Probability Formulae

Simon Ellersgaard

ABSTRACT. A very condensed overview of "must know" probability formulae - principally aimed at financial practitioners.

Contents

CHAPTER 1

Conventional Notation

The following table summarizes the symbolism typically used in a probability context. Naturally there are plenty of variations within the literature: e.g. negation of a set, $\neg A$, can also be written as $\sim A$ or A' or A^c. However, I daresay that most of what follows is fairly unambiguous

Notation		*Explanation*
X	::	Random variables are denoted by upper case letters.
Z	::	Standard normal variables are mostly labelled by the letter Z.
$\mathbf{X}$	::	Random vectors are denoted by bold upper case letters.
x_i	::	A particular realization of a random variable X (a variate)
	::	is written in lower case letters.
$\sim$	::	Distributed as symbol. E.g. $Z \sim N(0,1)$.
$\cap$	::	Intersection. Interpreted probabilistically as *and*.
$\cup$	::	Union. Interpreted probabilistically as *or*.
$\neg$	::	Negation.
Ω	::	The universal set (the set of all possible events).
$\emptyset$	::	The empty set.
$\mathbb{P}(X = x_i)$	::	The probability that X takes on the value x_i.
$f(x)$	::	Probability density functions of continuous random variables s.t.
		$f(x)dx = \mathbb{P}(x \leq X \leq x + dx)$.
$p(x)$	::	Probability mass functions of discrete random variables s.t.
	::	$p(x_i) = \mathbb{P}(X = x_i)$.
$F_X(x)$	::	The cumulative distribution function.
$\phi(x)$	::	The pdf of the standard normal variate.
$\Phi(x)$	::	The cdf of the standard normal variate.
$\mathcal{F}$	::	Sigma algebras labelled with upper case calligraphic.
$\mathbb{E}$	::	The expectation operator.
$\mathbb{V}$ar	::	The variance operator.
$\mathbb{C}$ov	::	The covariance operator.
σ	::	The standard deviation.
ρ	::	The correlation coefficient.
$\mathbf{\Sigma}$	::	The covariance matrix.
$\bar{X}$	::	The sample mean.
s^2_{n-1}	::	The sample variance.

CHAPTER 2

Probability Theory

"We must become more comfortable with probability and uncertainty".
- Nate Silver

1. Foundations

DEFINITION 1. A **probability space** is a mathematical triplet $(\Omega, \mathcal{F}, \mathbb{P})$, which is used to model experiments with randomly occurring states. It consists of

(1) A sample space (universal set), Ω, which contains all possible realizations (outcomes) of the experiment.
(2) A sigma algebra (*see below*), $\mathcal{F}$, of all and only events, ω, we would like to consider of the experiment. Notice that the nomenclature *event* is more general than *outcome* in the sense that the former is a subset of Ω, whilst the latter is an element in Ω.
(3) The assignment of probabilities to the events, i.e. a function $\mathbb{P}$ from the events to the probability levels. This assignment is typically chosen in accordance with the Kolmogorov axioms (*see below*).

DEFINITION 2. A **sigma algebra** (or **sigma field** or simply **information set**) $\mathcal{F}$ of the universal set Ω is a non-empty collection of subsets of Ω such that the following hold

(1) $\mathcal{F}$ contains the empty set: $\emptyset \in \mathcal{F}$.
(2) $\mathcal{F}$ is closed under complements, i.e. if $A \in \mathcal{F}$ then $\neg A \in \mathcal{F}$.
(3) $\mathcal{F}$ is closed under countable unions, i.e. if $A_i \in \mathcal{F}$ for $i = 1, 2, ...$ then $\cup_i A_i \in \mathcal{F}$.

Remark. The cardinality (size) of a sigma algebra must necessarily be a subset of the power set (set of all subsets) of the universal set, i.e. $\mathcal{F} \subseteq \mathcal{P}(\Omega) = 2^{\Omega}$. The exact size depends on which events we care to consider.
As suggested the probability assignment, $\mathbb{P}$, is mosly taken to satisfy the following axioms:

DEFINITION 3. The **Kolmogorov axioms** of probability are

(1) The probability of every event A with a sigma algebra $\mathcal{F}$ is a non-negative real number. I.e. $\forall A \in \mathcal{F} : \mathbb{P}(A) \in [0, +\infty)$.
(2) If Ω is the universal set then $\mathbb{P}(\Omega) = 1$.
(3) Let $\{A_i\}_{i=1}^n$ be an arbitrary collection of pairwise disjoint events (i.e. $A_i \cap A_j = \emptyset$ if $i \neq j$) then

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i).$$

COROLLARY 1. Some corollaries of these axioms are:

- Monotinocity: If $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
- $\mathbb{P}(\emptyset) = 0$.

- $\forall A \in \mathcal{F} : \mathbb{P}(A) \in [0,1]$ (this follows immediately from monotonicity).
- $\mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cup B)$.
- $\mathbb{P}(\neg A) = 1 - \mathbb{P}(A)$. E.g. if X is a random variable which takes on real values then $\mathbb{P}(X \leq w) = 1 - \mathbb{P}(X > w)$.

Example: Consider the experiment of tossing an unbiased coin twice. During each toss the coin will come out heads H or tails T with equal likelihood. The sample space Ω is clearly

$$\Omega = \{HH, HT, TH, TT\}$$

where HT means a H comes up in the first toss, followed by a T in the second toss etc.
The *full* sigma algebra consists of $2^4 = 16$ events, viz. $\emptyset$ and its complement Ω, the outcomes ("elementary events") HH, HT, TH, TT, and all unions and complements of those events. It is easy to verify that

$$\begin{aligned}\mathcal{F} =\{&\emptyset, \Omega,\\ &HH, HT, TH, TT,\\ &HH \cup HT, HH \cup TH, HH \cup TT, HT \cup TH, HT \cup TT, TH \cup TT,\\ &HH \cup HT \cup TH, HH \cup HT \cup TT, HH \cup TH \cup TT, HT \cup TH \cup TT\}\end{aligned} \tag{1}$$

Notice that we have written our algebra without any complements. That's because complementary expressions are logically identical to union expressions through De Morgan's Law, e.g. $\neg(HH \cup HT \cup TH) = TT$.

Also, from a purely conceptual point of view, notice that there is no reason why we should operate we a complete sigma-algebra. In fact, it is perfectly possible to consider the above experiment, where we somehow have come to know the outcome of the second toss (a tail). The associated sigma algebra of interest will then be

$$\mathcal{F}_{?T} =\{\emptyset, \Omega, HT, TT, HT \cup TT, HH \cup TH, HH \cup HT \cup TH, HH \cup TH \cup TT\}$$

which evidently is a subset of $\mathcal{F}$.
Finally, the probability measure in our sample is

$$\mathbb{P}(HH) = \mathbb{P}(HT) = \mathbb{P}(TH) = \mathbb{P}(TT) = 0.25,$$

where $\mathbb{P}$ is assumed to obey the Kolmogorov axioms. This completes our specification for the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ of the coin tossing experiment.

2. Half-forgotten Memories from a High School Textbook

DEFINITION 4. If $A_1, A_2, ..., A_n$ are **mutually independent** events then

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} \mathbb{P}(A_i). \tag{2}$$

THEOREM 1. **Bayes' Theorem** Let A and B be two events in Ω. The conditional probability of A given B is

$$\mathbb{P}(A|B) \equiv \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{3}$$

or, identically, since $\mathbb{P}(B|A) = \mathbb{P}(B \cap A)/\mathbb{P}(A)$ we must have

$$\mathbb{P}(A|B) \equiv \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}. \tag{4}$$

Observe that if A and B are independent then $\mathbb{P}(A|B) = \mathbb{P}(A)$.

THEOREM 2. **Law of Total Probability:** Suppose we divide Ω into n pairwise disjoint events $\{A_i\}_{i=1}^n$ whose union is the entire sample space, i.e. if $i \neq j$ then $A_i \cap A_j = \emptyset$ and $\cup_{i=1}^n A_i = \Omega$, the the probability of B (which is the probability of $B \cap \Omega$) can be written

$$\begin{aligned} \mathbb{P}(B) &= \sum_{i=1}^{n} \mathbb{P}(B \cap A_i) \\ &= \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i). \end{aligned} \tag{5}$$

COROLLARY 2. Let $\{A_j\}_{i=1}^n$ be a partition of Ω. The **Extended Form of Bayes' Theorem** is

$$\begin{aligned} \mathbb{P}(A_i|B) &= \frac{\mathbb{P}(A_i \cap B)}{\sum_{j=1}^{n} \mathbb{P}(A_j \cap B)} \\ &= \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{j=1}^{n} \mathbb{P}(B|A_j)\mathbb{P}(A_j)}, \end{aligned} \tag{6}$$

which in the case of a binary partition of Ω reduces to

$$\mathbb{P}(A|B) \equiv \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|\neg A)\mathbb{P}(\neg A)}. \tag{7}$$

Example: Suppose a drug test is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose that 0.5% of people are users of the drug. If a randomly selected individual tests positive, what is the probability he or she is a user, $\mathbb{P}(U|+)$?

$$\mathbb{P}(U|+) = \frac{\mathbb{P}(+|U)\mathbb{P}(U)}{\mathbb{P}(+|U)\mathbb{P}(U) + \mathbb{P}(+|\neg U)\mathbb{P}(\neg U)} = \frac{0.99 \cdot 0.005}{0.99 \cdot 0.005 + 0.01 \cdot 0.995} = 33.2\%.$$

So it is in fact more likely that the person is NOT a user of the drug, despite the accuracy of the test.

CHAPTER 3

RVs, PMFs, PDFs and CDFs

It is remarkable that a science which began with the consideration of games of chance should have become the most important object of human knowledge.
-Pierre Simon Laplace

1. Random Variables

DEFINITION 5. A **random variable** (**RV**), X, on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a mapping from the sample space into the reals, $X : \Omega \mapsto \mathbb{R}$, such that the set $\{X \leq r\}$ is an event ($\in \mathcal{F}$) for any real number r[1].

Remark. Discrete random variables take on a finite (or countably infinite) number of realizations. Continuous variables take on values that vary continuously within one or more (possibly infinite) intervals. As a result there are an uncountably infinite number of individual realizations.

DEFINITION 6. A **random variate**, x, is a particular realization of a random variable, X.

