STAT W4413: Nonparametric Statistics

Paweł Polak

January 19, 2016

STAT W4413: Nonparametric Statistics - Lecture 1

Course Description

Meeting Time & Location

TR 7:40PM-8:55PM, 614 Schermerhorn Hall [SCH]

- Instructor: Paweł Polak,
  - Office Building: 1255 Amsterdam Ave, Room 928 (SSW, 9th floor)
  - Office Hours: 4:00 PM - 6:00 PM, Friday
    (please send me an email if you plan to come)
  - E-mail: pp2501@columbia.edu
    (please start the title of the email with [W4413])
- Teaching Assistant: Xiaoou Li
  - Office Hours: 12:30 PM - 2:30 PM, Thursday,
    at 10'th floor lounge of the SSW building.
  - E-mail: xiaoou@stat.columbia.edu
    (please start the title of the email with [W4413])

Course Description

Materials:

- Lecture notes & homework materials.

- Additional Readings:
  1. J. J. Higgins, "Introduction to modern nonparametric statistics".
  2. T. Hastie, R. Tibshirani, J. Friedman, "The Elements of Statistical Learning".
  3. L. Wasserman, "All of nonparametric statistics".
  4. B. Efron, R. Tibshirani, "Introduction to Bootstrap" E. Lehman, "Nonparametrics: statistical methods based on ranks".
  5. J. Gibbons, S. Chakraborti, "Nonparametric statistical inference".
  6. A. Tsybakov, "Introduction to nonparametric estimation".

Course Description

This course provides an introduction to nonparametric estimation and testing.

- Topics on estimation:
  - curve fitting, locally linear polynomial, bias/variance trade-off, splines, cross validation.
- Topics on testing:
  - goodness of fit tests including chi-square test, Kolmogorov-Smirnov test, Cramer-von Mises test, Anderson and Darling test, one and two sample nonparametric tests including sign test, rank test, Wilcoxon's test, Kruskal-Wallis test, and Friedman test. Measures of association such as Spearman's rank correlation and Kendall's Tau.
- If time permits, we will cover topics on parametric resampling, boosting, and bootstrap methods.

Course Outline

- Review of:
  - probability theory,
  - multivariate Gaussian distribution, and
  - linear algebra
- Estimation/testing
- Stochastic convergence
- Kernel/locally linear regression
- Bias/Var+Cross validation
- Splines/smoothing spline

Course Outline

- Testing/goodness of fit
- Goodness of fit (Kolmogorov-Smirnov / Cramer-von Mises tests)
- One/two sample tests
- k-sample tests
- Paired sample tests
- Block design
- Measures of association
- Parametric resampling, boosting and bootstrap methods.

Requirements

- Probability: basic knowledge of probability at the level of W4105
- Statistics: basic knowledge of statistics at the level of W4107
- Programming: basic knowledge of computer science, at a level sufficient to write a program in R, Matlab, or any other programming language.

Software

- MATLAB: The Language of Technical Computing.
- R: The R Project for Statistical Computing.
- All the examples in the lectures will be made in MATLAB.
- For homework you can use MATLAB or R depending on your preference.

Homeworks

- Homeworks (30%)
  - There will be 7-10 HW assignments.
  - Collaboration is allowed in solving the problems, but each student should hand in her or his own independently written solutions.
  - DUE by Thursday 12:00 noon.
  - Homework must be submitted into the Statistics Homework Boxes room 904 on the 9th floor of SSW building.
  - HW cannot be submitted to your TA by e-mail.
  - Please do not contact the TA or the grader directly concerning your grades.
  - Please write W4413 in the subject heading of all e-mail correspondence with instructor/TA. (This is in general effective in weeding out spam email.)
  - No late homework accepted.
  - Lowest score will be dropped.

Grading: 30/25/45

- 30% Homeworks.
- 25% midterm exam (in class):
  - Tuesday, March 8, 7:40PM-8:55PM, 614 Schermerhorn Hall [SCH]
- 45% final (in class):
  - TBA, (Consult Student Services Online for Final Exam Schedule).
- Exams are closed-book, closed-notes.
- An Important Note: no make-up exams will be given.
- The final letter grade depends on your performance in homeworks, midterm, and final exam.

