Chapter 1

Probability, Random Variables and Expectations

Note: The primary reference for these notes is Mittelhammer (1999). Other treatments of probability theory include Gallant (1997), Casella and Berger (2001) and Grimmett and Stirzaker (2001).

This chapter provides an overview of probability theory as it applied to both discrete and continuous random variables. The material covered in this chapter serves as a foundation of the econometric sequence and is useful throughout financial economics. The chapter begins with a discussion of the axiomatic foundations of probability theory and then proceeds to describe properties of univariate random variables. Attention then turns to multivariate random variables and important difference from univariate random variables. Finally, the chapter discusses the expectations operator and moments.

1.1 Axiomatic Probability

Probability theory is derived from a small set of axioms – a minimal set of essential assumptions. A deep understanding of axiomatic probability theory is *not* essential to financial econometrics or to the use of probability and statistics in general, although understanding these core concepts does provide additional insight.

The first concept in probability theory is the sample space, which is an abstract concept containing primitive probability events.

Definition 1.1 (Sample Space)**.** The sample space is a set, Ω, that contains all possible outcomes.

Example 1.1. Suppose interest is on a standard 6-sided die. The sample space is 1-dot, 2-dots, ..., 6-dots.

Example 1.2. Suppose interest is in a standard 52-card deck. The sample space is then $A\clubsuit$, $2\clubsuit$, $3\clubsuit$, ..., $J\clubsuit$, $Q\clubsuit$, $K\clubsuit$, $A\diamondsuit$, ..., $K\diamondsuit$, $A\heartsuit$, ..., $K\heartsuit$, $A\spadesuit$, ..., $K\spadesuit$.

Example 1.3. Suppose interest is in the logarithmic stock return, defined as $r_t = \ln P_t - \ln P_{t-1}$, then the sample space is $\mathbb{R}$, the real line.

The next item of interest is an event.

Definition 1.2 (Event)**.** An event, ω, is a subset of the sample space Ω.

An event may be any subsets of the sample space Ω (including the entire sample space), and the set of all events is known as the event space.

Definition 1.3 (Event Space)**.** The set of all events in the sample space Ω is called the event space, and is denoted $\mathcal{F}$.

Event spaces are a somewhat more difficult concept. For finite event spaces, the event space is usually the power set of the outcomes – that is, the set of all possible unique sets that can be constructed from the elements. When variables can take infinitely many outcomes, then a more nuanced definition is needed, although the main idea is to define the event space to be all non-empty intervals (so that each interval has infinitely many points in it).

Example 1.4. Suppose interest lies in the outcome of a coin flip. Then the sample space is $\{H, T\}$ and the event space is $\{\emptyset, \{H\}, \{T\}, \{H, T\}\}$ where $\emptyset$ is the empty set.

The first two axioms of probability are simple: all probabilities must be non-negative and the total probability of all events is one.

Axiom 1.1. *For any event $\omega \in \mathcal{F}$,*

$$\Pr(\omega) \geq 0. \tag{1.1}$$

Axiom 1.2. *The probability of all events in the sample space Ω is unity, i.e.*

$$\Pr(\Omega) = 1. \tag{1.2}$$

The second axiom is a normalization that states that the probability of the entire sample space is 1 and ensures that the sample space must contain all events that may occur. $\Pr(\cdot)$ is a set-valued function – that is, $\Pr(\omega)$ returns the probability, a number between 0 and 1, of observing an event ω.

Before proceeding, it is useful to refresh four concepts from set theory.

Definition 1.4 (Set Union)**.** Let A and B be two sets, then the union is defined

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

A union of two sets contains all elements that are in either set.

Definition 1.5 (Set Intersection)**.** Let A and B be two sets, then the intersection is defined

$$A \cap B = \{x : x \in A \text{ and } x \in B\}.$$

The intersection contains only the elements that are in both sets.

Definition 1.6 (Set Complement)**.** Let A be a set, then the complement set, denoted

$$A^c = \{x : x \notin A\}.$$

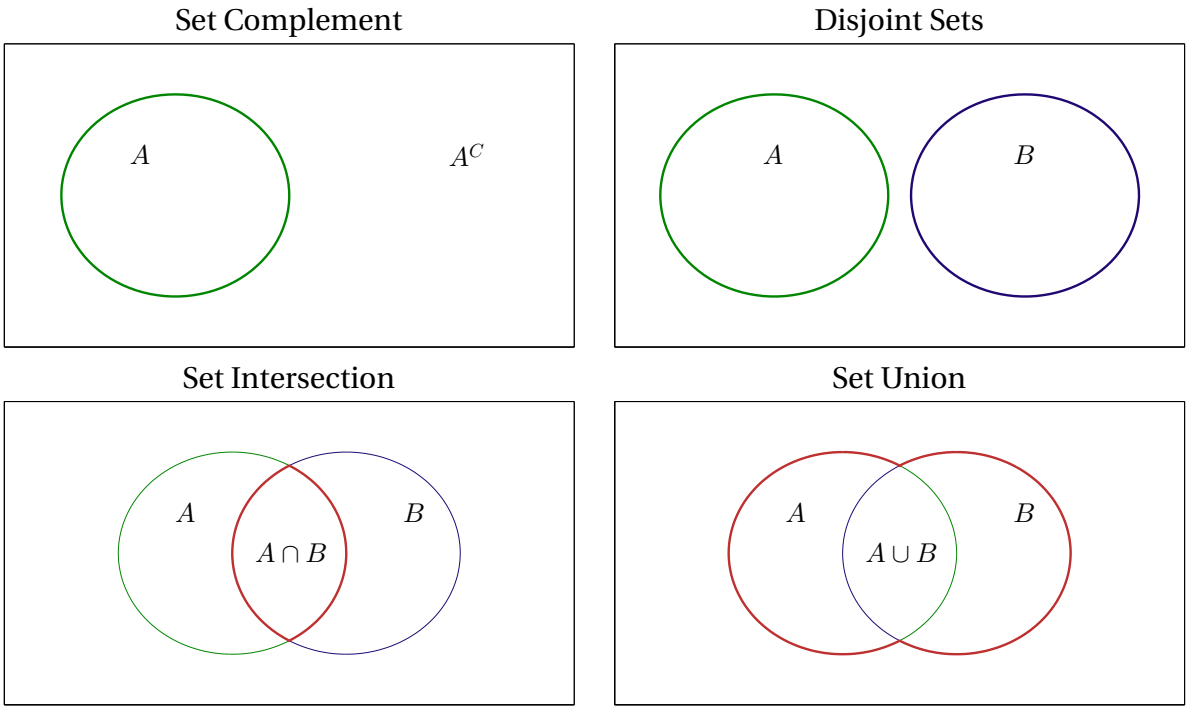


Figure 1.1: The four set definitions presented in $\mathbb{R}^2$. The upper left panel shows a set and its complement. The upper right shows two disjoint sets. The lower left shows the intersection of two sets (darkened region) and the lower right shows the union of two sets (darkened region). In all diagrams, the outer box represents the entire space.

The complement of a set contains all elements which are not contained in the set.

Definition 1.7 (Disjoint Sets)**.** Let A and B be sets, then A and B are disjoint if and only if $A \cap B = \emptyset$.

Figure 1.1 provides a graphical representation of the four set operations in a 2-dimensional space.

The third and final axiom states that probability is additive when sets are disjoint.

Axiom 1.3. *Let* $\{A_i\}$, $i = 1, 2, \ldots$ *be a finite or countably infinite set of disjoint events.*[1] *Then*

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i). \tag{1.3}$$

Assembling a sample space, event space and a probability measure into a set produces what is known as a probability space. Throughout the course, and in virtually all statistics, a complete

[1]

Definition 1.8. A S set is countably infinite if there exists a bijective (one-to-one) function from the elements of S to the natural numbers $\mathbb{N} = \{1, 2, \ldots\}$. Common sets that are countable infinite include the integers ($\mathbb{Z}$) and the rational numbers ($\mathbb{Q}$).

probability space is assumed (typically without explicitly stating this assumption).[2]

Definition 1.9 (Probability Space)**.** A probability space is denoted using the tuple $(\Omega, \mathcal{F}, \Pr)$ where Ω is the sample space, $\mathcal{F}$ is the event space and Pr is the probability set function which has domain $\omega \in \mathcal{F}$.

The three axioms of modern probability are very powerful, and a large number of theorems can be proven using only these axioms. A few simple example are provided, and selected proofs appear in the Appendix.

Theorem 1.1. *Let A be an event in the sample space Ω, and let A^c be the complement of A so that $\Omega = A \cup A^c$. Then $\Pr(A) = 1 - \Pr(A^c)$.*

Since A and A^c are disjoint, and by definition A^c is everything not in A, then the probability of the two must be unity.

Theorem 1.2. *Let A and B be events in the sample space Ω. Then $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.*

This theorem shows that for any two sets, the probability of the union of the two sets is equal to the probability of the two sets minus the probability of the intersection of the sets.

1.1.1 Conditional Probability

Conditional probability extends the basic concepts of probability to the case where interest lies in the probability of one event conditional on the occurrence of another event.

Definition 1.10 (Conditional Probability)**.** Let A and B be two events in the sample space Ω. If $\Pr(B) \neq 0$, then the conditional probability of the event A, given event B, is given by

$$\Pr\left(A|B\right) = \frac{\Pr(A \cap B)}{\Pr(B)}. \tag{1.4}$$

The definition of conditional probability is intuitive. The probability of observing an event in set A, given an event in the set B has occurred, is the probability of observing an event in the intersection of the two sets normalized by the probability of observing an event in set B.

Example 1.5. In the example of rolling a die, suppose $A = \{1, 3, 5\}$ is the event that the outcome is odd and $B = \{1, 2, 3\}$ is the event that the outcome of the roll is less than 4. Then the conditional probability of A given B is

$$\frac{\Pr\left(\{1, 3\}\right)}{\Pr\left(\{1, 2, 3\}\right)} = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3}$$

since the intersection of A and B is $\{1, 3\}$.

The axioms can be restated in terms of conditional probability, where the sample space consists of the events in the set B.

[2] A complete probability space is complete if and only if $B \in \mathcal{F}$ where $\Pr(B) = 0$ and $A \subset B$, then $A \in \mathcal{F}$. This condition ensures that probability can be assigned to any event.

1.1.2 Independence

Independence of two measurable sets means that any information about an event occurring in one set has no information about whether an event occurs in another set.

Definition 1.11. Let A and B be two events in the sample space Ω. Then A and B are independent if and only if

$$\Pr(A \cap B) = \Pr(A)\Pr(B) \tag{1.5}$$

, $A \perp\!\!\!\perp B$ is commonly used to indicate that A and B are independent.

One immediate implication of the definition of independence is that when A and B are independent, then the conditional probability of one given the other is the same as the unconditional probability of the random variable – i.e. $\Pr\left(A|B\right) = \Pr(A)$.

1.1.3 Bayes Rule

Bayes rule is frequently encountered in both statistics (known as Bayesian statistics) and in financial models where agents learn about their environment. Bayes rule follows as a corollary to a theorem that states that the total probability of a set A is equal to the conditional probability of A given a set of disjoint sets B which span the sample space.

Theorem 1.3. *Let $B_i, i = 1, 2\ldots$ be a finite or countably infinite partition of the sample space Ω so that $B_j \cap B_k = \emptyset$ for $j \neq k$ and $\bigcup_{i=1}^{\infty} B_i = \Omega$. Let $\Pr(B_i) > 0$ for all i, then for any set A,*

$$\Pr(A) = \sum_{i=1}^{\infty} \Pr\left(A|B_i\right)\Pr(B_i). \tag{1.6}$$

Bayes rule restates the previous theorem so that the probability of observing an event in B_j given an event in A is observed can be related to the conditional probability of A given B_j.