2. Single Variable PMFs, PDFs and CDFs

DEFINITION 7. A probability mass function (**PMF**) $p(x)$ for the discrete random variable X given the probability for each realization of $X : \{x_1, ..., x_n\}$. I.e. $p(x_i) = \mathbb{P}(X = x_i)\ \forall i$. A probability mass function is non-negative $\forall i$, and must sum to unity: $\sum_i p(x_i) = 1$. If A is some subset of all these x_i (i.e. A is an event) then

$$\mathbb{P}(X \in A) = \sum_{x_i \in A} p(x_i). \tag{8}$$

DEFINITION 8. For a continuous random variable, the probability of any single realization, $X = x$, is zero. Informally, we may define the probability density function (**PDF**) $f(x)$ of some random variable X, times the increment dx, as the probability that X is between x and $x + dx$: $f(x)dx = \mathbb{P}(x \leq X \leq x + dx)$. The density must be everywhere non-negative and must integrate to unity: $\int_{-\infty}^{+\infty} f(x)dx = 1$. More formally, the probability that $x \in [a, b]$ is

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx. \tag{9}$$

[1]The above definition is a special case of a more abstract definition, which describes a random variable X as a measurable function from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to the measurable space $(\Sigma, \mathcal{G})$. By measurable space we understand a set together with its sigma algebra. By measurable function we understand that for every subset $\omega_{\mathcal{G}} \in \mathcal{G}$, its pre-image $X^{-1}(\omega_{\mathcal{G}}) \in \mathcal{F}$, whence any subset in the target space can be measured by looking at the pre-image.

DEFINITION 9. The cumulative distribution function (**CDF**) $F_X(x)$ for the random variable X is the probability of X being less than or equal to x: $F_X(x) \equiv \mathbb{P}(X \leq x)$. For a discrete random variable this means that

$$F_X(x) = \sum_{x_i \leq x} p(x_i), \tag{10}$$

whilst for a continuous random variable

$$F_X(x) = \int_{-\infty}^{x} f(x')dx'. \tag{11}$$

2.1. Properties. Observe that

- $$\lim_{x \to -\infty} F_X(x) = 0, \qquad \lim_{x \to \infty} F_X(x) = 1$$
- $F_X(x)$ is non-decreasing in x.
- $F_X(x)$ is right-continuous.
- Finally, for continuous random variables, by Leibniz Integral Rule:

$$f(x) = \frac{dF_X(x)}{dx}. \tag{12}$$

Remark. It is not unusual to see the notation $d\mathbb{P}(X = x)$ for $f(x)dx$. If we define $d\mathbb{P}(X = x) \equiv \mathbb{P}(x \leq X \leq x + dx) = \mathbb{P}(X \leq x + dx) - \mathbb{P}(X \leq x)$ this makes quite good sense as the latter can be written as $F_X(x + dx) - F_X(x)$ or simply $dF_X(x)$. So $f(x)dx = dF_X(x)$ which is, of course, just a restatement of (12).

3. The Bivariate Case

DEFINITION 10. Let X and Y be discrete random variables. The **joint probability mass function** is $p(x_i, y_j) = \mathbb{P}(X = x_i \cap Y = y_j)$ and it satisfies the usual requirements of non-negativity $\forall i \forall j$, and that $\sum_i \sum_j p(x_i, y_j) = 1$. If $A \subset \mathbb{R}^2$ is some event, i.e. some collection of 2-tuples $\{(x_i, y_j)\}$, where all $x_i \in \Omega_X$ and $y_j \in \Omega_Y$ then the probability of A is

$$\mathbb{P}((X, Y) \in \mathcal{A}) = \sum_{x_i \in A} \sum_{y_j \in A} p(x_i, y_j). \tag{13}$$

DEFINITION 11. For continuous random variables, X and Y, the probability of the realizations $X = x$ and $Y = y$ is zero. Informally, the **joint probability density function** $f(x, y)$ of the random variables X and Y, times the infinitesimal area $dxdy$, gives the probability that X is between x and $x + dx$ and Y is between y and $y + dy$: $f(x, y)dxdy = \mathbb{P}(\{x \leq X \leq x + dx\} \cap \{y \leq Y \leq y + dy\})$. The joint density has the property of non-negativity and $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)dxdy = 1$. More formally, the probability that $(x, y) \in A \subseteq \mathbb{R}^2$ is

$$\mathbb{P}((X, Y) \in \mathcal{A}) = \iint_A f(x, y)dxdy \tag{14}$$

DEFINITION 12. The **joint cumulative distribution function**, $F_{X,Y}(x, y)$, of the random variates X and Y is the probability that X is less than x and Y is less than y: $F_{X,Y}(x, y) \equiv \mathbb{P}(X \leq x \cap Y \leq y)$. For discrete random variables this amounts to

$$F_{X,Y}(x,y) = \sum_{x_i \leq x} \sum_{y_j \leq y} p(x_i, y_j), \tag{15}$$

while, for continuous random variables, the joint CDF is

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x', y') dx' dy'. \tag{16}$$

For the latter, assuming continuity in x, y, we get

$$f(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}. \tag{17}$$

These results are straightforwardly generalizable to higher order mass and density functions.

4. Marginal Densities and The Law of Total Probability

DEFINITION 13. We define the **marginal density** as the bivariate density where we have *summed out* or *integrated out* one of the variables, such that only the density of one of the variables remains. In particular, for discrete random variables

$$p_X(x_i) = \sum_{\text{all } j} p(x_i, y_j) \tag{18}$$

For continuous random variables:

$$f_X(x) \equiv \int_{-\infty}^{+\infty} f(x,y) dy \tag{19}$$

Remark. Analogously we can define $p_Y(y_j)$ and $f_Y(y)$, by summing over all i and integrating over x.

Remark. Take care to note that the marginals $p_X(x_i)$, $f_X(x)$ correspond to the probability mass/density functions for X disregarding Y. To see this, consider the continuous case, where we are interested in $\mathbb{P}(a \leq X \leq b)$. This must be equal to $\mathbb{P}(\{a \leq X \leq b\} \cap \{-\infty \leq y \leq +\infty\})$ and thence $\int_a^b \int_{-\infty}^{+\infty} f(x,y) dx dy \equiv \int_a^b f_X(x) dx$. And $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x) dx$ should remind you of how we defined densities in the first place.

The Law of Total Probability. Recall that the law of total probability states that if $\{A_i\}_{i=1}^n$ is some partition of the sample space then $\mathbb{P}(B) = \sum_{i=1}^n \mathbb{P}(B|A_i)\mathbb{P}(A_i)$. We can extend this to random variables as follows:

THEOREM 3. Let X, Y be discrete random variables and let $A \subset \mathbb{R}^2$ then the law of total probability implies

$$\mathbb{P}((X,Y) \in A) = \sum_i \mathbb{P}((X,Y) \in A | X = x_i) \mathbb{P}(X = x_i) \tag{20}$$

THEOREM 4. Let X, Y be continuous random variables and let $A \subseteq \mathbb{R}^2$ then the law of total probability implies

$$\mathbb{P}((X,Y) \in A) = \int \mathbb{P}((X,Y) \in A | X = x) d\mathbb{P}(X = x) \tag{21}$$

where $d\mathbb{P}(X = x) = f_X(x)dx$.

Remark. Equation (21) emerges when we consider the limiting expression from the discrete law of total probability:

$$\mathbb{P}((X,Y) \in A) = \lim_{dx \to 0} \sum_{\text{intervals}} \mathbb{P}((X,Y) \in A | x \leq X \leq x + dx) \mathbb{P}(x \leq X \leq x + dx).$$

DEFINITION 14. If X and Y are **independent** random variables then the joint density can be written as a product of the marginals:

$$f(x,y) = f_X(x) f_Y(y). \tag{22}$$

5. Conditional PDFs

THEOREM 5. Let X be a continuous random variable with associated PDF $f(x)$ and CDF $F_X(x)$. The probability density for X given that X is greater than α is

$$f(x | x > \alpha) = \frac{f(x)}{1 - F_X(\alpha)}. \tag{23}$$

PROOF. Probability density functions are derived from cumulative distribution functions through a differentiation.

$$\begin{aligned} f(x|x>\alpha) &= \frac{\partial}{\partial x} F_{X|X>\alpha}(x) \equiv \frac{\partial}{\partial x} \mathbb{P}(X \leq x | X > \alpha) \\ &= \frac{\frac{\partial}{\partial x} \mathbb{P}(\{X \leq x\} \cap \{X > \alpha\})}{\mathbb{P}(X > \alpha)} = \frac{\frac{\partial}{\partial x} \mathbb{P}(\alpha < X \leq x)}{\mathbb{P}(X > \alpha)} \\ &= \frac{\frac{\partial}{\partial x} [\mathbb{P}(X \leq x) - \mathbb{P}(X \leq \alpha)]}{1 - \mathbb{P}(X \leq \alpha)} \equiv \frac{\frac{\partial}{\partial x} [F_X(x) - F_X(\alpha)]}{1 - F_X(\alpha)} \\ &= \frac{f(x)}{1 - F_X(\alpha)}. \end{aligned}$$

□

The step of deriving a PDF from a CDF (suitably converted through its definition in $\mathbb{P}$) is commonplace and will also aid us in the proof of the following theorem.

THEOREM 6. The **conditional density** of X given Y is defined as

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}, \tag{24}$$

where $f_Y(y)$ is the marginal density. This, of course, is a reminiscent of equations (6).