Some More Information

- Along the semester I will provide a set of lecture notes, and homeworks which will cover the whole material from the lecture and help you prepare for the exams.
- I encourage you to ask questions to make sure we are all on the same page during the lectures. Please ask questions!
- I try my best to be available, so that you can ask all your questions.
- I try my best to help you improve your background (linear algebra, inference, probability theory).
- You will ask question, when you do not understand something.
- If you feel you do not have a good background on a certain topic, stop by and let me know. I will for sure help.

Review of Probability Theory

Probability Theory

- The object of probability theory is to describe and investigate mathematical models of random phenomena.
- Statistics is closely related to probability theory. Statistics is concerned with creating principles, methods, and criteria in order to treat data pertaining to such (random) phenomena or data from experiments and other observations of the real world, by using the theories and knowledge available from the theory of probability.
- Through this class, we will be relying on concepts from probability theory and statistics. These notes attempt to cover the basics of probability theory without going into measure theory.

Probability Space

In order to define a probability on a set we need the probability space which consists of three basic elements $(\Omega, \mathcal{F}, P)$, where

- Ω is the **Sample space**: The set of all the outcomes of a random experiment. Here, each outcome $\omega \in \Omega$ can be thought of as a complete description of the state of the real world at the end of the experiment.
- $\mathcal{F}$ is the **Set of events** (or **event space**): A set whose elements $A \in \mathcal{F}$ (called **events**) are subsets of Ω (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment).[1]
- P is the **Probability measure**: A function $P : \mathcal{F} \to \mathbb{R}$ that satisfies the following properties:
    1. For an $A \in \mathcal{F}$, there exists a number $P(A) \geq 0$, the probability of A
    2. $P(\Omega) = 1$
    3. If $A_1, A_2, \ldots$ are disjoint events (i.e., $A_i \cap A_j = \emptyset$ whenever $i \neq j$), then

$$P(\cup_i A_i) = \sum_i P(A_i).$$

Axiom 3 is called countable additivity (in contrast to finite additivity).

These three properties are called the **Axioms of Probability** they are also known as Kolmogorov axioms.

[1] $\mathcal{F}$ has to satisfy three properties: (1) $\emptyset \in \mathcal{F}$; (2) $A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F}$; and (3) $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow \cup_i A_i \in \mathcal{F}$

Example

- Consider the event of tossing a six-sided die.
- The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- We can define different event spaces on this sample space.
- For example, the simplest event space is the trivial event space $\mathcal{F} = \{\emptyset, \Omega\}$.
- Another event space is the set of all subsets of Ω.
- For the first event space, the unique probability measure satisfying the requirements above is given by $P(\emptyset) = 0$, $P(\Omega) = 1$.
- For the second event space, one valid probability measure is to assign the probability of each set in the event space to be $\frac{i}{6}$ where i is the number of elements of that set; for example, $P(\{1, 2, 3, 4\}) = \frac{4}{6}$ and $P(\{1, 2, 3\}) = \frac{3}{6}$.

Properties of probability measure P

- If $A \subseteq B \Longrightarrow P(A) \leq P(B)$.
- $P(A \cap B) \leq \min(P(A), P(B))$.
- (Union Bound) $P(A \cup B) \leq P(A) + P(B)$.
- $P(\Omega \setminus A) = 1 - P(A)$.
- (Law of Total Probability) If $A_1, \ldots, A_k$ are a set of disjoint events such that $\cup_{i=1}^{k} A_i = \Omega$, then $\sum_{i=1}^{k} P(A_k) = 1$.

Conditional probability and independence

Let B be an event with non-zero probability. The conditional probability of any event A given B is defined as,

$$P(A \mid B) \triangleq \frac{P(A \cap B)}{P(B)}$$

In other words, $P(A \mid B)$ is the probability measure of the event A after observing the occurrence of event B. Two events are called independent if and only if $P(A \cap B) = P(A)P(B)$ (or equivalently, $P(A \mid B) = P(A)$). Therefore, independence is equivalent to saying that observing B does not have any effect on the probability of A.

Random Variables

- Consider an experiment in which we flip 10 coins, and we want to know the number of coins that come up heads.

- Here, the elements of the sample space Ω are 10-length sequences of heads and tails.

- For example, we might have $w_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$.

- However, in practice, we usually do not care about the probability of obtaining any particular sequence of heads and tails. Instead we usually care about real-valued functions of outcomes, such as the number of heads that appear among our 10 tosses, or the length of the longest run of tails. These functions, under some technical conditions, are known as **random variables**.