Corollary 1.1 (Bayes Rule)**.** *Let $B_i, i = 1, 2\ldots$ be a finite or countably infinite partition of the sample space Ω so that $B_j \cap B_k = \emptyset$ for $j \neq k$ and $\bigcup_{i=1}^{\infty} B_i = \Omega$. Let $\Pr(B_i) > 0$ for all i, then for any set A where $\Pr(A) > 0$,*

$$\begin{aligned}\Pr\left(B_j|A\right) &= \frac{\Pr\left(A|B_j\right)\Pr\left(B_j\right)}{\sum_{i=1}^{\infty}\Pr\left(A|B_i\right)\Pr(B_i)}. \\ &= \frac{\Pr\left(A|B_j\right)\Pr\left(B_j\right)}{\Pr(A)}\end{aligned}$$

An immediate consequence of the definition of conditional probability is the

$$\Pr(A \cap B) = \Pr\left(A|B\right)\Pr(B),$$

which is referred to as the multiplication rule. Also notice that the order of the two sets is arbitrary, so that the rule can be equivalently stated as $\Pr(A \cap B) = \Pr\left(B|A\right)\Pr(A)$. Combining these two (as long as $\Pr(A) > 0$),

$$\begin{aligned}\Pr\left(A|B\right)\Pr(B) &= \Pr\left(B|A\right)\Pr(A) \\ \Rightarrow \Pr\left(B|A\right) &= \frac{\Pr\left(A|B\right)\Pr(B)}{\Pr(A)}.\end{aligned} \tag{1.7}$$

Example 1.6. Suppose a family has 2 children and one is a boy, and that the probability of having a child of either sex is equal and independent across children. What is the probability that they have 2 boys?

Before learning that one child is a boy, there are 4 equally probable possibilities: $\{B,B\}$, $\{B,G\}$, $\{G,B\}$ and $\{G,G\}$. Using Bayes rule,

$$\begin{aligned}\Pr\left(\{B,B\}\,|B\geq 1\right) &= \frac{\Pr\left(B\geq 1|\,\{B,B\}\right)\times\Pr\left(\{B,B\}\right)}{\sum_{S\in\{\{B,B\},\{B,G\},\{G,B\},\{G,B\}\}}\Pr\left(B\geq 1|S\right)\Pr(S)}\\ &= \frac{1\times\frac{1}{4}}{1\times\frac{1}{4}+1\times\frac{1}{4}+1\times\frac{1}{4}+0\times\frac{1}{4}}\\ &= \frac{1}{3}\end{aligned}$$

so that knowing one child is a boy increases the probability of 2 boys from $\frac{1}{4}$ to $\frac{1}{3}$. Note that

$$\sum_{S\in\{\{B,B\},\{B,G\},\{G,B\},\{G,B\}\}}\Pr\left(B\geq 1|S\right)\Pr(S) \;=\; \Pr(B\geq 1).$$

Example 1.7. The famous Monte Hall *Let's Make a Deal* television program is an example of Bayes rule. Contestants competed for one of three prizes, a large one (e.g. a car) and two uninteresting ones (duds). The prizes were hidden behind doors numbered 1, 2 and 3. Before the contest starts, the contestant has no information about the which door has the large prize, and to the initial probabilities are all $\frac{1}{3}$. During the negotiations with the host, it is revealed that one of the non-selected doors does *not* contain the large prize. The host then gives the contestant the chance to switch from the door initially chosen to the one remaining door. For example, suppose the contestant choose door 1 initially, and that the host revealed that the large prize is not behind door 3. The contestant then has the chance to choose door 2 or to stay with door 1. In this example, B is the event where the contestant chooses the door which hides the large prize, and A is the event that the large prize is not behind door 2.

Initially there are three equally likely outcomes (from the contestant's point of view), where D indicates dud, L indicates the large prize, and the order corresponds to the door number.

$$\{D,D,L\},\{D,L,D\},\{L,D,D\}$$

The contestant has a $\frac{1}{3}$ chance of having the large prize behind door 1. The host will never remove the large prize, and so applying Bayes rule we have

$$\begin{aligned}\Pr\left(L=2|H=3,S=1\right) &= \frac{\Pr\left(H=3|S=1,L=2\right)\times\Pr\left(L=2|S=1\right)}{\sum_{i=1}^{3}\Pr\left(H=3|S=1,L=i\right)\times\Pr\left(L=i|S=1\right)}\\ &= \frac{1\times\frac{1}{3}}{\frac{1}{2}\times\frac{1}{3}+1\times\frac{1}{3}+0\times\frac{1}{3}}\\ &= \frac{\frac{1}{3}}{\frac{1}{2}}\\ &= \frac{2}{3}.\end{aligned}$$

where H is the door the host reveals, S is initial door selected, and L is the door containing the large prize. This shows that the probability the large prize is behind door 2, given that the player initially selected door 1 and the host revealed door 3 can be computed using Bayes rule.

$\Pr\left(H = 3|S = 1, L = 2\right)$ is the probability that the host shows door 3 given the contestant selected door 1 and the large prize is behind door 2, which always happens since the host will never reveal the large prize. $P\left(L = 2|S = 1\right)$ is the probability that the large is in door 2 given the contestant selected door 1, which is $\frac{1}{3}$. $\Pr\left(H = 3|S = 1, L = 1\right)$ is the probability that the host reveals door 3 given that door 1 was selected and contained the large prize, which is $\frac{1}{2}$, and $P\left(H = 3|S = 1, L = 3\right)$ is the probability that the host reveals door 3 given door 3 contains the prize, which never happens.

Bayes rule shows that it is always optimal to switch doors. This is a counter-intuitive result and occurs since the host's action reveals information about the location of the large prize. Essentially, the two doors not selected by the host have combined probability $\frac{2}{3}$ of containing the large prize before the doors are opened – opening the third assigns its probability to the door not opened.

1.2 Univariate Random Variables

Studying the behavior of random variables, and more importantly functions of random variables (i.e. statistics) is essential for both the theory and practice of financial econometrics. This section covers univariate random variables and multivariate random variables are discussed later.

The previous discussion of probability is set based and so includes objects which cannot be described as random variables, which are a limited (but highly useful) sub-class of all objects that can be described using probability theory. The primary characteristic of a random variable is that it takes values on the real line.

Definition 1.12 (Random Variable)**.** Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. If $X : \Omega \to \mathbb{R}$ is a real-valued function have as its domain elements of Ω, then X is called a random variable.

A random variable is essentially a function which takes $\omega \in \Omega$ as an input and returns a value $x \in \mathbb{R}$, where $\mathbb{R}$ is the symbol for the real line. Random variables come in one of three forms: discrete, continuous and mixed. Random variables which mix discrete and continuous distributions are generally less important in financial economics and so here the focus is on discrete and continuous random variables.

Definition 1.13 (Discrete Random Variable)**.** A random variable is called discrete if its range consists of a countable (possibly infinite) number of elements.

While discrete random variables are less useful than continuous random variables, they are still commonly encountered.

Example 1.8. A random variable which takes on values in $\{0, 1\}$ is known as a Bernoulli random variable, and is the simplest non-degenerate random variable (see Section 1.2.3.1).[3] Bernoulli random variables are often used to model "success" or "failure", where success is loosely defined – a large negative return, the existence of a bull market or a corporate default.

[3] A degenerate random variable always takes the same value, and so is not meaningfully random.

The distinguishing characteristic of a discrete random variable is not that it takes only finitely many values, but that the values it takes are distinct in the sense that it is possible to fit small intervals around each point without the overlap.

Example 1.9. Poisson random variables take values in $\{0, 1, 2, 3, \ldots\}$ (an infinite range), and are commonly used to model hazard rates (i.e. the number of occurrences of an event in an interval). They are especially useful in modeling trading activity (see Section 1.2.3.2).

1.2.1 Mass, Density, and Distribution Functions

Discrete random variables are characterized by a probability mass function (pmf) which gives the probability of observing a particular value of the random variable.

Definition 1.14 (Probability Mass Function)**.** The probability mass function, f, for a discrete random variable X is defined as $f(x) = \Pr(x)$ for all $x \in R(X)$, and $f(x) = 0$ for all $x \notin R(X)$ where $R(X)$ is the range of X (i.e. the values for which X is defined).

Example 1.10. The probability mass function of a Bernoulli random variable takes the form

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

where $p \in [0, 1]$ is the probability of success.

Figure 1.2 contains a few examples of Bernoulli pmfs using data from the FTSE 100 and S&P 500 over the period 1984–2012. Both weekly returns, using Friday to Friday prices and monthly returns, using end-of-month prices, were constructed. Log returns were used ($r_t = \ln(P_t/P_{t-1})$) in both examples. Two of the pmfs defined success as the return being positive. The other two define the probability of success as a return larger than -1% (weekly) or larger than -4% (monthly). These show that the probability of a positive return is much larger for monthly horizons than for weekly.

Example 1.11. The probability mass function of a Poisson random variable is

$$f(x; \lambda) = \frac{\lambda^x}{x!} \exp(-\lambda)$$

where $\lambda \in [0, \infty)$ determines the intensity of arrival (the average value of the random variable).

The pmf of the Poisson distribution can be evaluated for every value of $x \geq 0$, which is the support of a Poisson random variable. Figure 1.4 shows empirical distribution tabulated using a histogram for the time elapsed where .1% of the daily volume traded in the S&P 500 tracking ETF SPY on May 31, 2012. This data series is a good candidate for modeling using a Poisson distribution.

Continuous random variables, on the other hand, take a continuum of values – technically an uncountable infinity of values.

Definition 1.15 (Continuous Random Variable)**.** A random variable is called continuous if its range is uncountably infinite and there exists a non-negative-valued function $f(x)$ defined or all $x \in (-\infty, \infty)$ such that for any event $B \subset R(X)$, $\Pr(B) = \int_{x \in B} f(x)\, \mathrm{d}x$ and $f(x) = 0$ for all $x \notin R(X)$ where $R(X)$ is the range of X (i.e. the values for which X is defined).

Figure 1.2: These four charts show examples of Bernoulli random variables using returns on the FTSE 100 and S&P 500. In the top two, a success was defined as a positive return. In the bottom two, a success was a return above -1% (weekly) or -4% (monthly).

The pmf of a discrete random variable is replaced with the probability density function (pdf) for continuous random variables. This change in naming reflects that the probability of a single point of a continuous random variable is 0, although the probability of observing a value inside an arbitrarily small interval in $R(X)$ is not.

Definition 1.16 (Probability Density Function)**.** For a continuous random variable, the function f is called the probability density function (pdf).

Before providing some examples of pdfs, it is useful to characterize the properties that any pdf should have.

Definition 1.17 (Continuous Density Function Characterization)**.** A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is a member of the class of continuous density functions if and only if $f(x) \geq 0$ for all $x \in (-\infty, \infty)$ and $\int_{-\infty}^{\infty} f(x)\, \mathrm{d}x = 1$.

There are two essential properties. First, that the function is non-negative, which follows from the axiomatic definition of probability, and second, that the function integrates to 1, so that

the total probability across $R(X)$ is 1. This may seem like a limitation, but it is only a normalization since any non-negative integrable function can always be normalized to that it integrates to 1.

Example 1.12. A simple continuous random variable can be defined on $[0,1]$ using the probability density function

$$f(x) = 12\left(x - \frac{1}{2}\right)^2$$

and figure 1.3 contains a plot of the pdf.

This simple pdf has peaks near 0 and 1 and a trough at 1/2. More realistic pdfs allow for values in $(-\infty, \infty)$, such as in the density of a normal random variable.