PROOF. Formally we can prove the theorem in the following manner

$$\begin{aligned}
f_{X|Y}(x|y) &= \frac{\partial}{\partial x} F_{X|Y}(x) \equiv \frac{\partial}{\partial x} \lim_{\epsilon \to 0} \mathbb{P}(X \leq x | y \leq Y \leq y + \epsilon) \\
&= \frac{\partial}{\partial x} \lim_{\epsilon \to 0} \frac{\mathbb{P}(\{X \leq x\} \cap \{y \leq Y \leq y + \epsilon\})}{\mathbb{P}(y \leq Y \leq y + \epsilon)} \\
&= \frac{\partial}{\partial x} \lim_{\epsilon \to 0} \frac{[\mathbb{P}(\{X \leq x\} \cap \{Y \leq y + \epsilon\}) - \mathbb{P}(\{X \leq x\} \cap \{Y \leq y\})]}{\mathbb{P}(y \leq Y \leq y + \epsilon)} \\
&\equiv \frac{\partial}{\partial x} \lim_{\epsilon \to 0} \frac{[F_{X,Y}(x, y + \epsilon) - F_{X,Y}(x, y)]}{F_Y(y + \epsilon) - F_Y(y)} \\
&= \frac{\partial}{\partial x} \lim_{\epsilon \to 0} \frac{\frac{\partial}{\partial y} F_{X,Y}(x, y + \epsilon)\epsilon}{f_Y(y + \epsilon)\epsilon} \\
&= \frac{\frac{\partial^2}{\partial y \partial x} F_{X,Y}(x, y)}{f_Y(y)} \\
&= \frac{f(x, y)}{f_Y(y)}.
\end{aligned}$$

□

6. Bijective Transformations of Random Variables

THEOREM 7. Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ and $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ be two collections of random variables related by a bijective function $\mathcal{G} : \mathbf{X} \mapsto \mathbf{Y}$. If $\mathbf{X}$ has joint density $f_{\mathbf{X}}(x_1, x_2, ..., x_n)$ and $\mathbf{Y}$ has density $f_{\mathbf{Y}}(y_1, y_2, ..., y_n)$ then

$$f_{\mathbf{X}}(x_1, x_2, ..., x_n) = f_{\mathbf{Y}}(y_1, y_2, ..., y_n) \text{abs}\left[\mathcal{J}\right] \tag{25}$$

where the Jacobian is defined as

$$\mathcal{J} \equiv \frac{\partial(y_1, y_2, ..., y_n)}{\partial(x_1, x_2, ..., x_n)} \equiv \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{vmatrix}$$

This is the density transformation rule for random variables. A useful mnemonic for (25) is $f(y)dy = f(x)dx$.

Example: According to the Box-Müller Method, we can generate two independent normally distributed random variates $z_1, z_2 \sim N(0, 1)$ from two uniformly distributed random variates $u_1, u_2 \sim U(0, 1)$ through the equations $z_1 = \sqrt{-2 \ln u_1} \sin(2\pi u_2)$ and $z_2 = \sqrt{-2 \ln u_1} \cos(2\pi u_2)$. To see this, notice that $f_{\mathbf{U}}(u_1, u_2) = f_{\mathbf{U}}(u_1) f_{\mathbf{U}}(u_2) = 1$ while $f_{\mathbf{Z}}(z_1, z_2) = f_{\mathbf{Z}}(z_1) f_{\mathbf{Z}}(z_2) = \frac{1}{\sqrt{2\pi}} \exp(-z_1^2/2) \frac{1}{\sqrt{2\pi}} \exp(-z_2^2/2)$. Hence, by (25), to get $f_{\mathbf{Z}}(z_1, z_2) = f_{\mathbf{U}}(u_1, u_2) \text{abs}\left[\partial(u_1, u_2)/\partial(z_1, z_2)\right]$ we must require that

$$\text{abs}\left[\frac{\partial(u_1, u_2)}{\partial(z_1, z_2)}\right] = \frac{1}{2\pi} \exp\left(-\frac{z_1^2 + z_2^2}{2}\right)$$

which is fairly trivial to prove, by rewriting the Box-Müller equations as $u_1 = \exp(-(z_1^2 + z_2^2)/2)$ and $u_2 = \frac{1}{2\pi} \arctan(z_2/z_1)$.

7. Sum of Independent Random Variables

THEOREM 8. Let X and Y be two *independent* random variables with density functions $f_X(x)$ and $f_Y(y)$ defined for all x, y. Then the sum $Z \equiv X + Y$ is a random variate with density function $f_Z(z)$, where f_Z is the *convolution* of f_X and f_Y i.e.

$$f_Z(z) = (f_X * f_Y)(z) \tag{26}$$

where

$$(f_X * f_Y)(z) \equiv \int_{-\infty}^{+\infty} f_X(z-y) f_Y(y) dy \equiv \int_{-\infty}^{+\infty} f_X(x) f_Y(z-x) dx.$$

PROOF. The proof takes its natural starting point the the CDF of Z.

$$F_Z(z) = \mathbb{P}(Z \leq z) \equiv \mathbb{P}(X + Y \leq z) = \mathbb{P}(x \leq z - Y)$$

Now using the law of total probability, (21), on the last equality we get

$$F_Z(z) = \int_{-\infty}^{+\infty} \mathbb{P}(X \leq z - Y | Y = y) d\mathbb{P}(Y = y)$$

or, as X and Y are independent variables simply

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^{+\infty} \mathbb{P}(X \leq z - Y) d\mathbb{P}(Y = y) \\ &= \int_{-\infty}^{+\infty} F_X(z-y) f_Y(y) dy \end{aligned}$$

In the last line we have used the definitions of the CDF and PDF respectively. Finally, using (12), and the fact derivative operators can be taken under the integral sign when the integration limits are independent of z:

$$f_Z(z) = \frac{\partial}{\partial z} F_Z(z) = \int_{-\infty}^{+\infty} \frac{\partial}{\partial z} \left(F_X(z-y) f_Y(y) \right) dy = \int_{-\infty}^{+\infty} f_X(z-y) f_Y(y) dy.$$

□

Example: The sum of two independent normal distributions $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ is also normal with distribution $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. To see this, use the convolution theorem to establish

$$\int_{-\infty}^{+\infty} \frac{\exp\left(-\frac{(z-x_2-\mu_1)^2}{2\sigma_1^2}\right)}{\sqrt{2\pi\sigma_1^2}} \cdot \frac{\exp\left(-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right)}{\sqrt{2\pi\sigma_2^2}} dx_2 = \frac{\exp\left(-\frac{(z-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right)}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}}.$$

CHAPTER 4

Convergence of Random Variables

1. Convergence in probability

THEOREM 9. A random variable $X_n \in \mathbb{R}$ **converges in probability** to X as $n \to \infty$, denoted $X_n \xrightarrow{p} X$ (X is the **plim** of X_n), if for all $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \leq \delta) = 1. \tag{27}$$

The convergence is **strong** or **almost sure**, denoted $X_n \xrightarrow{a.s.} X$, if

$$\mathbb{P}\left(\lim_{n\to\infty} |X_n - X| \leq \delta\right) = 1. \tag{28}$$

Almost sure convergence is stronger than convergence in probability in the sense $X_n \xrightarrow{a.s.} X$ implies $X_n \xrightarrow{p} X$, but not vice versa.

THEOREM 10. The **weak law of large numbers** states that if $\mathbb{E}|X| < \infty$ then as $n \to \infty$,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mathbb{E}[X_i]. \tag{29}$$

THEOREM 11. The **strong law of large numbers** states that if $\mathbb{E}|X| < \infty$ then as $n \to \infty$,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{a.s.} \mathbb{E}[X_i]. \tag{30}$$

2. Convergence in distribution

THEOREM 12. A random variate X_n **converges in distribution** to a random variable X, denoted $X_n \xrightarrow{d} X$, if

$$\lim_{n\to\infty} F_n(x) = F(x), \tag{31}$$

for every number $x \in \mathbb{R}$ at which F is continuous.

THEOREM 13. The **Lindeberg-Levy Central Limit Theorem** states that if $\{X_i\}_{i=1}^n$ is a sequence of independent and identically distributed (i.i.d.) random variables (**not** necessarily normal) with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}\text{ar}[X_i] = \sigma^2 < \infty$, then as $n \to \infty$ then

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2), \tag{32}$$

where $\bar{X} = \sum_{i=1}^n X_i/n$. This generalizes to the multivariate case in the following manner. Let $\{\mathbf{X}_i\}_{i=1}^n$ be i.i.d. vectors with $\mathbb{E}[\mathbf{X}_i] = \boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ then

$$\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}), \tag{33}$$

where $\bar{\mathbf{X}} = \sum_{i=1}^{n} \mathbf{X}_i/n$.

The following extension is also noteworthy

THEOREM 14. According to the **Delta Method**, if $\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$, then the distribution of a function h of $\bar{\mathbf{X}}$ is

$$\sqrt{n}(h(\bar{\mathbf{X}}) - h(\boldsymbol{\mu})) \xrightarrow{d} N(\mathbf{0}, \nabla h(\boldsymbol{\mu})^\intercal \boldsymbol{\Sigma} \nabla h(\boldsymbol{\mu})), \tag{34}$$

The proof of this is quite easy. It follows by making the Taylor expansion $h(\bar{\mathbf{X}}) \approx h(\boldsymbol{\mu}) + \nabla h(\boldsymbol{\mu})^\intercal (\bar{\mathbf{X}} - \boldsymbol{\mu})$ and then computing the expectation and variance.

CHAPTER 5

Expectations

"How dare we speak of the laws of chance? Is not chance the antithesis of all law? ".
- Joseph Bertrand

1. Unconditional Expectations

DEFINITION 15. Let X be a random variable. If X is discrete then its **expected value** is given by the formula

$$\mathbb{E}[X] \equiv \sum_{x_i \in X} x_i p(x_i). \tag{35}$$

If X is continuous then the equation is given by

$$\mathbb{E}[X] \equiv \int_{-\infty}^{+\infty} x f(x) dx. \tag{36}$$

THEOREM 15. The **Law of the Unconscious Statistician** decrees how to compute expected values of a function $g(X)$ of a random variable X, when one knows the distribution of X (but not of $g(X)$):

$$\text{(discrete)} \qquad \mathbb{E}[g(X)] \equiv \sum_{x_i \in X} g(x_i) p(x_i) \tag{37}$$

$$\text{(continuous)} \qquad \mathbb{E}[g(X)] \equiv \int_{-\infty}^{+\infty} g(x) f(x) dx \tag{38}$$

PROOF. Suppose the random variable X has pdf $f_X(x)$ and thence expectation value as in (36). Now define a new random variable $Y = g(X)$ then $\mathbb{E}[Y]$ is given by

$$\mathbb{E}[Y] = \int_{-\infty}^{+\infty} y f_Y(y) dy$$

where $f_Y(y)$ is the pdf of Y. However, from the law of the unconscious statistician $|f_X(x)dx| = |f_Y(y)dy|$, from which (38) follows immediately. □

DEFINITION 16. Expectation values can be extended to the multiple variables. For the bivariate discrete case

$$\mathbb{E}[X] = \sum_{x_i \in X} \sum_{y_j \in Y} x_i p(x_i, y_j). \tag{39}$$

Correspondingly, for continuous random variables X and Y:

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f_{X,Y}(x,y) dx dy. \tag{40}$$

The law of the unconscious statistician also holds in the bivariate case, i.e.

$$\text{(discrete)} \qquad \mathbb{E}[g(X,Y)] \equiv \sum_{x_i \in X} \sum_{y_j \in Y} g(x_i, y_j) p(x_i, y_j) \tag{41}$$

$$\text{(continuous)} \qquad \mathbb{E}[g(X,Y)] \equiv \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x,y) f(x,y) dx dy \tag{42}$$

Indeed we can generalize these results to density functions with an arbitrary number of random variables.