Random Variables

- More formally, a (real valued) random variable X is a measurable function* $X : \Omega \longrightarrow \mathbb{R}$.
- Typically, we will denote random variables using upper case letters $X(\omega)$ or more simply X (where the dependence on the random outcome ω is implied).
- We will denote the value that a random variable may take on using lower
case letters x.

(*)Technically speaking, not every function is acceptable as a random variable. From a measure-theoretic perspective, random variables must be Borel-measurable functions.

Example

- In our experiment in the previous slide, suppose that $X(\omega)$ is the number of heads which occur in the sequence of tosses ω.
- Given that only 10 coins are tossed, $X(\omega)$ can take only a finite number of values, so it is known as a **discrete random variable**.
- Here, the probability of the set associated with a random variable X taking on some specific value k is

$$P(X = k) := P(\{\omega : X(\omega) = k\}).$$

Example

- Suppose that $X(\omega)$ is a random variable indicating the amount of time it takes for a radioactive particle to decay.
- In this case, $X(\omega)$ takes on a infinite number of possible values, so it is called a **continuous random variable**.
- We denote the probability that X takes on a value between two real constants a and b (where $a < b$) as

$$P(a \leq X \leq b) := P(\{\omega : a \leq X(\omega) \leq b\}).$$

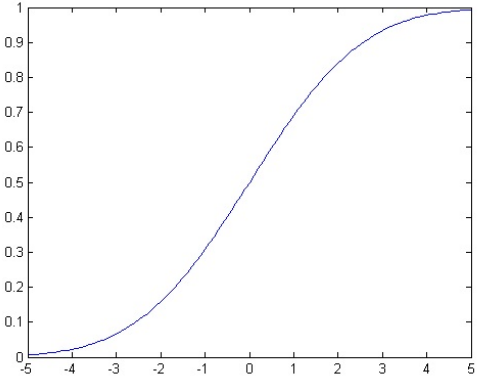
In order to specify the probability measures used when dealing with random variables, it is often convenient to specify alternative functions (CDFs, PDFs, and PMFs) from which the probability measure governing an experiment immediately follows.

Cumulative distribution functions CDF

A **cumulative distribution function (CDF)** is a function $F_X : \mathbb{R} \to [0,1]$ which specifies a probability measure as,

$$F_X(x) \triangleq P(X \leq x). \tag{1}$$

By using this function one can calculate the probability of any event in $\mathcal{F}$. Figure 1 shows a sample CDF function.



Properties of CDF

- $0 \leq F_X(x) \leq 1$.
- $\lim_{x \to -\infty} F_X(x) = 0$.
- $\lim_{x \to \infty} F_X(x) = 1$.
- $x \leq y \Longrightarrow F_X(x) \leq F_X(y)$.

Probability mass functions PMF

When a random variable X takes on a finite set of possible values (i.e., X is a discrete random variable), a simpler way to represent the probability measure associated with a random variable is to directly specify the probability of each value that the random variable can assume.
In particular, a *probability mass function (PMF)* is a function $p_X : \Omega \to \mathbb{R}$ such that

$$p_X(x) \triangleq P(X = x).$$

In the case of discrete random variable, we use the notation $Val(X)$ for the set of possible values that the random variable X may assume. For example, if $X(\omega)$ is a random variable indicating the number of heads out of ten tosses of coin, then $Val(X) = \{0, 1, 2, \ldots, 10\}$.

Properties of PMS

- $0 \leq p_X(x) \leq 1$.
- $\sum_{x \in Val(X)} p_X(x) = 1$.
- $\sum_{x \in A} p_X(x) = P(X \in A)$.

Probability density functions PDF

For some continuous random variables, the cumulative distribution function $F_X(x)$ is differentiable everywhere. In these cases, we define the **Probability Density Function** or **PDF** as the derivative of the CDF, i.e.,

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}. \tag{2}$$

Note here, that the PDF for a continuous random variable may not always exist (i.e., if $F_X(x)$ is not differentiable everywhere).
According to the properties of differentiation, for very small $\triangle x$,

$$P(x \leq X \leq x + \triangle x) \approx f_X(x) \triangle x. \tag{3}$$

Both CDFs and PDFs (when they exist!) can be used for calculating the probabilities of different events. But it should be emphasized that the value of PDF at any given point x is not the probability of that event, i.e., $f_X(x) \neq P(X = x)$. For example, $f_X(x)$ can take on values larger than one (but the integral of $f_X(x)$ over any subset of $\mathbb{R}$ will be at most one).

