Wednesday, May 28, 2014
2:08 PM

Probability Notes

1 Probability Space (Leon-Garcia 2.1-2.2)

A probability space is the collection $(\Omega, \mathscr{F}, \mathbb{P})$ of:

- a sample space Ω,
- a sigma algebra $\mathscr{F}$, which is a collection of subsets of Ω,
- and a probability measure $\mathbb{P}$ assigning probabilities to each set $\mathscr{A} \in \mathscr{F}$

The sigma algebra $\mathscr{F}$ must contain the empty set $\emptyset$ (as well as the entire sample space Ω) and be closed under complements and countably infinite unions and intersections, i.e.

- $\emptyset \in \mathscr{F}$
- $\mathscr{A} \in \mathscr{F} \implies \mathscr{A}^c \in \mathscr{F}$
- $\mathscr{A}_\mathtt{i} \in \mathscr{F},\ \mathtt{i} \in \{1, 2, \ldots\}, \implies (\bigcup_{\mathtt{i}=1}^{\infty} \mathscr{A}_\mathtt{i}) \in \mathscr{F}$
- $\mathscr{A}_\mathtt{i} \in \mathscr{F},\ \mathtt{i} \in \{1, 2, \ldots\}, \implies (\bigcap_{\mathtt{i}=1}^{\infty} \mathscr{A}_\mathtt{i}) \in \mathscr{F}$

The probability measure $\mathbb{P} : \mathscr{F} \to [0, 1]$, i.e. is a map between the sigma algebra $\mathscr{F}$ and the interval $[0, 1]$ which assigns probabilities to sets in $\mathscr{F}$, and must satisfy $\mathbb{P}(\Omega) = 1$ and *countable additivity*, which states that for any sequence of disjoint sets $\mathscr{A}_\mathtt{i} \in \mathscr{F},\ \mathtt{i} \in \mathbb{N}$, so that $\mathscr{A}_\mathtt{j} \bigcap \mathscr{A}_\mathtt{i} = \emptyset$ for all $\mathtt{j} \neq \mathtt{i}$ we have

$$\mathbb{P}\left(\bigcup_{\mathtt{i}=1}^{\infty} \mathscr{A}_\mathtt{i}\right) = \sum_{\mathtt{i}=1}^{\infty} \mathbb{P}\left(\mathscr{A}_\mathtt{i}\right)$$

From these requirements follow other familiar properties of a probability measure.

1.1 Total Probability

Suppose $\mathscr{F}$ contains sets $\{\mathscr{A}_i | i \in \{1, 2, \ldots\}\}$ which form a partition of Ω, so that $\Omega = \bigcup_{i=1}^{\infty} \mathscr{A}_i$, and $\mathscr{A}_i \bigcap \mathscr{A}_j = \emptyset\ i \neq j$. Then $\mathbb{P}(\mathscr{B}) = \sum_{i=1}^{\infty} \mathbb{P}(\mathscr{A}_i \bigcap \mathscr{B})$.

Any measurable subset E of sample space in the sigma algebra is called an event

Countable set vs Uncountable Set:

In mathematics, a **countable set** is a set with the same cardinality (number of elements) as some subset of the set of natural numbers. A countable set is either a finite set or a *countably infinite* set. Whether finite or infinite, the elements of a countable set can always be counted one at a time and, although the counting may never finish, every element of the set will eventually be associated with a natural number.

Pasted from <http://www.wikiwand.com/en/Countable_set>

Integers and any series similar to integers, primes, Rational numbers are countable.
Real numbers and any subset of real numbers is also uncountable.

Conditional Probability:
P(A|B)= P(AB)/P(B) (AB is the intersection of the sets A and B)
Independent events:
P(EF) = P(E)P(F)

Here we only talk about the events A and B which are the elements of sigma algebra

Bayes Theorem:
P(A|B) = P(B|A)P(A)/P(B) where P(B) can be found by law of total probability

---

2 Random Variables

A random variable X is a map from the sample space to the real numbers $X : \Omega \to \mathbb{R}$ such that for all $x \in \mathbb{R}$, $\{\omega \in \Omega | X(\omega) \leq x\} \in \mathscr{F}$. Notation: capital (red) letters for the random variable, lower case (black) letters for the value it takes. (We will not provide explicit constructions of sigma algebras for continuous random variables for this course, however, the interested reader can look up material on the Borel sigma algebra, which is the sigma algebra generated by the open sets on $\mathbb{R}$ and the idea of a measurable map. A good graduate level mathematics intensive book tailored to operations research audiences is *Probability Essentials, 2nd ed.* by Jean Jacod and Philip Protter (Springer).)