Properties: The exponential operator satisfies the following properties. Let X_i be random variables for all i (not necessarily statistically independent) and let α be an arbitrary constant, then

- $\mathbb{E}[X + \alpha] = \mathbb{E}[X] + \alpha$.
- $\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$.
- $\mathbb{E}[X_1 + X_2 + ... + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + ... + \mathbb{E}[X_n]$.
- If X_1 and X_2 are independent then $\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$ and indeed we have $\mathbb{E}[g(X_1)f(X_2)] = \mathbb{E}[g(X_1)]\mathbb{E}[g(X_2)]$ for any functions g and f (this follows straightforwardly by combining (22) and (42) for the continuous case and similarly for the discrete). The converse statement, that "$\mathbb{E}[X_1 X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$ implies independence of X_1 and X_2", is **false**.
- The Cauchy-Schwarz inequality for probabilities is: $|\mathbb{E}[XY]|^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$.

1.1. Expectation Theorems.

DEFINITION 17. Define the **Characteristic Function** $\varphi_X(t) = \mathbb{E}[e^{itX}]$ of a random variable X as

$$\mathbb{R} \mapsto \mathbb{C} : \varphi_X(t) \equiv \int_{-\infty}^{+\infty} e^{itx} f(x) dx, \tag{43}$$

where $i = \sqrt{-1}$. Then by Fourier transformation, supposing $\varphi_X(t)$ is integrable,

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \varphi_X(t) dt. \tag{44}$$

Properties.

- Characteristic functions can be used to find moments of a random variable. In particular
$$\mathbb{E}[X^n] = i^{-n} \left. \frac{d^n \varphi_X(t)}{dt^n} \right|_{t=0}$$
This is because of the Taylor expansion $e^{itX} = 1 + itX + i^2 t^2 X^2/2! + ... + i^n t^n X^n/n! + ...$ hence $\varphi_X(t) = 1 + itm_1 + i^2 t^2 m_2/2! + ... + i^n t^n m_n/n! + ...$ where m_n is the n^{th} moment.

- Consider a collection of *independent* and not necessarily identically distributed variables $\{X_1, X_2, ..., X_n\}$. Define $S_n = \sum_{i=1}^{n} a_i X_i$ where the a_i are constants. Then the characteristic function of S_n is given by

$$\varphi_{S_n}(t) = \varphi_{X_1}(a_1 t)\varphi_{X_2}(a_2 t)...\varphi_{X_n}(a_n t).$$

  This is useful if we wish to calculate PDFs of sums of random variables like $X + Y$. All we need to know is the characteristic functions of X and Y then by (44)

$$f_{X+Y}(x, y) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-itx} \varphi_X(t)\varphi_Y(t)dt.$$

DEFINITION 18. The **matrix expectation** straightforwardly operates on each of the matrix elements:

$$\mathbb{E}\begin{pmatrix} X_{1,1} & \cdots & X_{1,n} \\ \vdots & \ddots & \vdots \\ X_{m,1} & \cdots & X_{m,n} \end{pmatrix} = \begin{pmatrix} \mathbb{E}[X_{1,1}] & \cdots & \mathbb{E}[X_{1,n}] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[X_{m,1}] & \cdots & \mathbb{E}[X_{m,n}] \end{pmatrix} \tag{45}$$

THEOREM 16. Let $\xi : X \mapsto \mathbb{R}$ be a *convex* function, then **Jensen's Inequality** states that

$$\xi(\mathbb{E}[X]) \leq \mathbb{E}[\xi(X)]. \tag{46}$$

PROOF. (Discrete) The proof runs by induction. First, recall the definition of convexity, $\forall x_1, \forall x_2 \in X$ and $\{\forall p_1 \forall p_2 \in [0, 1] | p_1 + p_2 = 1\}$

$$\xi(p_1 x_1 + p_2 x_2) \leq p_1 \xi(x_1) + p_2 \xi(x_2)$$

Assume this result holds for $\{\forall p_1 \forall p_2 ... \forall p_k \in [0, 1] | p_1 + p_2 + ... + p_k = 1\}$:

$$\xi(p_1 x_1 + p_2 x_2 + ... + p_k x_k) \leq p_1 \xi(x_1) + p_2 \xi(x_2) + ... + p_k \xi(x_k)$$

then we need to show that the analogous result for $k + 1$ terms is implied. Clearly, $\exists p$ which is *strictly positive* (let us label this as p_1) then

$$\begin{aligned} \xi\left(p_1 x_1 + p_2 x_2 + ... + p_{k+1} x_{k+1}\right) &= \xi\left(p_1 x_1 + (1 - p_1) \sum_{i=2}^{k+1} \tfrac{p_i}{1-p_1} x_i\right) \\ &\leq p_1 \xi(x_1) + (1 - p_1) \xi\left(\sum_{i=2}^{k+1} \tfrac{p_i}{1-p_1} x_i\right). \end{aligned}$$

Since $\sum_{i=2}^{k+1} \frac{p_i}{1-p_1} = 1$ we can apply the induction hypothesis to the last term to get out the desired result. □

NB: In order to remember the direction of the inequality in Jensen's formula, one can conveniently recall that variance is non-negative. In particular $\mathbb{V}\text{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \geq 0$, where $\xi(x) = x^2$ is a convex function.

2. Conditional Expectations

DEFINITION 19. We define the **conditional expectation** for the discrete random variable X given Y as

$$\mathbb{E}[X|Y = y_j] = \sum_{x_i \in X} x_i \mathbb{P}(X = x_i|Y = y_j) \tag{47}$$

and for the continuous case

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{+\infty} x f_{X|Y}(x|y)dx. \tag{48}$$

Properties: Let X, X_i for $i = 1, 2, 3$ be random variable, let α, β be constants and let $f : \mathbb{R} \mapsto \mathbb{R}$ then

- $\mathbb{E}[\alpha|X] = \alpha$.
- $\mathbb{E}[\alpha X_1 + \beta X_2|X_3] = \alpha\mathbb{E}[X_1|X_3] + \beta\mathbb{E}[X_2|X_3]$.
- $\mathbb{E}[X_1|X_2] \geq 0$ if $X_1 \geq 0$.
- $\mathbb{E}[X_1|X_2] = \mathbb{E}[X_1]$ if X_1 and X_2 are independent.
- $\mathbb{E}[X_1 f(X_2)|X_2] = f(X_2)\mathbb{E}[X_1|X_2]$.
- $\mathbb{E}[X_1|X_2 \cap f(X_2)] = \mathbb{E}[X_1|X_2]$.

THEOREM 17. The **Law of Iterated Expectations** (or **Tower Rule** or **Law of Total Expectations**) states that if X is an integrable random variable and Y is another (not necessarily integrable) random variable then

$$\mathbb{E}[X] = \mathbb{E}_y[\mathbb{E}[X|Y]] \tag{49}$$

In a financial context, this law is often stated as follows: let $\mathcal{F}_t \subseteq \mathcal{F}_s$ be two sigma fields then

$$\mathbb{E}[X|\mathcal{F}_t] = \mathbb{E}[\mathbb{E}[X|\mathcal{F}_s]|\mathcal{F}_t]. \tag{50}$$

Conventionally, in a time series context when we have some information at time t and even more information at a later time s, we write

$$\mathbb{E}_t[X] = \mathbb{E}_t[\mathbb{E}_s[X]] \tag{51}$$

A heuristic way to remember this formulation is *today's forecast of tomorrow's forecast of some event is simply today's forecast of that event*.

PROOF. We shall prove (49) for the discrete case. The right hand side can be written as

$$\begin{aligned}\mathbb{E}_y[\mathbb{E}[X|Y]] &= \mathbb{E}_y\left[\sum_{x_j\in X} x_j\mathbb{P}(X=x_j|Y)\right]\\ &= \sum_{y_i\in Y}\sum_{x_j\in X} x_j\mathbb{P}(X=x_j|Y=y_i)\mathbb{P}(Y=y_i)\\ &= \sum_{x_j\in X} x_j\sum_{y_i\in Y}\mathbb{P}(X=x_j|Y=y_i)\mathbb{P}(Y=y_i)\\ &= \sum_{x_j\in X} x_j\mathbb{P}(X=x_j)\equiv\mathbb{E}[X].\end{aligned}$$

In the fourth line, we used the law of total probability (5). □

COROLLARY 3. By comparing the first and last line, observing that we can take $\sum x_j$ outside our expectation, we get

$$\mathbb{P}(X=x_j)=\mathbb{E}[\mathbb{P}(X=x_j|Y)]. \tag{52}$$

CHAPTER 6

Variance & Covariance

Certainty is the mother of quiet and repose, and uncertainty the cause of variance and contentions.
-Edward Coke

1. Of Random Variables

DEFINITION 20. We define the **variance** of a random variable X as

$$\mathbb{V}\text{ar}[X] \equiv \mathbb{E}[(X - \mathbb{E}[X])^2] \tag{53}$$

which trivially can be rewritten as

$$\mathbb{V}\text{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{54}$$

The **standard deviation** or **volatility** of X is defined as

$$\sigma(X) \equiv \sqrt{\mathbb{V}\text{ar}[X]}. \tag{55}$$

COROLLARY 4. Thus, for a discrete random variable X the variance is

$$\mathbb{V}\text{ar}[X] = \sum_{x_i \in X} (x_i - \mathbb{E}[X])^2 p(x_i) \tag{56}$$

and for the continuous case

$$\mathbb{V}\text{ar}[X] = \int_{-\infty}^{+\infty} (x - \mathbb{E}[X])^2 f(x) dx. \tag{57}$$

Properties: The variance operator satisfies the following properties. Let X_i be random variables for all i (not necessarily statistically independent) and let α be an arbitrary constant, then

- $\mathbb{V}\text{ar}[X] \geq 0$.
- $\mathbb{V}\text{ar}[\alpha] = 0$.
- $\mathbb{V}\text{ar}[X + \alpha] = \mathbb{V}\text{ar}[X]$.
- $\mathbb{V}\text{ar}[\alpha X] = \alpha^2 \mathbb{V}\text{ar}[X]$.
- The variance of n random variates is

$$\mathbb{V}\text{ar}[X_1 + X_2 + ... + X_n] = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{C}\text{ov}[X_i, X_j] = \sum_{i=1}^{n} \mathbb{V}\text{ar}[X_i] + 2 \sum_{i>j} \mathbb{C}\text{ov}[X_i, X_j].$$

In the last line we introduced the the covariance operator $\mathbb{C}\text{ov}[..., ...]$.