Properties of PDF

- $f_X(x) \geq 0$.
- $\int_{-\infty}^{\infty} f_X(x) = 1$.
- $\int_{x \in A} f_X(x) dx = P(X \in A)$.

Expectation

- Suppose that X is a discrete random variable with PMF $p_X(x)$ and $g : \mathbb{R} \to \mathbb{R}$ is an arbitrary function. In this case, $g(X)$ can be considered a random variable, and we define the **expectation** or **expected value** of $g(X)$ as

$$E[g(X)] \triangleq \sum_{x \in Val(X)} g(x) p_X(x).$$

- If X is a continuous random variable with PDF $f_X(x)$, then the expected value of $g(X)$ is defined as,

$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Expectation

- Intuitively, the expectation of $g(X)$ can be thought of as a "weighted average" of the values that $g(x)$ can taken on for different values of x, where the weights are given by $p_X(x)$ or $f_X(x)$.
- As a special case of the above, note that the expectation, $E[X]$ of a random variable itself is found by letting $g(x) = x$; this is also known as the **mean** of the random variable X.

Properties of Expectation

- $E[a] = a$ for any constant $a \in \mathbb{R}$.
- $E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbb{R}$.
- (Linearity of Expectation) $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$.
- For a discrete random variable X, $E[1\{X = k\}] = P(X = k)$.

Variance

The variance of a random variable X is a measure of how concentrated the distribution of a random variable X is around its mean. Formally, the variance of a random variable X is defined as

$$Var[X] \triangleq E[(X - E(X))^2]$$

Using the properties in the previous section, we can derive an alternate expression for the variance:

$$\begin{aligned} E[(X - E[X])^2] &= E[X^2 - 2E[X]X + E[X]^2] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2; \end{aligned}$$

where the second equality follows from linearity of expectations and the fact that $E[X]$ is actually a constant with respect to the outer expectation.

Properties:

- $Var[a] = 0$ for any constant $a \in \mathbb{R}$.
- $Var[af(X)] = a^2 Var[f(X)]$ for any constant $a \in \mathbb{R}$.

Example

Calculate the mean and the variance of the uniform random variable X with PDF $f_X(x) = 1$, $\forall x \in [0, 1]$, 0 elsewhere.

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^1 x dx = \frac{1}{2}.$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_0^1 x^2 dx = \frac{1}{3}.$$

$$Var[X] = E[X^2] - E[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Example: Derivation for Normal Distribution.

Suppose $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, then

$$\begin{aligned}
\mathbb{E}(X) &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \mathrm{d}x \quad (\text{setting } z = x - \mu) \\
&= \underbrace{\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z}_{\text{expected value of } \mathrm{N}\left(0,\sigma^2\right)} + \mu \underbrace{\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2\sigma^2}} \mathrm{d}z}_{1} \\
&= 0 + \mu = \mu.
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(X) = \mathbb{E}(X-\mu)^2 &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \mathrm{d}x \\
&\quad (\text{setting } z = (x-\mu)/\sigma) \\
&= \sigma^2 \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} \mathrm{d}z}_{\text{variance of } \mathrm{N}(0,1)} \\
&= \sigma^2.
\end{aligned}$$

Example

Suppose that $g(x) = 1\{x \in A\}$ for some subset $A \subseteq \Omega$.

What is $E[g(X)]$?

Discrete case:

$$E[g(X)] = \sum_{x \in Val(X)} 1\{x \in A\} P_X(x) = \sum_{x \in A} P_X(x) = P(x \in A).$$

Continuous case:

$$E[g(X)] = \int_{-\infty}^{\infty} 1\{x \in A\} f_X(x) dx = \int_{x \in A} f_X(x) dx = P(x \in A).$$

Some common discrete random variables

- $X \sim \mathit{Bernoulli}(p)$ (where $0 \leq p \leq 1$): one if a coin with heads probability p comes up heads, zero otherwise

$$p(x) = \begin{cases} p & \text{if } X = 1 \\ 1 - p & \text{if } X = 0 \end{cases}$$

- $X \sim \mathit{Bernoulli}(n, p)$ (where $0 \leq p \leq 1$): the number of heads in n independent flips of a coin with heads probability p.

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- $X \sim \mathit{Geometric}(p)$ (where $p > 0$): the number of flips of a coin with heads probability p until the first heads.

$$p(x) = p(1-p)^{x-1}$$

- $X \sim \mathit{Poisson}(\lambda)$ (where $\lambda > 0$): a probability distribution over the nonnegative integers used for modeling the frequency of rare events.