Random variable denoted by capital letter X and value it takes by small x
The cumulative distribution characterizes the random variable. So does the characteristic function (They are duals to each other)

Examples of discrete random variables (Experiment with tossing coin; Balls picked randomly from an urn)
Examples of continous random variables (motion of an atom, time a train arrives at a station, imperfect measurement of oil)

Every Random variable is associated with a Cumulative Distribution Function (CDF):
F_X(x) = P(omega | X(omega) <= x) we know omega is in sigma algebra by definition.

Probability mass function: P(X = x) [This can be found by countable intersections in case of discrete random variable]

Expected Value:
Let *X* be a discrete random variable taking values *x*

1, *x*

2, ... with probabilities *p*

1, *p*

2, ... respectively. Then the expected value of this random variable is the infinite sum

$$\mathrm{E}[X] = \sum_{i=1}^{\infty} x_i \, p_i,$$

Pasted from <http://www.wikiwand.com/en/Expected_value>
Variance: Var[X] = E[(X-E[X])^2]
Eg: Bernouli random variable; Binomial Random variable

Continous Random Variable:
P(X <= x) = FX(x) Probability distribution function. Then if the function is absolutely continous we have density function. The density function is denoted by f_X(x) is given by \partial_x (FX(x)). The density is not a probability.

Expectation and Variance of Random Variables;

A normal distribution is:

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

Pasted from <http://www.wikiwand.com/en/Normal_distribution>

Transformation of Random Variables:

2.4 Transformations Between RVs (Leon-Garcia 3.5, 4.6)

Suppose we have a injective (1 to 1) transformation $\mathbf{G}$ from an open set $\mathscr{O} \subset \mathbb{R}^N$ to $\mathbb{R}^N$. Because it is injective, we may find a unique inverse transformation $\mathbf{G}^{-1} : \mathbf{G}(\mathscr{O}) \to \mathscr{O}$ so that $\mathbf{G}^{-1}(\mathbf{G}(\mathbf{x})) = \mathbf{x}$ for any $\mathbf{x} \in \mathscr{O}$. Suppose also that $\mathbf{G}$ is continuously differentiable, so that we can define a matrix of partial derivatives

$$\mathbf{J}(\mathbf{y}) := \left[\frac{\partial \mathbf{G}_i^{-1}(\mathbf{y})}{\partial y_j}\right]$$

and that the determinant of this matrix is not equal to zero anywhere on $\mathscr{O}$. Now suppose we define the random vector $\mathbf{Y}$ to be the result of operating on $\mathbf{X}$ which takes values in $\mathscr{O}$ with the transformation $\mathbf{G}$, so that

$$\mathbf{Y} := \mathbf{G}(\mathbf{X})$$

We wish to determine the probability density function for $\mathbf{Y}$ from the probability density function for $\mathbf{X}$. It is

$$f_{\mathbf{Y}}(\mathbf{y}) = \begin{cases} f_{\mathbf{X}}(\mathbf{G}^{-1}(\mathbf{y}))|\det \mathbf{J}(\mathbf{y})| & \mathbf{y} \in \mathbf{G}(\mathscr{O}) \\ 0 & \text{otherwise} \end{cases}$$

When the transformation of interest $\mathbf{G}$ is not injective (for instance it maps a two dimensional vector to a one dimensional vector), sometimes it can be made into an injection by properly augmenting it $\mathbf{y}' = (\mathbf{G}(\mathbf{x}), x_{i_1}, \ldots, x_{i_U})$. Then one can use the theorem above to calculate the distribution for $\mathbf{y}'$ and then integrate out $x_{i_1}, \ldots, x_{i_U}$ to get the desired distribution. When this does not work either, there is still another theorem that one can use when there are a finite number of smooth solutions. See your book.