DEFINITION 21. The **covariance** of two random variables X and Y is defined as

$$\mathbb{C}\text{ov}[X,Y] \equiv \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])], \tag{58}$$

or identically

$$\mathbb{C}\text{ov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]. \tag{59}$$

Properties: Let X, X_i be random variables for $i = 1, 2, 3, 4$ and let $\alpha, \beta, \gamma, \delta$ be constants, then

- $\mathbb{C}\text{ov}[X, \alpha] = 0$.
- $\mathbb{C}\text{ov}[X, X] = \mathbb{V}\text{ar}[X]$.
- $\mathbb{C}\text{ov}[X_1, X_2] = \mathbb{C}\text{ov}[X_2, X_1]$.
- $\mathbb{C}\text{ov}[\alpha X_1, \beta X_2] = \alpha\beta\mathbb{C}\text{ov}[X_1, X_2]$.
- $\mathbb{C}\text{ov}[X_1 + \alpha, X_2 + \beta] = \mathbb{C}\text{ov}[X_1, X_2]$.
- $\mathbb{C}\text{ov}[\alpha X_1 + \beta X_2, \gamma X_3 + \delta X_4] = \alpha\gamma\mathbb{C}\text{ov}[X_1, X_3] + \alpha\delta\mathbb{C}\text{ov}[X_1, X_4] + \beta\gamma\mathbb{C}\text{ov}[X_2, X_3] + \beta\delta\mathbb{C}\text{ov}[X_2, X_4]$.
- If X_1, X_2 are independent then $\mathbb{E}[X_1X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$ and thence $\mathbb{C}\text{ov}[X_1, X_2] = 0$ (the converse is again **false**).

DEFINITION 22. We define the **Pearson product-moment correlation coefficient** between random variables X and Y as

$$\rho_{XY} \equiv \frac{\mathbb{C}\text{ov}[X,Y]}{\sigma(X)\sigma(Y)}. \tag{60}$$

From the Cauchy-Schwarz inequality $|\mathbb{C}\text{ov}[X,Y]|^2 \leq \mathbb{V}\text{ar}[X]\mathbb{V}\text{ar}[Y]$ so $|\rho_{XY}| \leq 1$. Furthermore, clearly $\rho_{XY} = \rho_{YX}$.

2. Of Random Vectors

DEFINITION 23. Suppose we have n random variables, collectively encoded in the vector $\mathbf{X} = (X_1, X_2, ..., X_n)$. We define the **covariance matrix** $\Sigma \in \mathbb{R}^{n\times n}$ of $\mathbf{X}$ as

$$\Sigma \equiv \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\intercal] \tag{61}$$

such that

$$\begin{aligned}\Sigma &\equiv \begin{pmatrix} \mathbb{C}\text{ov}[X_1,X_1] & \mathbb{C}\text{ov}[X_1,X_2] & \cdots & \mathbb{C}\text{ov}[X_1,X_n] \\ \mathbb{C}\text{ov}[X_2,X_1] & \mathbb{C}\text{ov}[X_2,X_2] & \cdots & \mathbb{C}\text{ov}[X_2,X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{C}\text{ov}[X_n,X_1] & \mathbb{C}\text{ov}[X_n,X_2] & \cdots & \mathbb{C}\text{ov}[X_n,X_n] \end{pmatrix} \\ &= \begin{pmatrix} \mathbb{V}\text{ar}[X_1] & \rho_{1,2}\sigma(X_1)\sigma(X_2) & \cdots & \rho_{1,n}\sigma(X_1)\sigma(X_n) \\ \rho_{1,2}\sigma(X_1)\sigma(X_2) & \mathbb{V}\text{ar}[X_2] & \cdots & \rho_{2,n}\sigma(X_2)\sigma(X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1,n}\sigma(X_1)\sigma(X_n) & \rho_{2,n}\sigma(X_2)\sigma(X_n) & \cdots & \mathbb{V}\text{ar}[X_n] \end{pmatrix}\end{aligned} \tag{62}$$

Notationally, $\Sigma \equiv \mathbb{V}\text{ar}[\mathbf{X}] \equiv \mathbb{C}\text{ov}[\mathbf{X}]$. The latter is particularly used when considering the covariance matrix between two different random vectors: $\mathbb{C}\text{ov}[\mathbf{X}, \mathbf{Y}] \equiv \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^\intercal]$.

Properties: Let $\mathbf{X}, \mathbf{X}_i \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^m$ be random vectors. Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times n}$ be constant matrices and $\boldsymbol{k} \in \mathbb{R}^m$ a constant vector.

- $\Sigma = \mathbb{E}[\mathbf{X}\mathbf{X}^\intercal] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}^\intercal]$.
- Symmetry: $\Sigma = \Sigma^\intercal$.
- Positive-semidefiniteness: $\forall \mathbf{w} \in \mathbb{R}^n : \mathbf{w}^\intercal \Sigma \mathbf{w} \geq 0$. This is particularly important since every positive definite matrix is invertible and its inverse is also positive definite, i.e. Σ^{-1} exists and $\forall \mathbf{w} \in \mathbb{R}^n : \mathbf{w}^\intercal \Sigma^{-1} \mathbf{w} \geq 0$.
- $\mathbb{C}\text{ov}[\boldsymbol{A}\mathbf{X} + \boldsymbol{k}] = \boldsymbol{A}\mathbb{C}\text{ov}[\mathbf{X}]\boldsymbol{A}^\intercal$.
- $\mathbb{C}\text{ov}[\mathbf{X}, \mathbf{Y}] = \mathbb{C}\text{ov}[\mathbf{Y}, \mathbf{X}]^\intercal$.
- $\mathbb{C}\text{ov}[\mathbf{X}_1 + \mathbf{X}_2, \mathbf{Y}] = \mathbb{C}\text{ov}[\mathbf{X}_1, \mathbf{Y}] + \mathbb{C}\text{ov}[\mathbf{X}_2, \mathbf{Y}]$.
- $\mathbb{C}\text{ov}[\boldsymbol{A}\mathbf{X}, \boldsymbol{B}^\intercal \mathbf{Y}] = \boldsymbol{A}\mathbb{C}\text{ov}[\mathbf{X}, \mathbf{Y}]\boldsymbol{B}$.

3. Miscellaneous

DEFINITION 24. We define the **conditional variance** of a random variables X given Y as

$$\mathbb{V}\text{ar}[X|Y] \equiv \mathbb{E}[(X - \mathbb{E}[X|Y])^2|Y]. \tag{63}$$

THEOREM 18. The **law of total variance** states that

$$\mathbb{V}\text{ar}[X] = \mathbb{E}_y[\mathbb{V}\text{ar}[X|Y]] + \mathbb{V}\text{ar}_y[\mathbb{E}[X|Y]] \tag{64}$$

PROOF. The proof deploys the law of iterated expectations and the identity $\mathbb{V}\text{ar}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$:

$$\begin{aligned}
\mathbb{V}\text{ar}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \mathbb{E}_y[\mathbb{E}[X^2|Y]] - \mathbb{E}_y[\mathbb{E}[X|Y]]^2 \\
&= \mathbb{E}_y[\mathbb{V}\text{ar}[X|Y] + \mathbb{E}[X|Y]^2] - \mathbb{E}_y[\mathbb{E}[X|Y]]^2 \\
&= \mathbb{E}_y[\mathbb{V}\text{ar}[X|Y]] + \left(\mathbb{E}_y[\mathbb{E}[X|Y]]^2 - \mathbb{E}_y[\mathbb{E}[X|Y]]^2\right) \\
&= \mathbb{E}_y[\mathbb{V}\text{ar}[X|Y]] + \mathbb{V}\text{ar}_y[\mathbb{E}[X|Y]].
\end{aligned}$$

□

DEFINITION 25. Similarly we define the **conditional covariance** of a random variables X and Y given Z as

$$\mathbb{C}\text{ov}[X, Y|Z] \equiv \mathbb{E}[(X - \mathbb{E}[X|Z])(Y - \mathbb{E}[Y|Z])|Z]. \tag{65}$$

THEOREM 19. And the **law of total covariance** states that

$$\mathbb{C}\text{ov}[X, Y] = \mathbb{E}_z[\mathbb{C}\text{ov}[X, Y|Z]] + \mathbb{C}\text{ov}_z[\mathbb{E}[X|Z], \mathbb{E}[Y|Z]]. \tag{66}$$

CHAPTER 7

Martingales and Wiener Processes

1. Definitions and Properties

DEFINITION 26. Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A **stochastic process** is a collection of real valued random variables $\{X_t\}_{t\in T}$ on Ω indexed by a totally ordered set T, which we denote *time*. In discrete time one typically considers the time ordering $T : t = 0 < t = 1 < t = 2 < ...$ such that the stochastic process becomes $\{X_t\}_{t\in T} = \{X_0, X_1, X_2, ...\}$. In continuous time, the corresponding result is $T : t \geq 0$: $\{X_t\}_{t\in T} = \{X_t\}_{t\geq 0}$.

DEFINITION 27. A **filtered probability space** $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ is a probability space with a **filtration** $\{\mathcal{F}_t\}_{t\geq 0}$ i.e. an increasing sequence of sigma algebras. In particular $\forall t : \mathcal{F}_t \subseteq \mathcal{F}$ and $t \leq s \Rightarrow \mathcal{F}_t \subseteq \mathcal{F}_s$.

Remark. Think of the filtration as an increasing flow of information. If we e.g. toss a coin twice then initially (at, say, $t = 0$ -before the first toss) we have very little information to go by, and the sigma algebra is just $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Then, after the first toss (at $t = 1$) we at least know whether the first outcome is a head or tail, and our sigma algebra is $\mathcal{F}_1 = \{\emptyset, \Omega, HH \cup HT, TT \cup TH\}$. Finally, after the last coin toss (at $t = 2$) we have full information and the sigma algebra $\mathcal{F}_2$ will be of the form (1).