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

Some common continuous random variables

- $X \sim \mathit{Uniform}(a, b)$ (where $a < b$): equal probability density to every value between a and b on the real line.

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

- $X \sim \mathit{Exponential}(\lambda)$ (where $\lambda > 0$): decaying probability density over the nonnegative reals.

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if otherwise} \end{cases}$$

- $X \sim \mathit{Normal}(\mu, \sigma^2)$: also known as the Gaussian distribution

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

Shape of PDFs and CDFs

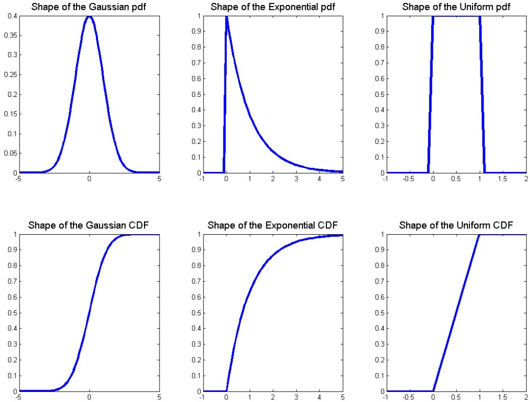


Figure : PDF and CDF of a couple of random variables.

Properties of distributions

The following table is the summary of some of the properties of these distributions.

Distribution	PDF or PMF	Mean	Variance
Bernoulli(p)	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0 \end{cases}$	p	$p(1-p)$
Binomial(n,p)	$\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
Geometric(p)	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$e^{-\lambda}\lambda^x / x!$ for $k = 1, 2, \ldots$	λ	λ
Uniform(a,b)	$\frac{1}{b-a} \forall x \in (a,b)$	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$
Gaussian(μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
Exponential(λ)	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Two random variables

- In many situations there may be more than one quantity that we are interested in knowing during a random experiment.
- For instance, in an experiment where we flip a coin ten times, we may care about both $X(\omega) =$ the number of heads that come up as well as $Y(\omega) =$ the length of the longest run of consecutive heads. In this section, we consider the setting of two random variables.

Joint and marginal distributions

- Suppose that we have two random variables X and Y.

- ne way to work with these two random variables is to consider each of them separately.

- If we do that we will only need $F_X(x)$ and $F_Y(y)$. But if we want to know about the values that X and Y assume simultaneously during outcomes of a random experiment, we require a more complicated structure known as the **joint cumulative distribution function** of X and Y , defined by

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

- It can be shown that by knowing the joint cumulative distribution function, the probability of any event involving X and Y can be calculated.

Joint and marginal distributions

The joint CDF $F_{XY}(x, y)$ and the joint distribution functions $F_X(x)$ and $F_Y(y)$ of each variable separately are related by

$$F_X(x) = \lim_{y \to \infty} F_{XY}(x, y)dy$$

$$F_Y(y) = \lim_{x \to \infty} F_{XY}(x, y)dx$$

Here, we call $F_X(x)$ and $F_Y(y)$ the **marginal cumulative distribution functions** of $F_{XY}(x, y)$.

Properties

- $0 \leq F_{XY}(x, y) \leq 1$.
- $\lim_{x,y \to \infty} F_{XY}(x, y) = 1$.
- $\lim_{x,y \to -\infty} F_{XY}(x, y) = 0$.
- $F_X(x) = \lim_{y \to \infty} F_{XY}(x, y)$.

Joint and marginal probability mass functions

- If X and Y are discrete random variables, then the **joint probability mass function** $p_{XY} : \mathbb{R} \times \mathbb{R} \to [0, 1]$ is defined by

$$p_{XY}(x, y) = P(X = x, Y = y).$$

- Here, $0 \leq P_{XY}(x, y) \leq 1$ for all x, y, and $\sum_{x \in Val(X)} \sum_{y \in Val(Y)} P_{XY}(x, y) = 1$.

- How does the joint PMF over two variables relate to the probability mass function for each variable separately? It turns out that

$$p_X(x) = \sum_y p_{XY}(x, y).$$

  and similarly for $p_Y(y)$.

- In this case, we refer to $p_X(x)$ as the **marginal probability mass function** of X. In statistics, the process of forming the marginal distribution with respect to one variable by summing out the other variable is often known as "marginalization."