Example 1.13. The pdf of a normal random variable with parameters μ and σ^2 is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{1.8}$$

$N\left(\mu, \sigma^2\right)$ is used as a shorthand notation for a random variable with this pdf. When $\mu = 0$ and $\sigma^2 = 1$, the distribution is known as a standard normal. Figure 1.3 contains a plot of the standard normal pdf along with two other parameterizations.

For large values of x (in the absolute sense), the pdf of a standard normal takes very small values, and peaks at $x = 0$ with a value of 0.3989. The shape of the normal distribution is that of a bell (and is occasionally referred to a bell curve).

A closely related function to the pdf is the cumulative distribution function, which returns the total probability of observing a value of the random variable *less* than its input.

Definition 1.18 (Cumulative Distribution Function)**.** The cumulative distribution function (cdf) for a random variable X is defined as $F(c) = \Pr(x \leq c)$ for all $c \in (-\infty, \infty)$.

Cumulative distribution function is used for both discrete and continuous random variables.

Definition 1.19 (Discrete cdf)**.** When X is a discrete random variable, the cdf is

$$F(x) = \sum_{s \leq x} f(s) \tag{1.9}$$

for $x \in (-\infty, \infty)$.

Example 1.14. The cdf of a Bernoulli is

$$F(x; p) = \begin{cases} 0 & \text{if} \quad x < 0 \\ p & \text{if} \quad 0 \leq x < 1 \\ 1 & \text{if} \quad x \geq 1 \end{cases}.$$

The Bernoulli cdf is simple since it only takes 3 values. The cdf of a Poisson random variable relatively simple since it is defined as sum the probability mass function for all values less than or equal to the function's argument.

Example 1.15. The cdf of a Poisson(λ)random variable is given by

$$F(x;\lambda) = \exp(-\lambda)\sum_{i=0}^{\lfloor x \rfloor}\frac{\lambda^i}{i!},\ x \geq 0.$$

where $\lfloor \cdot \rfloor$ returns the largest integer smaller than the input (the floor operator).

Continuous cdfs operate much like discrete cdfs, only the summation is replaced by an integral since there are a continuum of values possible for X.

Definition 1.20 (Continuous cdf)**.** When X is a continuous random variable, the cdf is

$$F(x) = \int_{-\infty}^{x} f(s)\,\mathrm{d}s \tag{1.10}$$

for $x \in (-\infty, \infty)$.

The integral computes the total area under the pdf starting from $-\infty$ up to x.

Example 1.16. The cdf of the random variable with pdf given by $12(x-1/2)^2$ is

$$F(x) = 4x^3 - 6x^2 + 3x.$$

and figure 1.3 contains a plot of this cdf.

This cdf is the integral of the pdf, and checking shows that $F(0) = 0$, $F(1/2) = 1/2$ (since it is symmetric around 1/2) and $F(1) = 1$, which must be 1 since the random variable is only defined on [0, 1].h

Example 1.17. The cdf of a normally distributed random variable with parameters μ and σ^2 is given by

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{x}\exp\left(-\frac{(s-\mu)^2}{2\sigma^2}\right)\mathrm{d}s. \tag{1.11}$$

Figure 1.3 contains a plot of the standard normal cdf along with two other parameterizations.

In the case of a standard normal random variable, the cdf is not available in closed form, and so when computed using a computer (i.e. in Excel or MATLAB), fast, accurate numeric approximations based on polynomial expansions are used (Abramowitz and Stegun, 1964).

The cdf can be similarly derived from the pdf as long as the cdf is continuously differentiable. At points where the cdf is not continuously differentiable, the pdf is defined to take the value 0.[4]

Theorem 1.4 (Relationship between cdf and pdf)**.** *Let $f(x)$ and $F(x)$ represent the pdf and cdf of a continuous random variable X, respectively. The density function for X can be defined as $f(x) = \frac{\partial F(x)}{\partial x}$ whenever $f(x)$ is continuous and $f(x) = 0$ elsewhere.*

[4]Formally a pdf does not have to exist for a random variable, although a cdf always does. In practice, this is a technical point and distributions which have this property are rarely encountered in financial economics.

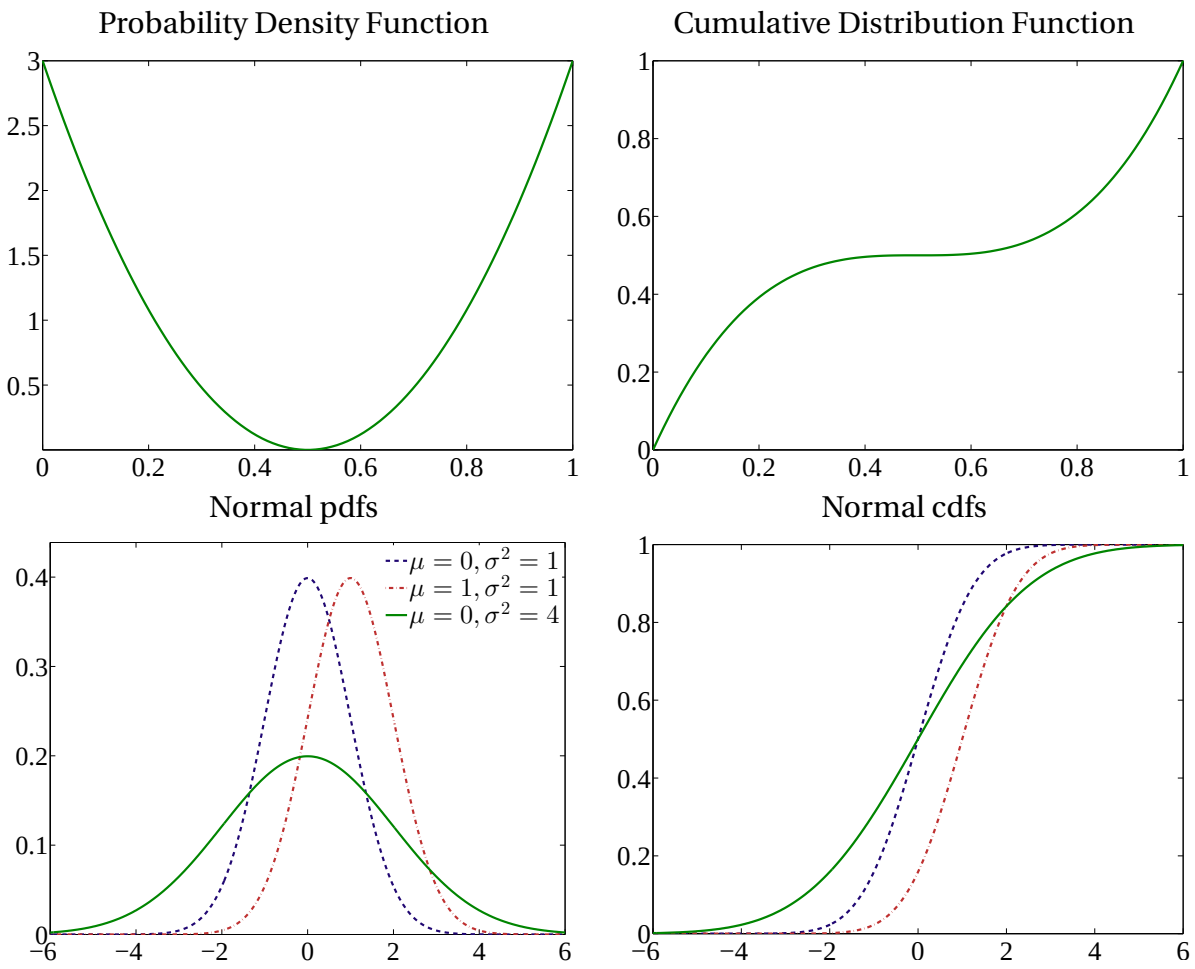


Figure 1.3: The top panels show the pdf for the density $f(x) = 12\left(x - \frac{1}{2}\right)^2$ and its associated cdf. The bottom left panel shows the probability density function for normal distributions with alternative values for μ and σ^2. The bottom right panel shows the cdf for the same parameterizations.

Example 1.18. Taking the derivative of the cdf in the running example,

$$\begin{aligned}\frac{\partial F(x)}{\partial x} &= 12x^2 - 12x + 3 \\ &= 12\left(x^2 - x + \frac{1}{4}\right) \\ &= 12\left(x - \frac{1}{2}\right)^2.\end{aligned}$$

1.2.2 Quantile Functions

The quantile function is closely related to the cdf – and in many important cases, the quantile function is the inverse (function) of the cdf. Before defining quantile functions, it is necessary to

define a quantile.

Definition 1.21 (Quantile)**.** Any number q satisfying $\Pr(x \leq q) = \alpha$ and $\Pr(x \geq q) = 1 - \alpha$ is known as the α-quantile of X and is denoted q_α.

A quantile is just the point on the cdf where the total probability that a random variable is smaller is α and the probability that the random variable takes a larger value is $1 - \alpha$. The definition of a quantile does not necessarily require uniqueness and non-unique quantiles are encountered when pdfs have regions of 0 probability (or equivalently cdfs are discontinuous). Quantiles are unique for random variables which have continuously differentiable cdfs. One common modification of the quantile definition is to select the *smallest* number which satisfies the two conditions to impose uniqueness of the quantile.

The function which returns the quantile is known as the quantile function.

Definition 1.22 (Quantile Function)**.** Let X be a continuous random variable with cdf $F(x)$. The quantile function for X is defined as $G(\alpha) = q$ where $\Pr(x \leq q) = \alpha$ and $\Pr(x > q) = 1 - \alpha$. When $F(x)$ is one-to-one (and hence X is strictly continuous) then $G(\alpha) = F^{-1}(\alpha)$.

Quantile functions are generally set-valued when quantiles are not unique, although in the common case where the pdf does not contain any regions of 0 probability, the quantile function is the inverse of the cdf.

Example 1.19. The cdf of an exponential random variable is

$$F(x;\lambda) = 1 - \exp\left(-\frac{x}{\lambda}\right)$$

for $x \geq 0$ and $\lambda > 0$. Since $f(x;\lambda) > 0$ for $x > 0$, the quantile function is

$$F^{-1}(\alpha;\lambda) = -\lambda \ln(1-\alpha).$$

The quantile function plays an important role in simulation of random variables. In particular, if $u \sim U(0,1)$[5], then $x = F^{-1}(u)$ is distributed F. For example, when u is a standard uniform ($U(0,1)$), and $F^{-1}(\alpha)$ is the quantile function of an exponential random variable with shape parameter λ, then $x = F^{-1}(u;\lambda)$ follows an exponential(λ) distribution.

Theorem 1.5 (Probability Integral Transform)**.** *Let U be a standard uniform random variable, $F_X(x)$ be a continuous, increasing cdf. Then* $\Pr\left(F^{-1}(U) < x\right) = F_X(x)$ *and so $F^{-1}(U)$ is distributed F.*

Proof. Let U be a standard uniform random variable, and for an $x \in R(X)$,

$$\Pr(U \leq F(x)) = F(x),$$

which follows from the definition of a standard uniform.

$$\begin{aligned}\Pr(U \leq F(x)) &= \Pr\left(F^{-1}(U) \leq F^{-1}(F(x))\right)\\ &= \Pr\left(F^{-1}(U) \leq x\right)\\ &= \Pr(X \leq x).\end{aligned}$$

□

[5]The mathematical notation $\sim$ is read "distributed as". For example, $x \sim U(0,1)$ indicates that x is distributed as a standard uniform random variable.