DEFINITION 28. A stochastic process $\{X_t\}$ is **adapted** to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$ if $X_t \in \mathcal{F}_t \; \forall t \geq 0$.

DEFINITION 29. A stochastic process $\{X_t\}$ is called an $\mathcal{F}_t$ **martingale** if the following conditions hold:

- X is adapted to the filtration $\{\mathcal{F}_t\}_{t\geq 0}$.
- For all t, $\mathbb{E}[|X_t|] < \infty$.
- For all t and s with $t \leq s$, the inequality $\mathbb{E}[X_t|\mathcal{F}_s] = X_s$.

DEFINITION 30. A **Wiener process** or **Brownian motion** W_t is a continuous time stochastic process which obeys the following properties:

i The initial element is zero, $W_0 = 0$.
ii The function $t \mapsto W_t$ is almost surely (i.e. with probability one) everywhere continuous.
iii W_t has independent increments, which are normally distributed with mean zero and a variance equal to the temporal difference:

$$W_t - W_s \sim N(0, t - s),$$

for $0 \leq s < t$.

Properties From the definition, the following properties immediately follow:

(1) $W_t = W_t - W_0 \sim N(0, t)$ i.e. the random variable W_t has the unconditional density function

$$f_{W_t}(x) = \frac{1}{\sqrt{2\pi t}} \exp\left(-\frac{x^2}{2t}\right)$$

where $\mathbb{E}[W_t] = 0$ and $\mathbb{V}\text{ar}[W_t] = t$.
(2) If $0 \leq t_1 < t_2 \leq t_3 < t_4$ then $\mathbb{E}[(W_{t_2} - W_{t_1})(W_{t_4} - W_{t_3})] = \mathbb{E}[W_{t_2} - W_{t_1}]\mathbb{E}[W_{t_4} - W_{t_3}]$.
(3) $\mathbb{C}\text{ov}[W_t, W_s] = \min\{t, s\}$.
(4) $\mathbb{C}\text{orr}[W_t, W_s] = \sqrt{\min\{t, s\}/\max\{t, s\}}$.
(5) The Wiener process satisfies the conditions for being a **Martingale** since
   - $\forall t : \mathbb{E}[|W_t|] < \infty$.
   - $\mathbb{E}[W_t|\mathcal{F}_s] = W_s$ where $s \leq t$, where $\mathcal{F}_s$ is the information (**sigma field**) available at time s.
(6) $\mathbb{V}\text{ar}[W_t|\mathcal{F}_s] = t - s$ where $s \leq t$.

Proof:

(1) Follows immediately from combining (i) and (iii).
(2) Is the independence condition given in (iii).
(3) Assume $t > s \geq 0$ and write $W_t = W_t - W_s + W_s$ and $W_s = W_s - W_0$. Then $\mathbb{C}\text{ov}[W_t - W_s + W_s, W_s - W_0] = \mathbb{C}\text{ov}[W_t - W_s, W_s - W_0] + \mathbb{C}\text{ov}[W_s, W_s - W_0] = 0 + \mathbb{V}\text{ar}[W_s] = s$. If we assume $s > t \geq 0$ then the covariance will correspondingly be t.
(4) Follows by the definition,

$$\mathbb{C}\text{orr}[W_t, W_s] \equiv \frac{\mathbb{C}\text{ov}[W_t, W_s]}{\sqrt{\mathbb{V}\text{ar}[W_t]}\sqrt{\mathbb{V}\text{ar}[W_s]}}.$$

where $\mathbb{V}\text{ar}[W_t] = s, \mathbb{V}\text{ar}[W_s] = s$. If $t > s$ then $\mathbb{C}\text{orr}[W_t, W_s] = s/(\sqrt{t \cdot s}) = \sqrt{s/t}$. Conversely, if $s > t$ then the correlation can be shown to be $\sqrt{t/s}$.
(5) The latter bullet point is easy demonstrated as $W_t = W_t - W_s + W_s$. Now W_s is known at time s so $\mathbb{E}[W_s|\mathcal{F}_s] = W_s$. Furthermore $W_t - W_s|\mathcal{F}_s \sim \mathcal{N}(0, t - s)$ as the increment is independent of any information about the Wiener process up to and including time s: i.e. $\mathbb{E}[W_t - W_s|\mathcal{F}_s] = 0$. All in all,

$$\begin{aligned}\mathbb{E}[W_t|\mathcal{F}_s] &= \mathbb{E}[W_t - W_s + W_s|\mathcal{F}_s] \\ &= \mathbb{E}[W_t - W_s|\mathcal{F}_s] + \mathbb{E}[W_s|\mathcal{F}_s] \\ &= W_s.\end{aligned}$$

(6) $\mathbb{V}\text{ar}[W_t|\mathcal{F}_s] = \mathbb{V}\text{ar}[W_t - W_s + W_s|\mathcal{F}_s] = \mathbb{V}\text{ar}[W_t - W_s|\mathcal{F}_s] + \mathbb{V}\text{ar}[W_s|\mathcal{F}_s] + 2\mathbb{C}\text{ov}[W_t - W_s, W_s]$. But by independent increments, the covariance is $\mathbb{C}\text{ov}[W_t - W_s, W_s - W_0] = 0$ so $\mathbb{V}\text{ar}[W_t|\mathcal{F}_s] = (t - s) + W_s \cdot \mathbb{V}\text{ar}[1|\mathcal{F}_s] = t - s$.

2. Examples

Example: Show that W_t^3 is *not* a martingale.

Solution: We must show that for some $s < t$, $\mathbb{E}[W_t^3|\mathcal{F}_s] \neq W_s^3$. Let us write $W_t^3 = (\Delta W + W_s)^3$ where $\Delta W \equiv W_t - W_s$ then

$$\begin{aligned}\mathbb{E}[W_t^3|\mathcal{F}_s] &= \mathbb{E}[(\Delta W + W_s)^3|\mathcal{F}_s] \\ &= \mathbb{E}[(\Delta W)^3 + 3(\Delta W)^2 W_s + 3(\Delta W)W_s^2 + W_s^3|\mathcal{F}_s] \\ &= \mathbb{E}[(\Delta W)^3|\mathcal{F}_s] + 3W_s\mathbb{E}[(\Delta W)^2|\mathcal{F}_s] + 3W_s^2\mathbb{E}[\Delta W|\mathcal{F}_s] + W_s^3 \\ &= 3W_s(t - s) + W_s^3 \neq W_s^3\end{aligned}$$

In the last line we have use that $\Delta W \sim N(0, t-s)$, which implies that $\mathbb{E}[(\Delta W)^p]$ equals 0 for all odd p, and $(t-s)^{p/2}(p-1)!!$ for all even p. Also, notice that $W_t^3 - 3W_t t$ *is* a martingale.

Example: A classical quant interview question goes as follows: suppose $t_n = t_0 + n \cdot \delta t$ and W_t is a Wiener process, where $\Delta W_{ij} \equiv W_{t_j} - W_{t_i}$. What is the probability that *both* W_{t_1} and W_{t_2} are positive?

Solution:

$$\begin{aligned}\mathbb{P}(\{W_{t_1} > 0\} \cap \{W_{t_2} > 0\}) &= \mathbb{P}(\{W_{t_1} > 0\} \cap \{W_{t_2} - W_{t_1} > -W_{t_1}\}) \\ &= \mathbb{P}(\{W_{t_1} - W_{t_0} > 0\} \cap \{W_{t_2} - W_{t_1} > -(W_{t_1} - W_{t_0})\}) \\ &\equiv \mathbb{P}(\{\Delta W_{01} > 0\} \cap \{\Delta W_{12} > -\Delta W_{01}\})\end{aligned}$$

Next, we use **Baye's theorem** which, by definition, yields

$$\begin{aligned}\mathbb{P}(\{\Delta W_{01} > 0\} \cap \{\Delta W_{12} > -\Delta W_{01}\}) &\equiv \mathbb{P}(\Delta W_{01} > 0) \cdot \mathbb{P}(\Delta W_{12} > -\Delta W_{01} | \Delta W_{01} > 0) \\ &= 0.5 \cdot \mathbb{P}(\Delta W_{12} > -\Delta W_{01} | \Delta W_{01} > 0)\end{aligned}$$

since $\Delta W_{01} \sim N(0, \delta t)$ is symmetrically distributed around zero.
Subjecting the $\mathbb{P}$ multiplying 0.5 to the law of total probability, where we have partitioned ΔW_{12} into two sets $\mathbb{R}^+$ and $\mathbb{R}_0^-$:

$$\begin{aligned}\mathbb{P}(\Delta W_{12} > -\Delta W_{01} | \Delta W_{01} > 0) = \sum_{\Delta W_{12}} &\mathbb{P}(\Delta W_{12}) \cdot \\ &\mathbb{P}(\Delta W_{12} > -\Delta W_{01} | \{\Delta W_{01} > 0\} \cap \{\Delta W_{12}\}) \\ = 0.5 \cdot \mathbb{P}(\Delta W_{12} > -\Delta W_{01} | \{\Delta W_{01} > 0\} &\cap \{\Delta W_{12} > 0\}) + \\ &0.5 \cdot \mathbb{P}(\Delta W_{12} > -\Delta W_{01} | \{\Delta W_{01} > 0\} \cap \{\Delta W_{12} \leq 0\}) \\ = 0.5 \cdot 1 + 0.5 \cdot 0.5 \\ = 0.75\end{aligned}$$

whence $\mathbb{P}(\{W_{t_1} > 0\} \cap \{W_{t_2} > 0\}) = 0.5 \cdot 0.75 = \underline{\underline{0.375}}$.

In the second to last line we have used that when $X, Y > 0$ then X is bound to be greater than $-Y$: i.e. $\mathbb{P}(X > -Y) = 1$. Also if X, Y are i.i.d. $N(0, \delta t)$ and $X \leq 0, Y > 0$ then the probability that $X > -Y$ must be 0.5 by symmetry.

CHAPTER 8

Popular Distributions

1. The Bernoulli Distribution

DEFINITION 31. The **Bernoulli distribution** is a discrete distribution having two possible outcomes viz. $x = 1$ (″success″ - with probability $p \in (0, 1)$) and $x = 0$ (″failure″ - with probability $1 - p$). The **PMF** is clearly

$$p_{\text{bnl}}(x) = \mathbb{P}(X = x) = p^x(1 - p)^{1-x}. \tag{67}$$

where $x \in \mathbb{B} = \{0, 1\}$.