Joint and marginal probability density functions

Let X and Y be two continuous random variables with joint distribution function F_{XY}. In the case that $F_{XY}(x,y)$ is everywhere differentiable in both x and y, then we can define the **joint probability density function**,

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}.$$

Like in the single-dimensional case, $f_{XY}(x,y) \neq P(X = x, Y = y)$, but rather

$$\iint_{x \in A} f_{XY}(x,y) dx dy = P((X,Y) \in A).$$

Note that the values of the probability density function $f_{XY}(x,y)$ are always nonnegative, but they may be greater than 1. Nonetheless, it must be the case that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y) = 1$.
Analagous to the discrete case, we define

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y) dy,$$

as the **marginal probability density function** (or **marginal density**) of X, and similarly for $f_Y(y)$.

Conditional distributions: discrete case

- Conditional distributions seek to answer the question, what is the probability distribution over Y, when we know that X must take on a certain value x?
- In the discrete case, the conditional probability mass function of X given Y is simply

$$p_{Y|X}(y \mid x) = \frac{p_{XY}(x, y)}{p_X(x)},$$

assuming that $p_X(x) \neq 0$.

Conditional distributions: continuous case

- In the continuous case, the situation is technically a little more complicated because the probability that a continuous random variable X takes on a specific value x is equal to zero. To get around this, a more reasonable way to calculate the conditional CDF is,

$$F_{Y|X}(y,x) = \lim_{\triangle x \to 0} P(Y \leq y \mid x \leq X \leq x + \triangle x).$$

- It can be easily seen that if $F(x,y)$ is differentiable in both x, y then,

$$F_{Y|X}(y,x) = \int_{-\infty}^{y} \frac{f_{X,Y}(x,\alpha)}{f_X(x)} d\alpha$$

and therefore we define the conditional PDF of Y given $X = x$ in the following way,

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}.$$

provided $f_X(x) \neq 0$.

Bayes's rule

A useful formula that often arises when trying to derive expression for the conditional probability of one variable given another, is **Bayes's rule**.
In the case of discrete random variables X and Y,

$$P_{Y|X}(y \mid x) = \frac{P_{X,Y}(x,y)}{P_X(x)} = \frac{P_{X|Y}(x \mid y)P_Y(y)}{\sum_{y' \in Val(Y)} P_{X|Y}(x \mid y')P_Y(y')}.$$

If the random variables X and Y are continuous,

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{f_{X|Y}(x \mid y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x \mid y')f_Y(y')dy'}.$$

Independence

Two random variables X and Y are **independent** if $F_{XY}(x, y) = F_X(x)F_Y(y)$ for all values of x and y. Equivalently,

- For discrete random variables, $p_{XY}(x, y) = p_X(x)p_Y(y)$ for all $x \in Val(X)$, $y \in Val(Y)$.
- For discrete random variables, $p_{Y|X}(y \mid x) = p_Y(y)$ whenever $p_X(x) \neq 0$ for all $y \in Val(Y)$.
- For continuous random variables, $f_{XY}(x, y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$.
- For continuous random variables, $f_{Y|X}(y \mid x) = f_Y(y)$ whenever $f_X(x) \neq 0$ for all $y \in \mathbb{R}$.

Independence

Informally, two random variables X and Y are independent if "knowing" the value of one variable will never have any effect on the conditional probability distribution of the other variable, that is, you know all the information about the pair (X, Y) by just knowing $f(x)$ and $f(y)$. The following lemma formalizes this observation:

Lemma

If X and Y are independent then for any subsets $A, B \subseteq \mathbb{R}$, we have,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

By using the above lemma one can prove that if X is independent of Y then any function of X is independent of any function of Y.

Expectation and covariance

Suppose that we have two discrete random variables X, Y and $g : \mathbf{R}^2 \to \mathbf{R}$ is a function of these two random variables. Then the expected value of g is defined in the following way,

$$E[g(X,Y)] \triangleq \sum_{x \in Val(X)} \sum_{y \in Val(Y)} g(x,y) p_{XY}(x,y).$$

For continuous random variables X, Y , the analogous expression is

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y) dxdy.$$

We can use the concept of expectation to study the relationship of two random variables with each other. In particular, the **covariance** of two random variables X and Y is defined as

$$Cov[X,Y] \triangleq E[(X - E[X])(Y - E[Y])]$$

Expectation and covariance

Using an argument similar to that for variance, we can rewrite this as,

$$
\begin{aligned}
Cov[X, Y] &= E[(X - E[X])(Y - E[Y])] \\
&= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\
&= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y]] \\
&= E[XY] - E[X]E[Y].
\end{aligned}
$$

Here, the key step in showing the equality of the two forms of covariance is in the third equality, where we use the fact that $E[X]$ and $E[Y]$ are actually constants which can be pulled out of the expectation. When $Cov[X, Y] = 0$, we say that X and Y are **uncorrelated**. However, this is not the same thing as stating that X and Y are independent! For example, if $X \sim Uniform(-1, 1)$ and $Y = X^2$, then one can show that X and Y are uncorrelated, even though they are not independent.