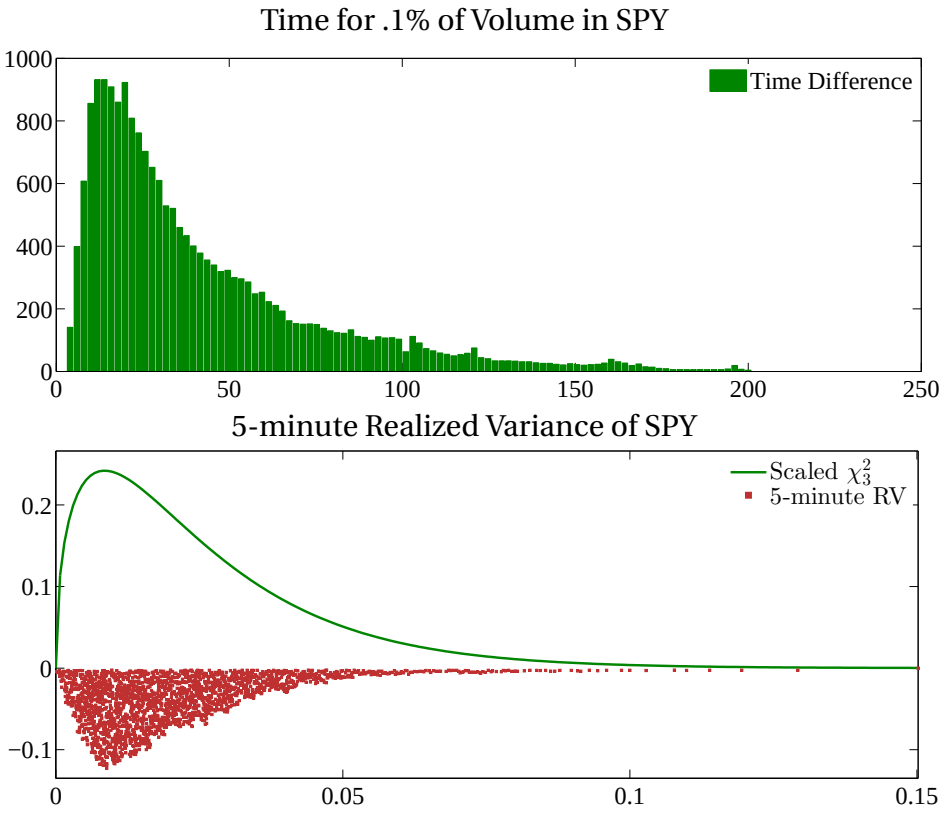


Figure 1.4: The left panel shows a histogram of the elapsed time in seconds required for .1% of the daily volume being traded to occur for SPY on May 31, 2012. The right panel shows both the fitted scaled χ^2 distribution and the raw data (mirrored below) for 5-minute "realized variance" estimates for SPY on May 31, 2012.

The key identity is that $\Pr\left(F^{-1}(U) \leq x\right) = \Pr(X \leq x)$, which shows that the distribution of $F^{-1}(U)$ is F by definition of the cdf. The right panel of figure 1.8 shows the relationship between the cdf of a standard normal and the associated quantile function. Applying $F(X)$ produces a uniform U through the cdf and applying $F^{-1}(U)$ produces X through the quantile function.

1.2.3 Common Univariate Distributions

Discrete

1.2.3.1 Bernoulli

A Bernoulli random variable is a discrete random variable which takes one of two values, 0 or 1. It is often used to model success or failure, where success is loosely defined. For example, a success may be the event that a trade was profitable net of costs, or the event that stock market volatility as measured by VIX was greater than 40%. The Bernoulli distribution depends on a single parameter p which determines the probability of success.

Parameters

$p \in [0, 1]$

Support

$x \in \{0, 1\}$

Probability Mass Function

$f(x; p) = p^x (1 - p)^{1-x}, \; p \geq 0$

Moments

Mean	p
Variance	$p(1 - p)$

1.2.3.2 Poisson

A Poisson random variable is a discrete random variable taking values in $\{0, 1, \ldots\}$. The Poisson depends on a single parameter λ (known as the intensity). Poisson random variables are often used to model counts of events during some interval, for example the number of trades executed over a 5-minute window.

Parameters

$\lambda \geq 0$

Support

$x \in \{0, 1, \ldots\}$

Probability Mass Function

$f(x; \lambda) = \frac{\lambda^x}{x!} \exp(-\lambda)$

Moments

Mean	λ
Variance	λ

Continuous

1.2.3.3 Normal (Gaussian)

The normal is the most important univariate distribution in financial economics. It is the familiar "bell-shaped" distribution, and is used heavily in hypothesis testing and in modeling (net) asset returns (e.g. $r_t = \ln P_t - \ln P_{t-1}$ or $r_t = \frac{P_t - P_{t-1}}{P_{t-1}}$ where P_t is the price of the asset in period t).

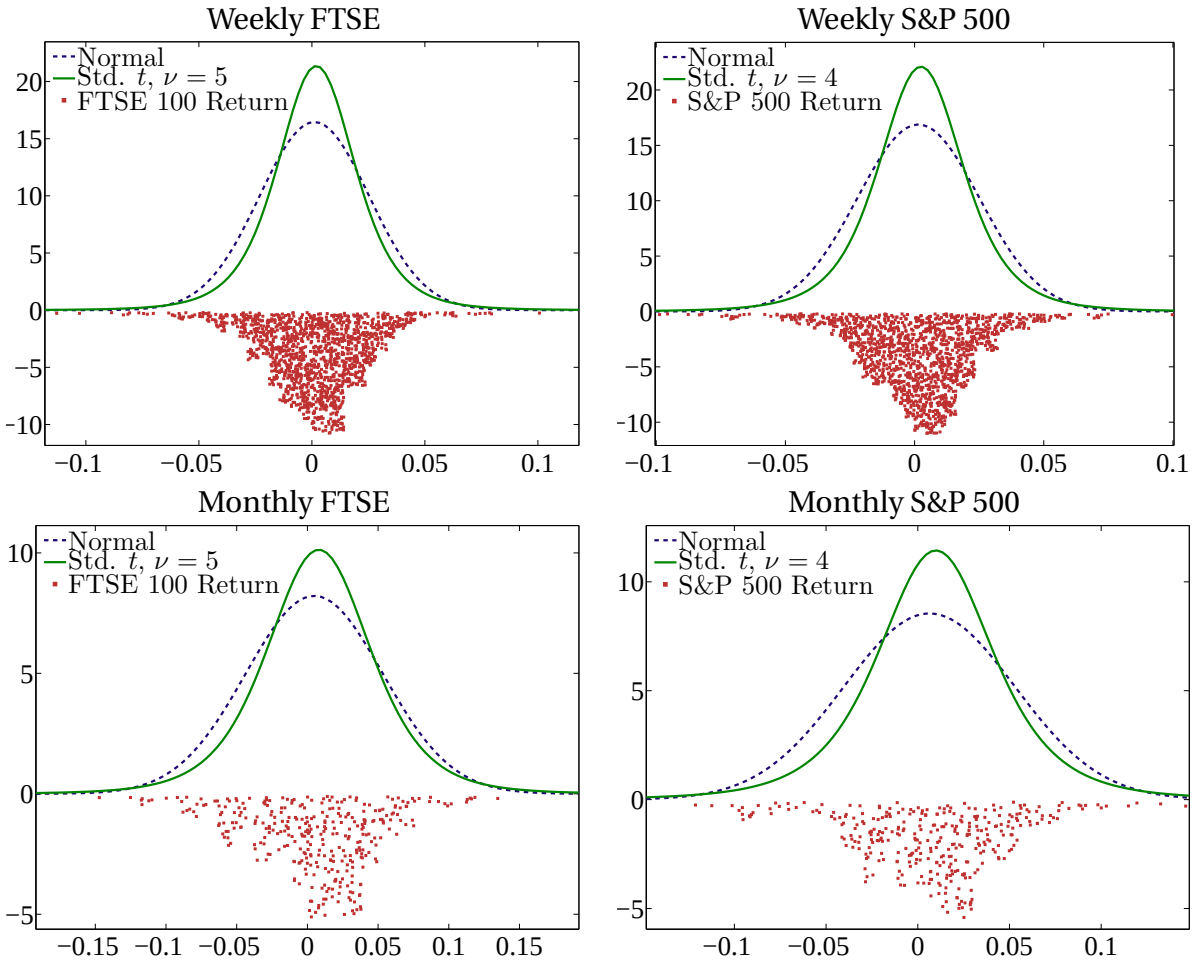


Figure 1.5: Weekly and monthly densities for the FTSE 100 and S&P 500. All panels plot the pdf of a normal and a standardized Student's t using parameters estimated with maximum likelihood estimation (See Chapter1). The points below 0 on the y-axis show the actual returns observed during this period.

Parameters

$\mu \in (-\infty, \infty)$, $\sigma^2 \geq 0$

Support

$x \in (-\infty, \infty)$

Probability Density Function

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

Cumulative Distribution Function

$F\left(x; \mu, \sigma^2\right) = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{1}{\sqrt{2}}\frac{x-\mu}{\sigma}\right)$ where erf is the error function.[6]

Moments

Mean	μ
Variance	σ^2
Median	μ
Skewness	0
Kurtosis	3

Notes

The normal with mean μ and variance σ^2 is written $N\left(\mu, \sigma^2\right)$. A normally distributed random variable with $\mu = 0$ and $\sigma^2 = 1$ is known as a standard normal. Figure 1.5 shows the fit normal distribution to the FTSE 100 and S&P 500 using both weekly and monthly returns for the period 1984–2012. Below each figure is a plot of the raw data.

1.2.3.4 Log-Normal

Log-normal random variables are closely related to normals. If X is log-normal, then $Y = \ln(X)$ is normal. Like the normal, the log-normal family depends on two parameters, μ and σ^2, although unlike the normal these parameters do not correspond to the mean and variance. Log-normal random variables are commonly used to model gross returns, P_{t+1}/P_t (although it is often simpler to model $r_t = \ln P_t - \ln P_{t-1} = \ln(P_t/P_{t-1})$ which is normally distributed).

Parameters

$\mu \in (-\infty, \infty)$, $\sigma^2 \geq 0$

[6]The error function does not have a closed form and is defined

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-s^2\right) \, \text{d}s.$$

Support

$x \in (0, \infty)$

Probability Density Function

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

Cumulative Distribution Function

Since $Y = \ln(X) \sim N\left(\mu, \sigma^2\right)$, the cdf is the same as the normal only using $\ln x$ in place of x.

Moments

Mean	$\exp\left(\mu + \frac{\sigma^2}{2}\right)$
Median	$\exp(\mu)$
Variance	$\left\{\exp\left(\sigma^2\right) - 1\right\} \exp\left(2\mu + \sigma^2\right)$

1.2.3.5 χ^2 (Chi-square)

χ^2_ν random variables depend on a single parameter ν known as the degree-of-freedom. They are commonly encountered when testing hypotheses, although they are also used to model continuous variables which are non-negative such as conditional variances. χ^2_ν random variables are closely related to standard normal random variables and are defined as the sum of ν independent standard normal random variables which have been squared. Suppose $Z_1, \ldots, Z_\nu$ are standard normally distributed and independent, then $x = \sum_{i=1}^{\nu} z_i^2$ follows a χ^2_ν.[7]

Parameters

$\nu \in [0, \infty)$

Support

$x \in [0, \infty)$

Probability Density Function

$f(x; \nu) = \frac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu-2}{2}} \exp\left(-\frac{x}{2}\right)$, $\nu \in \{1, 2, \ldots\}$ where $\Gamma(a)$ is the Gamma function.[8]

Cumulative Distribution Function

$F(x; \nu) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right)} \gamma\left(\frac{\nu}{2}, \frac{x}{2}\right)$ where $\gamma(a, b)$ is the lower incomplete gamma function.

[7] ν does not need to be an integer,

[8] The χ^2_ν is related to the gamma distribution which has pdf $f(x; \alpha, b) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} \exp(-x/\beta)$ by setting $\alpha = \nu/2$ and $\beta = 2$.