Characterisic Function:

$$\varphi_X(t) = \mathrm{E}\left[e^{itX}\right]$$

Pasted from <http://www.wikiwand.com/en/Characteristic_function_(probability_theory)>

Moment Generating Functions:

E[e^{tX}]

d/dt of moment generating functions evaluated at t = 0 give moments of a distribution

Joint Distributions:

If two RV are mappings on the same sample set, we can talk about joint distribution of the random variables:

F(a,b) = P(X <= a, Y <= b)

Density and mass similar to above;

Conditional Distribution, density;

3.2 Conditional Distribution

X, Y random variables with joint density $f_{X,Y}$. The conditional distribution for X given $Y = y$ is

$$F_{X|Y}(x|y) := \lim_{\Delta \downarrow 0} \frac{\mathbb{P}[X \le x, y < Y \le y + \Delta]}{\mathbb{P}[y < Y \le y + \Delta]}$$

The conditional density is then

$$f_{X|Y}(x, y) := \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

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

where $f_{X,Y}(x, y)$ gives the joint density of X and Y, while $f_X(x)$ gives the marginal density for X. Also in this case it is necessary that

$$f_X(x) > 0$$

.

The relation with the probability distribution of *X* given *Y* is given by:

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

Pasted from <http://www.wikiwand.com/en/Conditional_probability_distribution>

Marginal Distribution
Similarly for continuous random variables, the marginal probability density function can be written as *px*(*x*). This is

$$p_X(x) = \int_y p_{X,Y}(x, y)\ \mathrm{d}y = \int_y p_{X|Y}(x|y)\, p_Y(y)\ \mathrm{d}y,$$

Pasted from <http://www.wikiwand.com/en/Marginal_distribution>

Independent Random Variables:
$F_{X,Y}(x, y) = F_X(x) F_Y(y),$
or equivalently, a joint density

$$f_{X,Y}(x, y) = f_X(x) f_Y(y).$$

Pasted from <http://www.wikiwand.com/en/Independence_(probability_theory)>

(Try out E[X+Y])

If Independent E[XY] = E[X]E[Y] not the other way around
Variance, Covariance and inequalities:

$$\operatorname{COV}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

the **correlation coefficient** between two random variables X and Y is then defined as

$$\rho_{X,Y} := \frac{\operatorname{COV}(X, Y)}{\sqrt{\operatorname{VAR}(X)\operatorname{VAR}(Y)}}$$

For a column random vector $\mathbf{X}$, define the **covariance matrix**

$$\boldsymbol{\Sigma}_{\mathbf{X},\mathbf{X}} := \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\right]$$

so that the $\mathtt{i}, \mathtt{j}$th element of the matrix is

$$[\boldsymbol{\Sigma}_{\mathbf{X},\mathbf{X}}]_{\mathtt{i},\mathtt{j}} := \operatorname{COV}(X_{\mathtt{i}}, X_{\mathtt{j}})$$

The **Markov inequality** says that

$$\mathbb{P}\left[|X| \geq a\right] \leq \frac{\mathbb{E}[|X|]}{a}$$

The **Chebyshev inequality** says that for a random variable X with mean m, variance σ^2

$$\mathbb{P}\left[|X - m| \geq a\right] \leq \frac{\sigma^2}{a^2}$$

The proposition in probability theory known as the **law of total expectation**,[1] the **law of iterated expectations**, the **tower rule**, the **smoothing theorem**, **Adam's Law** among other names, states that if *X* is an integrable random variable (i.e., a random variable satisfying E(| *X* |) < ∞) and *Y* is any random variable, not necessarily integrable, on the same probability space, then

$$\mathrm{E}(X) = \mathrm{E}_Y(\mathrm{E}_{X|Y}(X \mid Y)),$$

Pasted from <http://www.wikiwand.com/en/Law_of_total_expectation>

Random Processes:

A random process (also known as a stochastic process) is a map $X : \mathscr{T} \times \Omega \to \mathbb{R}$ between the sample space and signals, so that to every outcome $\omega \in \Omega$ there is a signal (a function of time)

$$X(t, \omega) \quad t \in \mathscr{T}$$

As in deterministic signals and systems $\mathscr{T}$ can be the real numbers $\mathbb{R}$, in which case X is a continuous time random process, or the integers $\mathbb{Z}$, in which case X is a discrete time random process.

as well as the **auto-correlation** function

$$\mathsf{R}_X(t_1, t_2) := \mathbb{E}\left[X(t_1)X(t_2)\right] = \int xy \mathsf{f}_{X(t_1),X(t_2)}(x, y)\mathsf{d}x\mathsf{d}y$$

and the **auto-covariance** function

$$\mathsf{C}_X(t_1, t_2) := \mathbb{E}\left[(X(t_1) - m_X(t_1))(X(t_2) - m_X(t_2))\right] = \mathsf{R}_X(t_1, t_2) - m_X(t_1)m_X(t_2)$$

4.3 Markov Processes (Leon-Garcia 6.2)

A **Markov random process** is one for which the future is independent of the past given the present. Thus, for arbitrary times $t_1 < t_2 < \cdots < t_N$

$$\mathbb{P}\left[a < X(t_N) \leq b | X(t_1) = x_1, X(t_2) = x_2, \ldots, X(t_{N-1}) = x_{N-1}\right] = \mathbb{P}\left[a < X(t_N) \leq b | X(t_{N-1}) = x_{N-1}\right]$$

When such a density exists, this translates to

$$\mathsf{f}_{X(t_N)|X(t_1),X(t_2),\ldots,X(t_{N-1})}(x_N | x_1, x_2, \ldots, x_{N-1}) = \mathsf{f}_{X(t_N)|X(t_{N-1})}(x_N | x_{N-1})$$

A Interpreting the noise through Brownian motion

The noise w is uncorrelated in time and is such that

$$E[w(t)] = 0, \qquad E[w(t)w(\tau)^T] = W(t)\delta(t-\tau),$$

for a given positive semidefinite symmetric covariance $W(t)$. The above relationship is actually based on the heuristic notion that $w(t)$ corresponds to the time-derivative of a fundamental process denoted by $\beta(t)$ according to

$$w(t) = S(t)\dot{\beta}(t),$$

where S satisfies $SS^T = W$. We say "heuristic" because for stochastic processes even the notion of differentiability and smoothness is not strict. The stochastic process $\beta(t)$ is called a *Wiener process* or a *Brownian motion* if

i. $\beta(0) = 0$

ii. each sample path is continuous

iii. $\beta(t) \sim \mathcal{N}(0, t)$, i.e. it is a zero-mean Gaussian with variance t

iv. for all $0 < t_1 < t_2 < t_3 < \dots$ the random variables

$$\beta(t_1), \beta(t_2) - \beta(t_1), \beta(t_3) - \beta(t_2), \dots$$

are uncorrelated in time.

In the multidimensional case, we have $\beta(t) = [\beta_1(t), \beta_2(t), \dots]$ and it is assumed that these are uncorrelated so that $\beta(t)$ is $\mathcal{N}(0, tI)$, where I is the identity matrix.

The dynamics (1) can be written as the *Itô stochastic differential equation*

$$dx(t) = f(x(t), u(t))dt + L(t)S(t)d\beta(t),$$

which is obtained by dividing (1) by dt. This form suggests that we can regard the change in the state dx as a change in time dt and change in a Brownian motion $d\beta$.

Heuristically, since $\beta(t)$ is $\mathcal{N}(0, tI)$ we can think of $d\beta$ as $\mathcal{N}(0, dtI)$. This will be useful in deriving the HJB equation since using the relationship $wdt = d\beta$ we have

$$E[(dtw(t))(dtw(\tau))^T] = dt^2 E[S(t)d\beta(t)d\beta(\tau)^T S(t)^T] \quad \Rightarrow \quad E[w(t)w(\tau)^T] = \frac{1}{dt}W(t)\delta_{jk},$$

where δ_{jk} is the Kronecker in view of defining $t = t_0 + jdt$, $\tau = t_0 + kdt$.