Expectation and covariance

Properties:

- (Linearity of expectation) $E[f(X,Y) + g(X,Y)] = E[f(X,Y)] + E[g(X,Y)]$.
- $Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$.
- If X and Y are independent, then $Cov[X,Y] = 0$.
- If X and Y are independent, then $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$.

Example: Expectation of a Product of Random Variables.

If X, Y are random variables with joint density function $f(x,y)$, then the expectation of the product is given by

$$\mathbb{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\ f(x,y)\,\mathrm{d}x\mathrm{d}y.$$

- If X and Y are *independent* with density function f_X and f_Y, respectively, then $f(x,y) = f_X(x)\, f_Y(y)$. Hence,

$$\begin{aligned}
\mathbb{E}(XY) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\ f(x,y)\,\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\ f_X(x)\, f_Y(y)\,\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty} y\ f_Y(y)\left\{\int_{-\infty}^{\infty} x\ f_X(x)\,\mathrm{d}x\right\}\mathrm{d}y \\
&= \int_{-\infty}^{\infty} y\ f_Y(y)\,\mathbb{E}(X)\,\mathrm{d}y \\
&= \mathbb{E}(X)\int_{-\infty}^{\infty} y\ f_Y(y)\,\mathrm{d}y = \mathbb{E}(X)\,\mathbb{E}(Y).
\end{aligned}$$

Multiple random variables

The notions and ideas introduced in the previous slides can be generalized to more than two random variables.

In particular, suppose that we have n continuous random variables, $X_1(\omega), X_2(\omega), \ldots X_n(\omega)$.

For simplicity of presentation, we focus only on the continuous case, but the generalization to discrete random variables works similarly.

Multiple random variables: Basic properties

We can define

- the **joint distribution function** of $X_1, X_2, \dots, X_n$,
- the **joint probability density function** of $X_1, X_2, \dots, X_n$, and
- the **marginal probability density function** of X_1, and the conditional probability density function of X_1 given $X_2, \dots, X_n$,

as

$$
\begin{aligned}
F_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n) &= P(X_1 \le x_1, X_2 \le x_2, \dots X_n \le x_n) \\
f_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n) &= \frac{\partial^n F_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n)}{\partial x_1 \dots \partial x_n} \\
f_{X_1}(X_1) &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n) dx_2 \dots dx_n \\
f_{X_1|X_2,\dots,X_n}(x_1 \mid x_2, \dots, x_n) &= \frac{f_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n)}{f_{X_2,\dots,X_n}(x_1, x_2, \dots, x_n)}
\end{aligned}
$$

To calculate the probability of an event $A \subseteq \mathbb{R}^n$ we have,

$$
P((X_1, X_2, \dots, X_n) \in A) = \int_{x_1, x_2, \dots, x_n \in A} f_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n) dx_1 dx_2 \dots dx_n \quad (4)
$$

Multiple random variables: Chain rule

From the definition of conditional probabilities for multiple random variables, one can show that

$$\begin{aligned} f(x_1, x_2, \dots x_n) &= f(x_n \mid x_1, x_2, \dots, x_{n-1}) f(x_1, x_2, \dots x_{n-1}) \\ &= f(x_n \mid x_1, x_2, \dots, x_{n-1}) f(x_{n-1} \mid x_1, x_2, \dots x_{n-2}) f(x_1, x_2, \dots x_{n-2}) \\ &= \dots = f(x_1) \prod_{i=2}^{n} f(x_i \mid x_1, \dots x_{i-1}). \end{aligned}$$

Multiple random variables: Independence

For multiple events, $A_1, \ldots, A_k$, we say that $A_1, \ldots, A_k$ are **mutually independent** if for any subset $S \subseteq \{1, 2, \ldots, k\}$, we have

$$P(\cap_{i \in S} A_i) = \prod_{i \in S} P(A_i).$$

Likewise, we say that random variables $X_1, \ldots, X_n$ are independent if

$$f(x_1, \ldots, x_n) = f(x_1) f(x_2) \cdots f(x_n).$$

Here, the definition of mutual independence is simply the natural generalization of independence of two random variables to multiple random variables.