THEOREM 20. The mean and the variance of the Bernoulli distribution are

$$\mathbb{E}[X] = p, \qquad \mathbb{V}\text{ar}[X] = p(1 - p). \tag{68}$$

THEOREM 21. The characteristic function of the Bernouli distribution is

$$\varphi_X(t) = 1 + p(e^{it} - 1). \tag{69}$$

2. The Binomial Distribution

DEFINITION 32. The **binomial distribution** gives the discrete probability distribution $p_{\text{bin}}(x)$ of obtaining exactly x successes out of n Bernoulli trials (each of which has a probability p of ″success″). It is readily seen that the binomial distribution must have the **PMF**

$$p_{\text{bin}}(x) = \mathbb{P}(X = x) = \frac{n!}{x!(n - x)!}p^x(1 - p)^{n-x} \tag{70}$$

where $n \in \mathbb{Z}^+$ and $\{x \in \mathbb{N} | x \leq n\}$. The associated **CDF** is unsurprisingly

$$F_{X,\text{bin}}(x) = \mathbb{P}(X \leq x) = \sum_{x=0}^{\lfloor x \rfloor} \frac{n!}{x!(n - x)!}p^x(1 - p)^{n-x} \tag{71}$$

where $\lfloor x \rfloor$ is the floor function, i.e. the integer immediately below x.

THEOREM 22. The mean and the variance of the binomial distribution are

$$\mathbb{E}[X] = np, \qquad \mathbb{V}\text{ar}[X] = np(1 - p). \tag{72}$$

THEOREM 23. The characteristic function of the binomial distribution is

$$\varphi_X(t) = [1 + p(e^{it} - 1)]^n. \tag{73}$$

3. The Poisson Distribution

DEFINITION 33. A **Poisson process** is a discrete stochastic process satisfying the following properties:

(1) The numbers of successes in nonoverlapping intervals are independent for all intervals.
(2) The probability of exactly one success in a sufficiently small interval $\delta t = 1/n$ is $p = \lambda \delta t$, where λ is the probability of one success and n is the number of trials.
(3) The probability of two or more successes in a sufficiently small interval δt is essentially 0.

Given a Poisson process, the probability of obtaining exactly x successes in n trials is given by the $n \to \infty$ limit of a binomial distribution (70). The resulting density function is the so-called Poisson distribution[1].

DEFINITION 34. A discrete stochastic variable X is said to have a **Poisson distribution** with parameter $\lambda > 0$, if for $x = 0, 1, 2, ...$ the **PMF** of X is given by:

$$p_{\text{poi}}(x) = \mathbb{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}. \tag{74}$$

The associated **CDF** is

$$F_{X,\text{poi}}(x) = \mathbb{P}(X \leq x) = e^{-\lambda} \sum_{x=0}^{\lfloor x \rfloor} \frac{\lambda^x}{x!}. \tag{75}$$

PROOF. To prove (74) first, recall the elementary facts that (i) $\lambda = np$ by (72), (ii) $n!/(n-x)! = n \cdot (n-1) \cdot ... \cdot (n-x+1) = n^x + \mathcal{O}(n^{x-1})$, and (iii) by definition of Euler's constant: $e^x = \lim_{n\to\infty}(1 + \frac{x}{n})^n$. Then

$$\begin{aligned}
p_{\text{poi}}(x) &= \lim_{n\to\infty} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \lim_{n\to\infty} \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \lim_{n\to\infty} \frac{n \cdot (n-1) \cdot ... \cdot (n-x+1)}{n^x} \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^{n} \left(1 - \frac{\lambda}{n}\right)^{-x} \\
&= \lim_{n\to\infty} \frac{n^x + \mathcal{O}(n^{x-1})}{n^x} \frac{\lambda^x}{x!} \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{n} \lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-x} \\
&= 1 \cdot \frac{\lambda^x}{x!} \cdot e^{-\lambda} \cdot 1 = \frac{\lambda^x e^{-\lambda}}{x!}
\end{aligned}$$

[1] An example helps. Suppose we are interested in the probability distribution of the number of cars that pass us in a given hour on a highway. Assume further that the mean is known $\mathbb{E}[X] = \lambda$. This experiment is binomially distributed as we may think of partitioning an hour into n sufficiently small subintervals and counting the number of success intervals in which we observe a car. The PDF will then be of the form $p_{\text{bin}}(x) = \frac{n!}{x!(n-x)!} (p)^x (1-p)^{n-x}$ where the probability p can be related by the known quantities λ and n through (72). But what about the magnitude of n? How fine grained should we choose our temporal discretization? If we expect to see no more than one car per minute a natural choice would be $n = 60$. However, if this is an exceptionally busy high way, where we see multiple cars per minute, we might like to measure those seconds where we observe cars (so $n = 60 \cdot 60 = 3600$). It turns out that as we push n towards infinity, our PDF (70) converges to the simple (n-independent) expression, the Poisson distribution (74), which is much easier to work with.

□

THEOREM 24. The parameter $\lambda \in \mathbb{R}^+$ can be shown to equal the mean and the variance of a Poisson distribution

$$\mathbb{E}[X] = \mathbb{V}\text{ar}[X] = \lambda. \tag{76}$$

THEOREM 25. The characteristic function of the Poisson distribution is

$$\varphi_X(t) = \exp(\lambda(e^{it} - 1)). \tag{77}$$

4. The Normal Distribution

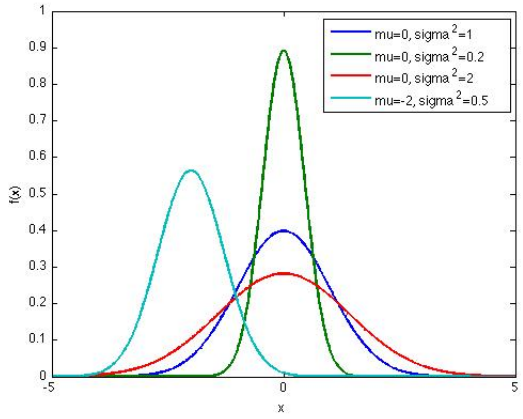

FIGURE 1. The normal density $f(x) = \exp(-(x-\mu)^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$ for various values of mean μ and variance σ^2. Notice that the curves are symmetric around μ.

DEFINITION 35. A random variable X is said to be **normally distributed** with mean μ and variance σ^2, $X \sim N(\mu, \sigma^2)$, if X has the probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{78}$$

and the random variablee Z is said to be **standard normally distributed** with mean 0 and variance 1, $Z \sim N(0,1)$, if Z has the probability density function

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right). \tag{79}$$

The associated cumulative distribution functions are labelled

$$F_X(w) = \int_{-\infty}^{w} f(x)dx, \qquad \text{and} \qquad \Phi(w) = \int_{-\infty}^{w} \phi(z)dz,$$

and they cannot be solved analytically (only numerically). However, one may sometimes see them expressed in terms of the (equally non-analytic) error function, $\text{erf}(w) \equiv \frac{1}{\sqrt{\pi}} \int_{-\infty}^{+\infty} e^{-x^2/2} dx$, such that, inter alia,

$$\Phi(w) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{w}{\sqrt{2}}\right)\right).$$

THEOREM 26. The characteristic function of the normal distribution is

$$\varphi_X(t) = \exp\left(i\mu t - \tfrac{1}{2}\sigma^2 t^2\right). \tag{80}$$

Symmetry: Notice that $f(x)$ is symmetric around μ (and thus, $\phi(z)$ is symmetric around 0). Another way of stating this result for the standard normal is

$$\Phi(z) = 1 - \Phi(-z).$$

Standardization: a normal random variablee X can be transformed into standard normal form $z = \frac{x-\mu}{\sigma}$, such that

$$F_X(w) = \Phi\left(\frac{w-\mu}{\sigma}\right).$$

The importance of this result is hard to over-estimate considering that normal statistical tables are given for the standard normal random variate only. E.g. suppose $X \sim N(\mu, \sigma^2)$ and we are interested in the threshold x s.t. $\mathbb{P}(X > x) = 0.05$. From the standard normal table we can instantly read of the z which satisfies $\mathbb{P}(Z > z) = 0.05$ viz. $z =$ "$\Phi^{-1}(0.95)$". Hence, $x = \mu + \sigma z$.

Properties:
Interesting properties of the normal distribution include

- Let α, β be constants and $X \sim N(\mu, \sigma^2)$ then $\alpha X + \beta \sim N(\alpha\mu + \beta, \alpha^2\sigma^2)$.
- Suppose X_1 and X_2 are *independent* random variates $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ then

$$Z \equiv \alpha X_1 + \beta X_2 \sim N(\alpha\mu_1 + \beta\mu_2, \alpha^2\sigma_1^2 + \beta^2\sigma_2^2). \tag{81}$$

  Use the convolution theorem or characteristic functions to prove this result.
- The central moment of $X \sim N(\mu, \sigma^2)$ is

$$\mathbb{E}[(X-\mu)^p] = \begin{cases} 0, & p \text{ is odd} \\ \sigma^p (p-1)!!, & p \text{ is even} \end{cases} \tag{82}$$

  where !! is the double factorial (which runs in decrements of two): e.g. $7!! = 7 \cdot 5 \cdot 3 \cdot 1$.

Remark. Proving the latter assumes some familiarity with the *gamma function* and should perhaps ideally just be memorized. However, it remains an important result, as it is not unusual to compute things like $\mathbb{E}[\epsilon_t^4]$, where $\epsilon_t \sim N(0, \sigma^2)$, in time series analysis. Here, the answer would clearly be $3\sigma^4$ according to our formula.