Moments

Mean	ν
Variance	2ν

Notes

Figure 1.4 shows a χ^2 pdf which was used to fit some simple estimators of the 5-minute variance of the S&P 500 from May 31, 2012. These were computed by summing and squaring 1-minute returns within a 5-minute interval (all using log prices). 5-minute variance estimators are important in high-frequency trading and other (slower) algorithmic trading.

1.2.3.6 Student's t and standardized Student's t

Student's t random variables are also commonly encountered in hypothesis testing and, like χ^2_ν random variables, are closely related to standard normals. Student's t random variables depend on a single parameter, ν, and can be constructed from two other independent random variables. If Z a standard normal, W a χ^2_ν and $Z \perp\!\!\!\perp W$, then $x = z/\sqrt{\frac{w}{\nu}}$ follows a Student's t distribution. Student's t are similar to normals except that they are heavier tailed, although as $\nu \to \infty$ a Student's t converges to a standard normal.

Support

$x \in (-\infty, \infty)$

Probability Density Function

$f(x;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$ where $\Gamma(a)$ is the Gamma function.

Moments

Mean	$0,\ \nu > 1$
Median	0
Variance	$\frac{\nu}{\nu-2},\ \nu > 2$
Skewness	$0,\ \nu > 3$
Kurtosis	$3\frac{(\nu-2)}{\nu-4},\ \nu > 4$

Notes

When $\nu = 1$, a Student's t is known as a Cauchy random variable. Cauchy random variables are so heavy-tailed that even the mean does not exist.

The standardized Student's t extends the usual Student's t in two directions. First, it removes the variance's dependence on ν so that the scale of the random variable can be established separately from the degree of freedom parameter. Second, it explicitly adds location and scale parameters so that if Y is a Student's t random variable with degree of freedom ν, then

$$x = \mu + \sigma\frac{\sqrt{\nu-2}}{\sqrt{\nu}}y$$

follows a standardized Student's t distribution ($\nu > 2$ is required). The standardized Student's t is commonly used to model heavy-tailed return distributions such as stock market indices.

Figure 1.5 shows the fit (using maximum likelihood) standardized t distribution to the FTSE 100 and S&P 500 using both weekly and monthly returns from the period 1984–2012. The typical degree of freedom parameter was around 4, indicating that (unconditional) distributions are heavy-tailed with a large kurtosis.

1.2.3.7 Uniform

The continuous uniform is commonly encountered in certain test statistics, especially those testing whether assumed densities are appropriate for a particular series. Uniform random variables, when combined with quantile functions, are also useful for simulating random variables.

Parameters

a, b the end points of the interval, where $a < b$

Support

$x \in [a, b]$

Probability Density Function

$f(x) = \frac{1}{b-a}$

Cumulative Distribution Function

$F(x) = \frac{x-a}{b-a}$ for $a \leq x \leq b$, $F(x) = 0$ for $x < a$ and $F(x) = 1$ for $x > b$

Moments

Mean	$\frac{b-a}{2}$
Median	$\frac{b-a}{2}$
Variance	$\frac{(b-a)^2}{12}$
Skewness	0
Kurtosis	$\frac{9}{5}$

Notes

A standard uniform has $a = 0$ and $b = 1$. When $x \sim F$, then $F(x) \sim U(0, 1)$

1.3 Multivariate Random Variables

While univariate random variables are very important in financial economics, most applications require the use multivariate random variables. Multivariate random variables allow the relation-

ship between two or more random quantities to be modeled and studied. For example, the joint distribution of equity and bond returns is important for many investors.

Throughout this section, the multivariate random variable is assumed to have n components,

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

which are arranged into a column vector. The definition of a multivariate random variable is virtually identical to that of a univariate random variable, only mapping $\omega \in \Omega$ to the n-dimensional space $\mathbb{R}^n$.

Definition 1.23 (Multivariate Random Variable)**.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. If $X : \Omega \to \mathbb{R}^n$ is a real-valued vector function having its domain the elements of Ω, then $X : \Omega \to \mathbb{R}^n$ is called a (multivariate) n-dimensional random variable.

Multivariate random variables, like univariate random variables, are technically functions of events in the underlying probability space $X(\omega)$, although the function argument ω (the event) is usually suppressed.

Multivariate random variables can be either discrete or continuous. Discrete multivariate random variables are fairly uncommon in financial economics and so the remainder of the chapter focuses exclusively on the continuous case. The characterization of a what makes a multivariate random variable continuous is also virtually identical to that in the univariate case.

Definition 1.24 (Continuous Multivariate Random Variable)**.** A multivariate random variable is said to be continuous if its range is uncountably infinite and if there exists a non-negative valued function $f(x_1, \ldots, x_n)$ defined for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$ such that for any event $B \subset R(X)$,

$$\Pr(B) = \int \ldots \int_{\{x_1,\ldots,x_n\} \in B} f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \ldots \mathrm{d}x_n \tag{1.12}$$

and $f(x_1, \ldots, x_n) = 0$ for all $(x_1, \ldots, x_n) \notin R(X)$.

Multivariate random variables, at least when continuous, are often described by their probability density function.

Definition 1.25 (Continuous Density Function Characterization)**.** A function $f : \mathbb{R}^n \to \mathbb{R}$ is a member of the class of multivariate continuous density functions if and only if $f(x_1, \ldots, x_n) \geq 0$ for all $x \in \mathbb{R}^n$ and

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, \mathrm{d}x_1 \ldots \mathrm{d}x_n = 1. \tag{1.13}$$

Definition 1.26 (Multivariate Probability Density Function)**.** The function $f(x_1, \ldots, x_n)$ is called a multivariate probability function (pdf).

A multivariate density, like a univariate density, is a function which is everywhere non-negative and which integrates to unity. Figure 1.7 shows the fit joint probability density function to weekly returns on the FTSE 100 and S&P 500 (assuming that returns are normally distributed). Two views are presented – one shows the 3-dimensional plot of the pdf and the other shows the iso-probability contours of the pdf. The figure also contains a scatter plot of the raw weekly data for comparison. All parameters were estimated using maximum likelihood.

Example 1.20. Suppose X is a bivariate random variable, then the function $f(x_1, x_2) = \frac{3}{2}\left(x_1^2 + x_2^2\right)$ defined on $[0, 1] \times [0, 1]$ is a valid probability density function.

Example 1.21. Suppose X is a bivariate standard normal random variable. Then the probability density function of X is

$$f(x_1, x_2) = \frac{1}{2\pi} \exp\left(-\frac{x_1^2 + x_2^2}{2}\right).$$

The multivariate cumulative distribution function is virtually identical to that in the univariate case, and measure the total probability between $-\infty$ (for each element of X) and some point.

Definition 1.27 (Multivariate Cumulative Distribution Function)**.** The joint cumulative distribution function of an n-dimensional random variable X is defined by

$$F(x_1, \ldots, x_n) = \Pr(X_i \leq x_i, i = 1, \ldots, n)$$

for all $(x_1, \ldots, x_n) \in \mathbb{R}^n$, and is given by

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_n} \ldots \int_{-\infty}^{x_1} f(s_1, \ldots, s_n)\, \mathrm{d}s_1 \ldots \mathrm{d}s_n. \tag{1.14}$$

Example 1.22. Suppose X is a bivariate random variable with probability density function

$$f(x_1, x_2) = \frac{3}{2}\left(x_1^2 + x_2^2\right)$$

defined on $[0, 1] \times [0, 1]$. Then the associated cdf is

$$F(x_1, x_2) = \frac{x_1^3 x_2 + x_1 x_2^3}{2}.$$

Figure 1.6 shows the joint cdf of the density in the previous example. As was the case for univariate random variables, the probability density function can be determined by differentiating the cumulative distribution function with respect to each component.

Theorem 1.6 (Relationship between cdf and pdf)**.** *Let $f(x_1, \ldots, x_n)$ and $F(x_1, \ldots, x_n)$ represent the pdf and cdf of an n-dimensional continuous random variable X, respectively. The density function for X can be defined as $f(x_1, \ldots, x_n) = \frac{\partial^n F(\mathbf{x})}{\partial x_1 \partial x_2 \ldots \partial x_n}$ whenever $f(x_1, \ldots, x_n)$ is continuous and $f(x_1, \ldots, x_n) = 0$ elsewhere.*

Example 1.23. Suppose X is a bivariate random variable with cumulative distribution function $F(x_1, x_2) = \frac{x_1^3 x_2 + x_1 x_2^3}{2}$. The probability density function can be determined using

$$\begin{aligned} f(x_1, x_2) &= \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2} \\ &= \frac{1}{2} \frac{\partial \left(3x_1^2 x_2 + x_2^3\right)}{\partial x_2} \\ &= \frac{3}{2}\left(x_1^2 + x_2^2\right). \end{aligned}$$

1.3.1 Marginal Densities and Distributions

The marginal distribution is the first concept unique to multivariate random variables. Marginal densities and distribution functions summarize the information in a subset, usually a single component, of X by averaging over all possible values of the components of X which are not being marginalized. This involves integrating out the variables which are not of interest. First, consider the bivariate case.

Definition 1.28 (Bivariate Marginal Probability Density Function)**.** Let X be a bivariate random variable comprised of X_1 and X_2. The marginal distribution of X_1 is given by

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2)\ \mathrm{d}x_2. \tag{1.15}$$

The marginal density of X_1 is a density function where X_2 has been integrated out. This integration is simply a form of averaging – varying x_2 according to the probability associated with each value of x_2 – and so the marginal is only a function of x_1. Both probability density functions and cumulative distribution functions have marginal versions.

Example 1.24. Suppose X is a bivariate random variable with probability density function

$$f(x_1, x_2) = \frac{3}{2}\left(x_1^2 + x_2^2\right)$$

and is defined on $[0,1] \times [0,1]$. The marginal probability density function for X_1 is

$$f_1(x_1) = \frac{3}{2}\left(x_1^2 + \frac{1}{3}\right),$$

and by symmetry the marginal probability density function of X_2 is

$$f_2(x_2) = \frac{3}{2}\left(x_2^2 + \frac{1}{3}\right).$$

Example 1.25. Suppose X is a bivariate random variable with probability density function $f(x_1, x_2) = 6\left(x_1 x_2^2\right)$ and is defined on $[0,1] \times [0,1]$. The marginal probability density functions for X_1 and X_2 are

$$f_1(x_1) = 2x_1 \text{ and } f_2(x_2) = 3x_2^2.$$

Example 1.26. Suppose X is bivariate normal with parameters $\boldsymbol{\mu} = [\mu_1\ \mu_2]'$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix},$$

then the marginal pdf of X_1 is $N\left(\mu_1, \sigma_1^2\right)$, and the marginal pdf of X_2 is $N\left(\mu_2, \sigma_2^2\right)$.

Figure 1.7 shows the fit marginal distributions to weekly returns on the FTSE 100 and S&P 500 assuming that returns are normally distributed. Marginal pdfs can be transformed into marginal cdfs through integration.

Definition 1.29 (Bivariate Marginal Cumulative Distribution Function)**.** The cumulative marginal distribution function of X_1 in bivariate random variable X is defined by

$$F_1(x_1) = \Pr(X_1 \leq x_1)$$

for all $x_1 \in \mathbb{R}$, and is given by

$$F_1(x_1) = \int_{-\infty}^{x_1} f_1(s_1)\, \mathrm{d}s_1.$$

The general j-dimensional marginal distribution partitions the n-dimensional random variable X into two blocks, and constructs the marginal distribution for the first j by integrating out (averaging over) the remaining $n - j$ components of X. In the definition, both X_1 and X_2 are vectors.

