- A marginal distribution for X tells you about the probability of different values of X, averaged over all possible values of Y.

Independent variables

- Two random variables are **independent** if

$$\mathrm{Pr}(\mathrm{X} \in \mathrm{A} \text{ and } \mathrm{Y} \in \mathrm{B}) = \mathrm{Pr}(\mathrm{X} \in \mathrm{A}) \cdot \mathrm{Pr}(\mathrm{Y} \in \mathrm{B})$$

  for all sets of real numbers A and B (such that $\{X \in A\}$ and $\{Y \in B\}$ are events).
- Two variables are **independent** if and only if the joint density can be factorized as

$$f_{X,Y}(x, y) = h_1(x) \cdot h_2(y)$$

- Note: this factorization must hold for **all** values of x and y. Watch out for non-rectangular support!
- X and Y are independent if learning something about X (e.g. $X > 2$) has no effect on the probabilities for different values of Y.

Multivariate distributions

- Obvious extension to more than two random variables, $X_1, X_2, ..., X_n$.
- Joint p.d.f.

$$f(x_1, x_2, ..., x_n)$$

- Marginal distribution of x_1

$$f_1(x_1) = \int_{x_2} \cdots \int_{x_n} f(x_1, x_2, ..., x_n) dx_2 \cdots dx_n$$

- Marginal distribution of x_1 and x_2

$$f_{12}(x_1, x_2) = \int_{x_3} \cdots \int_{x_n} f(x_1, x_2, ..., x_n) dx_3 \cdots dx_n$$

and so on.

Functions of random vectors

- Let $\mathbf{X}$ be an n-dimensional continuous random variable
- Let $\mathbf{X}$ have density $f_{\mathbf{X}}(\mathbf{x})$ on support $S \subset \mathbb{R}^n$.
- Let $Y = g(X)$, where $g : S \to T \subset \mathbb{R}^n$ is a bijection (1:1 and onto).
- Assume g and g^{-1} are continuously differentiable with Jacobian

$$\mathbf{J} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix}$$

Theorem

("The transformation theorem") The density of Y is

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}\left[h_1(\mathbf{y}), h_2(\mathbf{y}), ..., h_n(\mathbf{y})\right] \cdot |\mathbf{J}|$$

where $h = (h_1, h_2, ..., h_n)$ is the unique inverse of $g = (g_1, g_2, ..., g_n)$.