5. Multivariate Normal Distribution

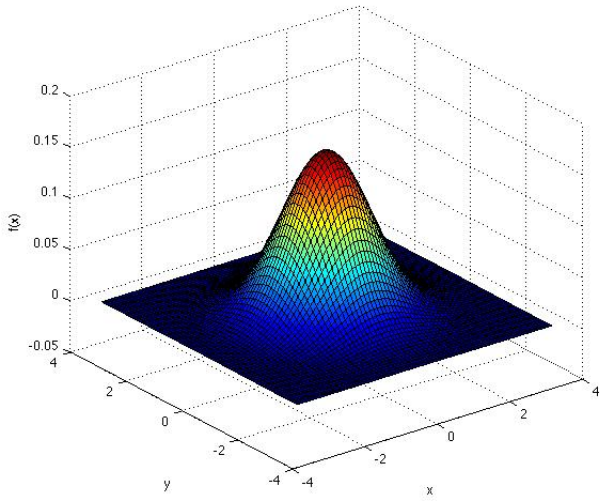

FIGURE 2. The uncorrelated bivariate normal density $f(x,y) = \exp(-(x^2+y^2)/2)/(2\pi)$. Observe that the intersection with the hyper-planes $x = 0$ and $y = 0$ give rise to single variable normal PDFs $f(y)$ and $f(x)$.

DEFINITION 36. Suppose we have a vector $\mathbf{X} = (X_1, X_2, ..., X_n)$ of normal random variables with mean $\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$ and covariance matrix $\boldsymbol{\Sigma}$, i.e $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. In this case the PDF will be

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^{\intercal}\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right) \tag{83}$$

where $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$.

COROLLARY 5. When $n = 1$, $\boldsymbol{\Sigma} = \mathbb{V}\text{ar}[X_1]$ and we recover expression (78).

COROLLARY 6. In the **bivariate** case $n = 2$ where $(X_1, X_2) \equiv (X, Y)$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}$$

where ρ is the correlation coefficient, the density function becomes

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right)$$

Properties

- Observe that an affine transformation $\boldsymbol{A}\mathbf{X} + \boldsymbol{b}$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$ has the distribution
$$\boldsymbol{A}\mathbf{X} + \boldsymbol{b} \sim N(\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^{\intercal})$$
which follows immediately from the properties outlined in Unconditional Expectations and Variance/Covariance.

- If all X_is are independent, the PDF reduces to the product of n normaly distributed PDFs as desired.

THEOREM 27. Suppose we have n normal random variables as before $\mathbf{X} = (X_1, X_2, ..., X_n) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ which we partition into two vectors $\mathbf{X}_1 = (X_1, X_2, ..., X_q) \in \mathbb{R}^q$ and likewise $\mathbf{X}_2 = (X_{q+1}, X_{q+2}, ..., X_n) \in \mathbb{R}^{n-q}$. The corresponding partitioning of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ will be

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{1,1} & \boldsymbol{\Sigma}_{1,2} \\ \boldsymbol{\Sigma}_{2,1} & \boldsymbol{\Sigma}_{2,2} \end{pmatrix}$$

where $\boldsymbol{\mu}_1 \in \mathbb{R}^q, \boldsymbol{\mu}_2 \in \mathbb{R}^{n-q}, \boldsymbol{\Sigma}_{1,1} = \mathbb{R}^{q\times q}, \boldsymbol{\Sigma}_{1,2} \in \mathbb{R}^{q\times(n-q)}, \boldsymbol{\Sigma}_{2,1} \in \mathbb{R}^{(n-q)\times q}$ and $\boldsymbol{\Sigma}_{2,2} \in \mathbb{R}^{(n-q)\times(n-q)}$. Furthermore, suppose $\mathbf{X}_2$ has yielded the following observations $\mathbf{x}_2$ i.e. $\mathbf{X}_2 = (X_{q+1} = x_{q+1}, X_{q+2} = x_{q+2}, ..., X_n = x_n)$. The **conditional normal expectation** of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$ is

$$\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2 \sim N(\boldsymbol{\mu}_{\mathbf{X}_1|\mathbf{X}_2=\mathbf{x}_2}, \boldsymbol{\Sigma}_{\mathbf{X}_1|\mathbf{X}_2=\mathbf{x}_2}) \tag{84}$$

$$\boldsymbol{\mu}_{\mathbf{X}_1|\mathbf{X}_2=\mathbf{x}_2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{1,2}\boldsymbol{\Sigma}_{2,2}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \qquad \boldsymbol{\Sigma}_{\mathbf{X}_1|\mathbf{X}_2=\mathbf{x}_2} = \boldsymbol{\Sigma}_{1,1} - \boldsymbol{\Sigma}_{1,2}\boldsymbol{\Sigma}_{2,2}^{-1}\boldsymbol{\Sigma}_{2,1}.$$

PROOF. For a proof of this theorem see
http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html. □

COROLLARY 7. For the single variable cases $\mathbf{X}_1 = X \in \mathbb{R}$ and $\mathbf{X} = Y \in \mathbb{R}$ (84) reduces to

$$X|Y = y \sim N\left(\mu_x + \frac{\sigma_x}{\sigma_y}\rho(y - \mu_y), (1 - \rho^2)\sigma_x^2\right). \tag{85}$$

6. Log-normal Distribution

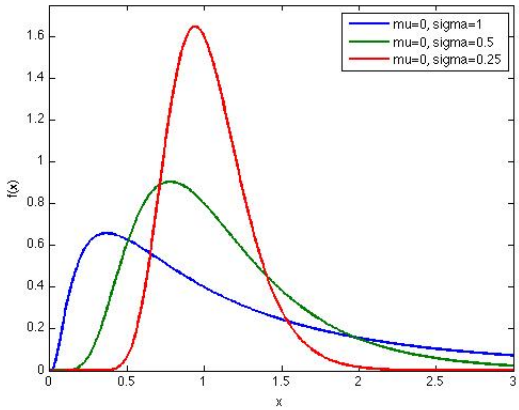

FIGURE 3. The log-normal density $f(x) = \exp(-(\ln(x) - \mu)^2/(2\sigma^2))/(x\sqrt{2\pi\sigma^2})$ for various values of mean μ and variance σ^2. Unlike the normal distribution, there is no axis of symmetry here.

DEFINITION 37. Suppose $X \sim N(\mu, \sigma^2)$ then $Y = e^X$ is **log-normally distributed**, $Y \sim \ln N(\mu, \sigma^2)$, with PDF

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma y}\exp\left(-\frac{(\ln y - \mu)^2}{2\sigma^2}\right) \tag{86}$$

which follows immediately from (25).

THEOREM 28. If $X \sim N(\mu, \sigma^2)$ and $Y = e^X$ then

$$\mathbb{E}[Y] = e^{\mu+\frac{1}{2}\sigma^2}, \qquad \mathbb{V}\text{ar}[Y] = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}. \tag{87}$$

PROOF. We'll restrict our attention to the expectation. Using (38) and the fact that X is normal

$$\begin{aligned}
\mathbb{E}[e^X] &= \int_{-\infty}^{+\infty} e^x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_{-\infty}^{+\infty} e^{\mu+\sigma z} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
&= e^{\mu+\frac{1}{2}\sigma^2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-\sigma)^2} dz = e^{\mu+\frac{1}{2}\sigma^2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\tilde{z}^2} d\tilde{z} \\
&= e^{\mu+\frac{1}{2}\sigma^2}
\end{aligned}$$

where we have used the fact the PDF integrates to unity over the interval $(-\infty, +\infty)$. □

THEOREM 29. We can generalize these results to conditional expectations such that

$$\mathbb{E}[Y|Z] = e^{\mathbb{E}[X|Z]+\frac{1}{2}\mathbb{V}\text{ar}[X|Z]}, \qquad \mathbb{V}\text{ar}[Y|Z] = (e^{\mathbb{V}\text{ar}[X|Z]} - 1)e^{2\mathbb{E}[X|Z]+\mathbb{V}\text{ar}[X|Z]}. \tag{88}$$

.

THEOREM 30. Let $X \sim N(\mu, \sigma^2)$ and $Y = e^X$ then

$$\mathbb{E}[\max\{Y - K, 0\}] = \mathbb{E}[Y]\Phi\left(\frac{\mu - \ln K}{\sigma} + \sigma\right) - K\Phi\left(\frac{\mu - \ln K}{\sigma}\right) \tag{89}$$

where K is a positive constant. This is possibly one of the most important theorems used in the pricing of financial derivatives.

PROOF. Let us introduce the indicator function $\mathbf{1}_{\{e^X > K\}}$ which is 1 iff $e^X > K \Leftrightarrow X > \ln K$ and 0 otherwise.

$$\begin{aligned}
\mathbb{E}[\max\{e^X - K, 0\}] &= \mathbb{E}[(e^X - K)\mathbf{1}_{\{e^X > K\}}] \\
&= \mathbb{E}[e^X \mathbf{1}_{\{e^X > K\}}] - K\mathbb{E}[\mathbf{1}_{\{e^X > K\}}] \\
&= \int_{\ln K}^{+\infty} e^x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx - K\int_{\ln K}^{+\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\
&= \int_{\frac{\ln K-\mu}{\sigma}}^{+\infty} e^{\mu+\sigma z} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz - K\int_{\frac{\ln K-\mu}{\sigma}}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz
\end{aligned}$$

$$
\begin{aligned}
&= e^{\mu+\frac{1}{2}\sigma^2}\int_{\frac{\ln K-\mu}{\sigma}}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-\sigma)^2}dz - K\left\{1-\Phi\left(\tfrac{\ln K-\mu}{\sigma}\right)\right\}\\
&= \mathbb{E}[Y]\int_{\frac{\ln K-\mu}{\sigma}-\sigma}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\tilde{z}^2}d\tilde{z} - K\Phi\left(\tfrac{\mu-\ln K}{\sigma}\right)\\
&= \mathbb{E}[Y]\left\{1-\Phi\left(\tfrac{\ln K-\mu}{\sigma}-\sigma\right)\right\} - K\Phi\left(\tfrac{\mu-\ln K}{\sigma}\right)\\
&= \mathbb{E}[Y]\Phi\left(\tfrac{\mu-\ln K}{\sigma}+\sigma\right) - K\Phi\left(\tfrac{\mu-\ln K}{\sigma}\right).
\end{aligned}
$$

□

COROLLARY 8. Let $X \sim N(\mu, \sigma^2)$ and $Y = e^X$ then

$$\mathbb{E}[\max\{K-Y,0\}] = K\Phi\left(-\frac{\mu-\ln K}{\sigma}\right) - \mathbb{E}[Y]\Phi\left(-\frac{\mu-\ln K}{\sigma}-\sigma\right) \tag{90}$$

where K is a positive constant.

PROOF. Follows from the identity

$$\mathbb{E}[\max\{Y-K,0\}] - \mathbb{E}[\max\{K-Y,0\}] \equiv \mathbb{E}[Y] - K.$$

□