Definition 1.30 (Marginal Probability Density Function)**.** Let X be a n-dimensional random variable and partition the first j ($1 \leq j < n$) elements of X into X_1, and the remainder into X_2 so that $X = \left[X_1' \, X_2'\right]'$. The marginal probability density function for X_1 is given by

$$f_{1,\ldots,j}\left(x_1, \ldots, x_j\right) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, \mathrm{d}x_{j+1} \ldots \mathrm{d}x_n. \tag{1.16}$$

The marginal cumulative distribution function is related to the marginal probability density function in the same manner as the joint probability density function is related to the cumulative distribution function. It also has the same interpretation.

Definition 1.31 (Marginal Cumulative Distribution Function)**.** Let X be a n-dimensional random variable and partition the first j ($1 \leq j < n$) elements of X into X_1, and the remainder into X_2 so that $X = \left[X_1' \, X_2'\right]'$. The marginal cumulative distribution function for X_1 is given by

$$F_{1,\ldots,j}\left(x_1, \ldots, x_j\right) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_j} f_{1,\ldots,j}\left(s_1, \ldots, s_j\right)\, \mathrm{d}s_1 \ldots \mathrm{d}s_j. \tag{1.17}$$

1.3.2 Conditional Distributions

Marginal distributions provide the tools needed to model the distribution of a subset of the components of a random variable while averaging over the other components. Conditional densities and distributions, on the other hand, consider a subset of the components a random variable conditional on observing a specific value for the remaining components. In practice, the vast majority of modeling makes use of conditioning information where the interest is in understanding the distribution of a random variable conditional on the observed values of some other random variables. For example, consider the problem of modeling the expected return of an individual stock. Balance sheet information such as the book value of assets, earnings and return on equity are all available, and can be conditioned on to model the conditional distribution of the stock's return.

First, consider the bivariate case.

Definition 1.32 (Bivariate Conditional Probability Density Function)**.** Let X be a bivariate random variable comprised of X_1 and X_2. The conditional probability density function for X_1 given that $X_2 \in B$ where B is an event where $\Pr(X_2 \in B) > 0$ is

$$f\left(x_1 | X_2 \in B\right) = \frac{\int_B f(x_1, x_2)\, \mathrm{d}x_2}{\int_B f_2(x_2)\, \mathrm{d}x_2}. \tag{1.18}$$

When B is an elementary event (e.g. single point), so that $\Pr(X_2 = x_2) = 0$ and $f_2(x_2) > 0$, then

$$f\left(x_1 | X_2 = x_2\right) = \frac{f(x_1, x_2)}{f_2(x_2)}. \tag{1.19}$$

Conditional density functions differ slightly depending on whether the conditioning variable is restricted to a set or a point. When the conditioning variable is specified to be a set where $\Pr(X_2 \in B) > 0$, then the conditional density is the joint probability of X_1 and $X_2 \in B$ divided by the marginal probability of $X_2 \in B$. When the conditioning variable is restricted to a point, the conditional density is the ratio of the joint pdf to the margin pdf of X_2.

Example 1.27. Suppose X is a bivariate random variable with probability density function

$$f(x_1, x_2) = \frac{3}{2}\left(x_1^2 + x_2^2\right)$$

and is defined on $[0,1] \times [0,1]$. The conditional probability of X_1 given $X_2 \in \left[\frac{1}{2}, 1\right]$

$$f\left(x_1 | X_2 \in \left[\frac{1}{2}, 1\right]\right) = \frac{1}{11}\left(12x_1^2 + 7\right),$$

the conditional probability density function of X_1 given $X_2 \in \left[0, \frac{1}{2}\right]$ is

$$f\left(x_1 | X_2 \in \left[0, \frac{1}{2}\right]\right) = \frac{1}{5}\left(12x_1^2 + 1\right),$$

and the conditional probability density function of X_1 given $X_2 = x_2$ is

$$f\left(x_1 | X_2 = x_2\right) = \frac{x_1^2 + x_2^2}{x_2^2 + 1}.$$

Figure 1.6 shows the joint pdf along with both types of conditional densities. The upper left panel shows that conditional density for $X_2 \in [0.25, 0.5]$. The highlighted region contains the components of the joint pdf which are averaged to produce the conditional density. The lower left also shows the pdf but also shows three (non-normalized) conditional densities of the form $f\left(x_1 | x_2\right)$. The lower right pane shows these three densities correctly normalized.

The previous example shows that, in general, the conditional probability density function differs as the region used changes.

Example 1.28. Suppose X is bivariate normal with mean $\boldsymbol{\mu} = [\mu_1\ \mu_2]'$ and covariance

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix},$$

then the conditional distribution of X_1 given $X_2 = x_2$ is $N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}\right)$.

Marginal distributions and conditional distributions are related in a number of ways. One obvious way is that $f\left(x_1|X_2 \in R(X_2)\right) = f_1(x_1)$ – that is, the conditional probability of X_1 given that X_2 is in its range is the marginal pdf of X_1. This holds since integrating over all values of x_2 is essentially not conditioning on anything (which is known as the unconditional, and a marginal density could, in principle, be called the unconditional density since it averages across all values of the other variable).

The general definition allows for an n-dimensional random vector where the conditioning variable has a dimension between 1 and $j < n$.

Definition 1.33 (Conditional Probability Density Function)**.** Let $f(x_1, \ldots, x_n)$ be the joint density function for an n-dimensional random variable $X = [X_1 \ \ldots X_n]'$ and and partition the first j ($1 \leq j < n$) elements of X into X_1, and the remainder into X_2 so that $X = \left[X_1' \ X_2'\right]'$. The conditional probability density function for X_1 given that $X_2 \in B$ is given by

$$f\left(x_1, \ldots, x_j|X_2 \in B\right) = \frac{\int_{(x_{j+1},\ldots,x_n)\in B} f(x_1, \ldots, x_n)\ \mathrm{d}x_n \ldots \mathrm{d}x_{j+1}}{\int_{(x_{j+1},\ldots,x_n)\in B} f_{j+1,\ldots,n}\left(x_{j+1}, \ldots, x_n\right)\ \mathrm{d}x_n \ldots \mathrm{d}x_{j+1}}, \tag{1.20}$$

and when B is an elementary event (denoted $\mathbf{x}_2$) and if $f_{j+1,\ldots,n}(\mathbf{x}_2) > 0$,

$$f\left(x_1, \ldots, x_j|X_2 = \mathbf{x}_2\right) = \frac{f\left(x_1, \ldots, x_j, \mathbf{x}_2\right)}{f_{j+1,\ldots,n}(\mathbf{x}_2)} \tag{1.21}$$

In general the simplified notation $f\left(x_1, \ldots, x_j|\mathbf{x}_2\right)$ will be used to represent $f\left(x_1, \ldots, x_j|X_2 = \mathbf{x}_2\right)$.

1.3.3 Independence

A special relationship exists between the joint probability density function and the marginal density functions when random variables are independent– the joint must be the product of each marginal.

Theorem 1.7 (Independence of Random Variables)**.** *The random variables $X_1, \ldots, X_n$ with joint density function $f(x_1, \ldots, x_n)$ are independent if and only if*

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i) \tag{1.22}$$

where $f_i(x_i)$ is the marginal distribution of X_i.

The intuition behind this result follows from the fact that when the components of a random variable are independent, any change in one component has no information for the others. In other words, both marginals and conditionals must be the same.

Example 1.29. Let X be a bivariate random variable with probability density function $f(x_1, x_2) = x_1 x_2$ on $[0, 1] \times [0, 1]$, then X_1 and X_2 are independent. This can be verified since

$$f_1(x_1) = x_1 \ \text{ and } \ f_2(x_2) = x_2$$

so that the joint is the product of the two marginal densities.

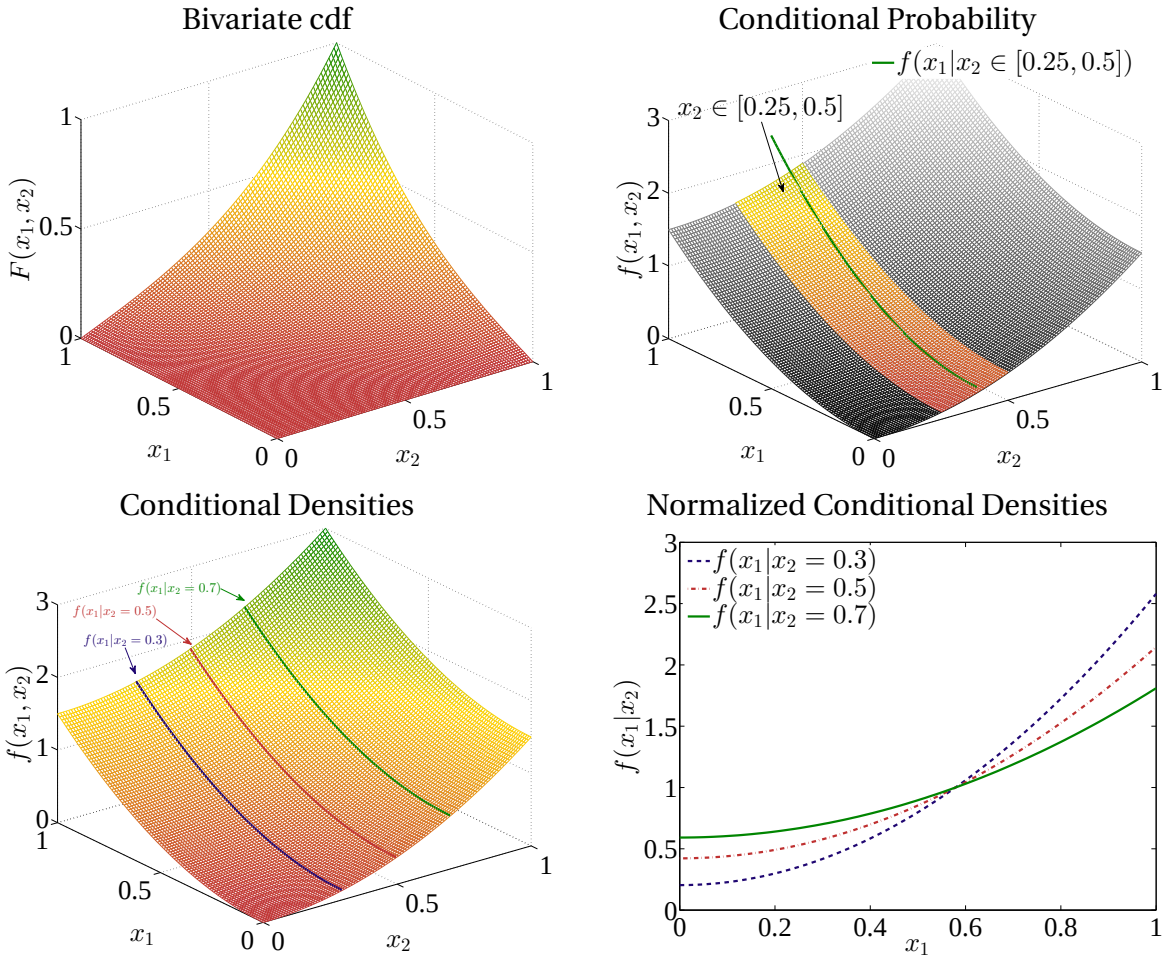


Figure 1.6: These four panels show four views of a distribution defined on $[0,1] \times [0,1]$. The upper left panel shows the joint cdf. The upper right shows the pdf along with the portion of the pdf used to construct a conditional distribution $f\left(x_1|x_2 \in [0.25, 0.5]\right)$. The line shows the actual correctly scaled conditional distribution which is only a function of x_1 plotted at $\mathrm{E}\left[X_2|X_2 \in [0.25, 0.5]\right]$. The lower left panel also shows the pdf along with three non-normalized conditional densities. The bottom right panel shows the correctly normalized conditional densities.