Random vectors

- Suppose that we have n random variables. When working with all these random variables together, we will often find it convenient to put them in a vector $X = [X_1, X_2, \ldots, X_n]^T$.
- We call the resulting vector a **random vector** (more formally, a random vector is a mapping from Ω to $\mathbb{R}^n$).
- It should be clear that random vectors are simply an alternative notation for dealing with n random variables, so the notions of joint PDF and CDF will apply to random vectors as well.

Expectation

- Consider an arbitrary function from $g : \mathbb{R}^n \to \mathbb{R}$. The expected value of this function is defined as

$$E[g(X)] = \int_{\mathbb{R}^n} g(x_1, x_2, \ldots, x_n) f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n, \tag{5}$$

where $\int_{\mathbb{R}^n}$ is n consecutive integrations from $-\infty$ to ∞.

- If g is a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, then the expected value of g is the element-wise expected values of the output vector, i.e., if g is

$$g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{bmatrix},$$

Then,

$$E[g(x)] = \begin{bmatrix} E[g_1(X)] \\ \vdots \\ E[g_m(X)] \end{bmatrix}.$$

Covariance matrix

For a given random vector $X : \Omega \to \mathbb{R}^n$, its covariance matrix Σ is the $n \times n$ square matrix whose entries are given by $\Sigma_{ij} = Cov[X_i, X_j]$. From the definition of covariance, we have

$$
\begin{aligned}
\Sigma &= \begin{bmatrix} Cov[X_1, X_1] & \cdots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \cdots & Cov[X_n, X_n] \end{bmatrix} \\
&= \begin{bmatrix} E[X_1^2] - E[X_1]E[X_1] & \cdots & E[X_1 X_n] - E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n X_1] - E[X_n]E[X_1] & \cdots & E[X_n^2] - E[X_n]E[X_n] \end{bmatrix} \\
&= \begin{bmatrix} E[X_1^2] & \cdots & E[X_1 X_n] \\ \vdots & \ddots & \vdots \\ E[X_n X_1] & \cdots & E[X_n^2] \end{bmatrix} - \begin{bmatrix} E[X_1]E[X_1] & \cdots & E[X_1]E[X_n] \\ \vdots & \ddots & \vdots \\ E[X_n]E[X_1] & \cdots & E[X_n]E[X_n] \end{bmatrix} \\
&= E[XX^T] - E[X]E[X]^T = \ldots = E[(X - E[X])(X - E[X])^T].
\end{aligned}
$$

where the matrix expectation is defined in the obvious way.

Covariance matrix: Properties

The covariance matrix has a number of useful properties:

- $\Sigma \succeq 0$; that is, Σ is positive semidefinite.
- $\Sigma = \Sigma^T$; that is, Σ is symmetric.

The multivariate Gaussian distribution

A random vector $X \in \mathbb{R}^n$ is said to have a multivariate normal (or Gaussian) distribution with mean $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{S}^n_{++}$,[2]

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T(\Sigma)^{-1}(x-\mu)\right).$$

We write this as $X \sim \mathcal{N}(\mu, \Sigma)$.

[2] $\mathbb{S}^n_{++}$ refers to the space of symmetric positive definite $n \times n$ matrices, i.e.,

$$\mathbf{S}^n_{++} = \{A \in \mathbf{R}^{n\times n} : A = A^T \text{ and } x^T A x > 0 \text{ for all } x \in \mathbf{R}^n \text{ such that } x \neq 0\}.$$

The multivariate Gaussian distribution

- Notice that in the case $n = 1$, this reduces the regular definition of a normal distribution with mean parameter μ_1 and variance $\sum_{11}$.

- Generally speaking, Gaussian random variables are extremely useful in machine learning and statistics for two main reasons.

    - they are extremely common when modeling "noise" in statistical algorithms. Quite often, noise can be considered to be the accumulation of a large number of small independent random perturbations affecting the measurement process; by the Central Limit Theorem, summations of independent random variables will tend to "look Gaussian."

    - Gaussian random variables are convenient for many analytical manipulations, because many of the integrals involving Gaussian distributions that arise in practice have simple closed form solutions.