Independence is a very strong concept, and it carries over from random variables to functions of random variables as long as each function involves only one random variable.[9]

Theorem 1.8 (Independence of Functions of Independent Random Variables)**.** *Let X_1 and X_2 be independent random variables and define $y_1 = Y_1(x_1)$ and $y_2 = Y_2(x_2)$, then the random variables Y_1 and Y_2 are independent.*

Independence is often combined with an assumption that the marginal distribution is the same to simplify the analysis of collections of random data.

Definition 1.34 (Independent, Identically Distributed)**.** Let $\{X_i\}$ be a sequence of random variables. If the marginal distribution for X_i is the same for all i and $X_i \perp\!\!\!\perp X_j$ for all $i \neq j$, then $\{X_i\}$ is said to be an independent, identically distributed (i.i.d.) sequence.

1.3.4 Bayes Rule

Bayes rule is used both in financial economics and econometrics. In financial economics, it is often used to model agents learning, and in econometrics it is used to make inference about unknown parameters given observed data (a branch known as Bayesian econometrics). Bayes rule follows directly from the definition of a conditional density so that the joint can be factored into a conditional and a marginal. Suppose X is a bivariate random variable, then

$$\begin{aligned} f(x_1, x_2) &= f\left(x_1|x_2\right) f_2(x_2) \\ &= f\left(x_2|x_1\right) f_1(x_2). \end{aligned}$$

The joint can be factored two ways, and equating the two factorizations results in Bayes rule.

Definition 1.35 (Bivariate Bayes Rule)**.** Let X by a bivariate random variable with components X_1 and X_2, then

$$f\left(x_1|x_2\right) = \frac{f\left(x_2|x_1\right) f_1(x_1)}{f_2(x_2)} \tag{1.23}$$

Bayes rule states that the probability of observing X_1 given a value of X_2 is equal to the joint probability of the two random variables divided by the marginal probability of observing X_2. Bayes rule is normally applied where there is a belief about X_1 ($f_1(x_1)$, called a *prior*), and the conditional distribution of X_1 given X_2 is a known density ($f\left(x_2|x_1\right)$, called the *likelihood*), which combine to form a belief about X_1 ($f\left(x_1|x_2\right)$, called the *posterior*). The marginal density of X_2 is not important when using Bayes rule since the numerator is still proportional to the conditional density of X_1 given X_2 since $f_2(x_2)$ is a number, and so it is common to express the non-normalized posterior as

$$f\left(x_1|x_2\right) \propto f\left(x_2|x_1\right) f_1(x_1),$$

where $\propto$ is read "is proportional to".

[9] This can be generalized to the full multivariate case where X is an n-dimensional random variable where the first j components are independent from the last $n - j$ components defining $y_1 = Y_1\left(x_1, \ldots, x_j\right)$ and $y_2 = Y_2\left(x_{j+1}, \ldots, x_n\right)$.

Example 1.30. Suppose interest lies in the probability a firm does bankrupt which can be modeled as a Bernoulli distribution. The parameter p is unknown but, given a value of p, the likelihood that a firm goes bankrupt is

$$f\left(x|p\right) = p^x \left(1-p\right)^{1-x}.$$

While p is known, a prior for the bankruptcy rate can be specified. Suppose the prior for p follows a Beta(α, β) distribution which has pdf

$$f\left(p\right) = \frac{p^{\alpha-1}\left(1-p\right)^{\beta-1}}{B\left(\alpha,\beta\right)}$$

where $B\left(a,b\right)$ is Beta function that acts as a normalizing constant.[10] The Beta distribution has support on $[0,1]$ and nests the standard uniform as a special case when $\alpha = \beta = 1$. The expected value of a random variable with a Beta(α,β) is $\frac{\alpha}{\alpha+\beta}$ and the variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ where $\alpha > 0$ and $\beta > 0$.

Using Bayes rule,

$$\begin{aligned} f\left(p|x\right) &\propto p^x\left(1-p\right)^{1-x} \times \frac{p^{\alpha-1}\left(1-p\right)^{\beta-1}}{B\left(\alpha,\beta\right)} \\ &= \frac{p^{\alpha-1+x}\left(1-p\right)^{\beta-x}}{B\left(\alpha,\beta\right)}. \end{aligned}$$

Note that this isn't a density since it has the wrong normalizing constant. However, the component of the density which contains p is $p^{(\alpha-x)-1}\left(1-p\right)^{(\beta-x+1)-1}$ (known as the *kernel*) is the same as in the Beta distribution, only with different parameters. Thus the posterior, $f\left(p|x\right)$ is Beta$(\alpha+x, \beta-x+1)$. Since the posterior is the same as the prior, it could be combined with another observation (and the Bernoulli likelihood) to produce an updated posterior. When a Bayesian problem has this property, the prior density said to be conjugate to the likelihood.

Example 1.31. Suppose M is a random variable representing the score on the midterm, and interest lies in the final course grade, C. The prior for C is normal with mean μ and variance σ^2, and that the distribution of M given C is also conditionally normal with mean C and variance τ^2. Bayes rule can be used to make inference on the final course grade given the midterm grade.

$$\begin{aligned} f\left(c|m\right) &\propto f\left(m|c\right) f_C\left(c\right) \\ &\propto \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(m-c)^2}{2\tau^2}\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(c-\mu)^2}{2\sigma^2}\right) \\ &= K \exp\left(-\frac{1}{2}\left\{\frac{(m-c)^2}{\tau^2} + \frac{(c-\mu)^2}{\sigma^2}\right\}\right) \\ &= K \exp\left(-\frac{1}{2}\left\{\frac{c^2}{\tau^2} + \frac{c^2}{\sigma^2} - \frac{2cm}{\tau^2} - \frac{2c\mu}{\sigma^2} + \frac{m^2}{\tau^2} + \frac{\mu^2}{\sigma^2}\right\}\right) \\ &= K \exp\left(-\frac{1}{2}\left\{c^2\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right) - 2c\left(\frac{m}{\tau^2}+\frac{\mu}{\sigma^2}\right) + \left(\frac{m^2}{\tau^2}+\frac{\mu^2}{\sigma^2}\right)\right\}\right) \end{aligned}$$

[10] The beta function can only be given as an indefinite integral,

$$B\left(a,b\right) = \int_0^1 s^{a-1}\left(1-s\right)^{b-1} \mathrm{d}s.$$

This (non-normalized) density can be shown to have the kernel of a normal by completing the square,[11]

$$f\left(c|m\right) \propto \exp\left(-\frac{1}{2\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)^{-1}}\left(c-\frac{\left(\frac{m}{\tau^2}+\frac{\mu}{\sigma^2}\right)}{\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)}\right)^2\right).$$

This is the kernel of a normal density with mean

$$\frac{\left(\frac{m}{\tau^2}+\frac{\mu}{\sigma^2}\right)}{\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)},$$

and variance

$$\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)^{-1}.$$

The mean is a weighted average of the prior mean, μ and the midterm score, m, where the weights are determined by the inverse variance of the prior and conditional distributions. Since the weights are proportional to the inverse of the variance, a small variance leads to a relatively large weight. If $\tau^2 = \sigma^2$,then the posterior mean is the average of the prior mean and the midterm score. The variance of the posterior depends on the uncertainty in the prior (σ^2) and the uncertainty in the data (τ^2). The posterior variance is always less than the smaller of σ^2 and τ^2. Like the Bernoulli-Beta combination in the previous problem, the normal distribution is a conjugate prior when the conditional density is normal.

1.3.5 Common Multivariate Distributions

1.3.5.1 Multivariate Normal

Like the univariate normal, the multivariate normal depends on 2 parameters, $\boldsymbol{\mu}$and n by 1 vector of means and $\boldsymbol{\Sigma}$ an n by n positive semi-definite covariance matrix. The multivariate normal is closed to both to marginalization and conditioning – in other words, if X is multivariate normal, then all marginal distributions of X are normal, and so are all conditional distributions of X_1 given X_2 for any partitioning.

Parameters

$\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Sigma}$ a positive semi-definite matrix

Support

$\mathbf{x} \in \mathbb{R}^n$

[11] Suppose a quadratic in x has the form $ax^2 + bx + c$. Then

$$ax^2 + bx + c = a\left(x - d\right)^2 + e$$

where $d = b/(2a)$ and $e = c - b^2/(4a)$.

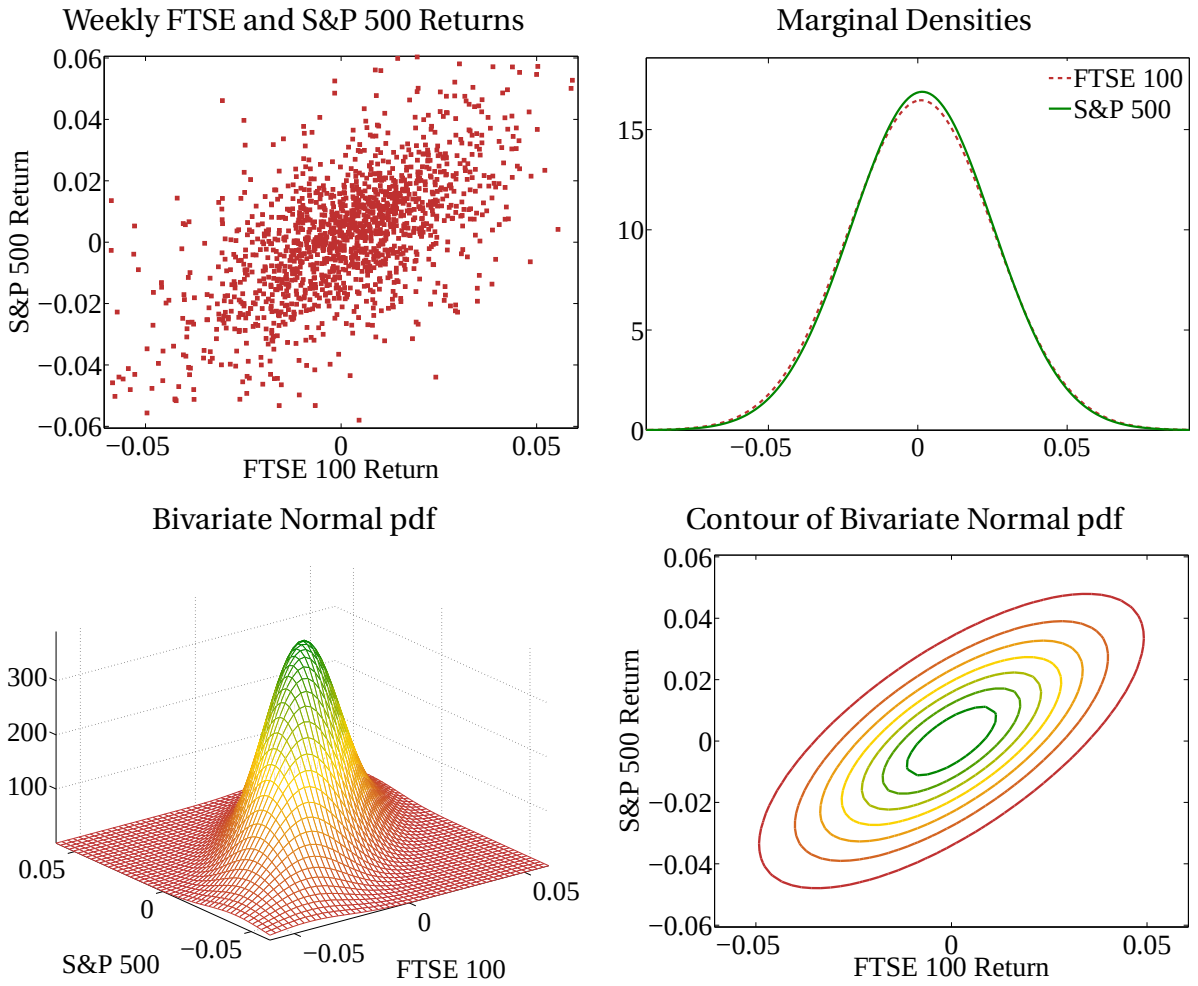


Figure 1.7: These four figures show different views of the weekly returns of the FTSE 100 and the S&P 500. The top left contains a scatter plot of the raw data. The top right shows the marginal distributions from a fit bivariate normal distribution (using maximum likelihood). The bottom two panels show two representations of the joint probability density function.

Probability Density Function

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Cumulative Distribution Function

Can be expressed as a series of n univariate normal cdfs using repeated conditioning.

Moments

Mean	$\boldsymbol{\mu}$
Median	$\boldsymbol{\mu}$
Variance	$\boldsymbol{\Sigma}$
Skewness	0
Kurtosis	3

Marginal Distribution

The marginal distribution for the first j components is

$$f_{X_1,\ldots X_j}\left(x_1,\ldots,x_j\right) = (2\pi)^{-\frac{j}{2}} |\boldsymbol{\Sigma}_{11}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\mathbf{x}_1-\boldsymbol{\mu}_1\right)'\boldsymbol{\Sigma}_{11}^{-1}\left(\mathbf{x}_1-\boldsymbol{\mu}_1\right)\right),$$

where it is assumed that the marginal distribution is that of the first j random variables[12], $\boldsymbol{\mu} = [\boldsymbol{\mu}_1' \; \boldsymbol{\mu}_2']'$ where $\boldsymbol{\mu}_1$ correspond to the first j entries, and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}' & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

In other words, the distribution of $\left[X_1,\ldots X_j\right]'$ is $N\left(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_{11}\right)$. Moreover, the marginal distribution of a single element of X is $N\left(\mu_i,\sigma_i^2\right)$ where μ_i is the i$^{\text{th}}$ element of μ and σ_i^2 is the i$^{\text{th}}$ diagonal element of $\boldsymbol{\Sigma}$.

Conditional Distribution

The conditional probability of X_1 given $X_2 = \mathbf{x}_2$ is

$$N\left(\boldsymbol{\mu}_1 + \boldsymbol{\beta}'\left(\mathbf{x}_2-\boldsymbol{\mu}_2\right), \boldsymbol{\Sigma}_{11} - \boldsymbol{\beta}'\boldsymbol{\Sigma}_{22}\boldsymbol{\beta}\right)$$

where $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}'$.

When X is a bivariate normal random variable,

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right),$$

[12] Any two variables can be reordered in a multivariate normal by swapping their means and reordering the corresponding rows and columns of the covariance matrix.

the conditional distribution is

$$X_1|X_2 = x_2 \sim N\left(\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{\sigma_{12}^2}{\sigma_2^2}\right),$$

where the variance can be seen to always be positive since $\sigma_1^2\sigma_2^2 \geq \sigma_{12}^2$ by the Cauchy-Schwarz inequality (see 1.15).

Notes

The multivariate Normal has a number of novel and useful properties:

- A standard multivariate normal has $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}_n$.
- If the covariance between elements i and j equals zero (so that $\sigma_{ij} = 0$), they are independent.
- For the normal, zero covariance (or correlation) implies independence. This is not true of most other multivariate random variables.
- Weighted sums of multivariate normal random variables are normal. In particular is $\mathbf{c}$ is a n by 1 vector of weights, then $Y = \mathbf{c}'X$ is normal with mean $\mathbf{c}'\boldsymbol{\mu}$ and variance $\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}$.

1.4 Expectations and Moments

Expectations and moments are (non-random) functions of random variables that are useful in both understanding properties of random variables – e.g. when comparing the dispersion between two distributions – and when estimating parameters using a technique known as the method of moments (see Chapter 1).

1.4.1 Expectations

The expectation is the value, on average, of a random variable (or function of a random variable). Unlike common English language usage, where one's expectation is not well defined (e.g. could be the mean or the mode, another measure of the tendency of a random variable), the expectation in a probabilistic sense *always* averages over the possible values weighting by the probability of observing each value. The form of an expectation in the discrete case is particularly simple.

Definition 1.36 (Expectation of a Discrete Random Variable)**.** The expectation of a discrete random variable, defined $\mathrm{E}[X] = \sum_{x \in R(X)} x f(x)$, exists if and only if $\sum_{x \in R(X)} |x| f(x) < \infty$.

When the range of X is finite then the expectation always exists. When the range is infinite, such as when a random variable takes on values in the range 0, 1, 2, . . ., the probability mass function must be sufficiently small for large values of the random variable in order for the expectation to exist.[13] Expectations of continuous random variables are virtually identical, only replacing the sum with an integral.

[13] An expectation is said to be nonexistent when the sum converges to $\pm\infty$ or oscillates. The use of the $|x|$ in the definition of existence is to rule out both the $-\infty$ and the oscillating cases.

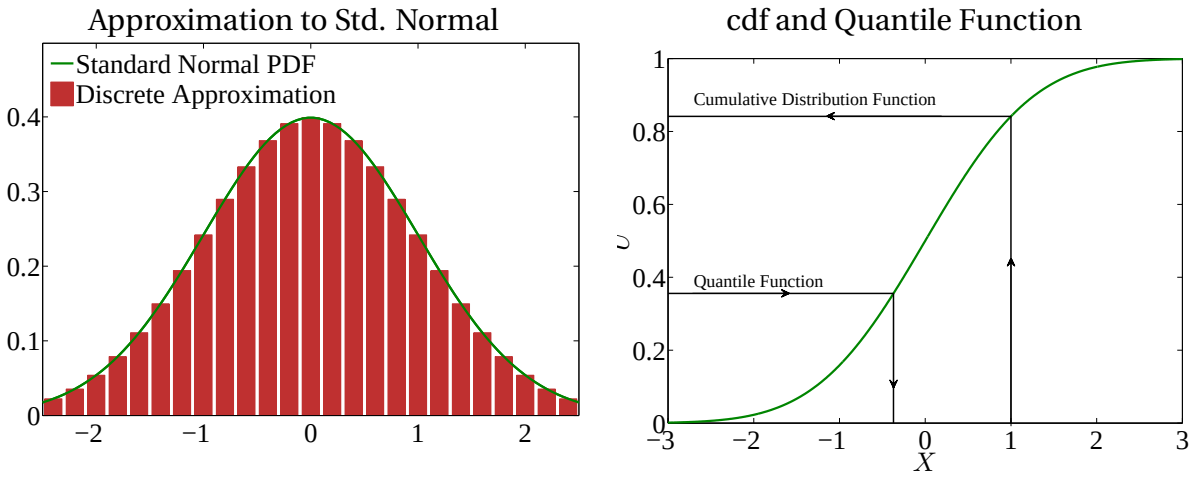


Figure 1.8: The left panel shows a standard normal and a discrete approximation. Discrete approximations are useful for approximating integrals in expectations. The right panel shows the relationship between the quantile function and the cdf.

Definition 1.37 (Expectation of a Continuous Random Variable)**.** The expectation of a continuous random variable, defined $\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$, exists if and only if $\int_{-\infty}^{\infty} |x|\, f(x)\,\mathrm{d}x < \infty$.

The existence of an expectation is a somewhat difficult concept. For continuous random variables, expectations may not exist if the probability of observing an arbitrarily large value (in the absolute sense) is very high. For example, in a Student's t distribution when the degree of freedom parameter ν is 1 (also known as a Cauchy distribution), the probability of observing a value with size $|x|$ is proportional to x^{-1} for large x (in other words, $f(x) \propto c\,x^{-1}$) so that $x f(x) \approx c$ for large x. The range is unbounded, and so the integral of a constant, even if very small, will not converge, and so the expectation does not exist. On the other hand, when a random variable is bounded, it's expectation always exists.

Theorem 1.9 (Expectation Existence for Bounded Random Variables)**.** *If* $|x| < c$ *for all* $x \in R(X)$, *then* $\mathrm{E}[X]$ *exists.*

The expectation operator, $\mathrm{E}[\cdot]$ is generally defined for arbitrary functions of a random variable, $g(x)$. In practice, $g(x)$ takes many forms – x, x^2, x^p for some p, $\exp(x)$ or something more complicated. Discrete and continuous expectations are closely related. Figure 1.8 shows a standard normal along with a discrete approximation where each bin has a width of 0.20 and the height is based on the pdf value at the mid-point of the bin. Treating the normal as a discrete distribution based on this approximation would provide reasonable approximations to the correct (integral) expectations.

Definition 1.38 (Expectation of a Function of Random Variable)**.** The expectation of a random variable defined as a function of X, $Y = g(x)$, is $\mathrm{E}[Y] = \mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\,\mathrm{d}x$ exists if and only if $\int_{-\infty}^{\infty} |g(x)|\,\mathrm{d}x < \infty$.

When $g(x)$ is either concave or convex, Jensen's inequality provides a relationship between the expected value of the function and the function of the expected value of the underlying random variable.

Theorem 1.10 (Jensen's Inequality)**.** *If $g(\cdot)$ is a continuous convex function on an open interval containing the range of X, then $\mathrm{E}[g(X)] \geq g(\mathrm{E}[X])$. Similarly, if $g(\cdot)$ is a continuous concave function on an open interval containing the range of X, then $\mathrm{E}[g(X)] \leq g(\mathrm{E}[X])$.*

The inequalities become strict if the functions are strictly convex (or concave) as long as X is not degenerate.[14] Jensen's inequality is common in economic applications. For example, standard utility functions ($U(\cdot)$) are assumed to be concave which reflects the idea that marginal utility ($U'(\cdot)$) is decreasing in consumption (or wealth). Applying Jensen's inequality shows that if consumption is random, then $\mathrm{E}[U(c)] < U(E[c])$ – in other words, the economic agent is worse off when facing uncertain consumption. Convex functions are also commonly encountered, for example in option pricing or in (production) cost functions. The expectations operator has a number of simple and useful properties:

- If c is a constant, then $\mathrm{E}[c] = c$. This property follows since the expectation is an integral against a probability density which integrates to unity.
- If c is a constant, then $\mathrm{E}[cX] = c\mathrm{E}[X]$. This property follows directly from passing the constant out of the integral in the definition of the expectation operator.
- The expectation of the sum is the sum of the expectations,

$$\mathrm{E}\left[\sum_{i=1}^{k} g_i(X)\right] = \sum_{i=1}^{k} \mathrm{E}[g_i(X)].$$

  This property follows directly from the distributive property of multiplication.
- If a is a constant, then $\mathrm{E}[a + X] = a + \mathrm{E}[X]$. This property also follows from the distributive property of multiplication.
- $\mathrm{E}[f(X)] = f(\mathrm{E}[X])$ when $f(x)$ is affine (i.e. $f(x) = a + bx$ where a and b are constants). For general non-linear functions, it is usually the case that $\mathrm{E}[f(X)] \neq f(\mathrm{E}[X])$ when X is non-degenerate.
- $\mathrm{E}[X^p] \neq \mathrm{E}[X]^p$ except when $p = 1$ when X is non-degenerate.

These rules are used throughout financial economics when studying random variables and functions of random variables.

The expectation of a function of a multivariate random variable is similarly defined, only integrating across all dimensions.

Definition 1.39 (Expectation of a Multivariate Random Variable)**.** Let $(X_1, X_2, \ldots, X_n)$ be a continuously distributed n-dimensional multivariate random variable with joint density function $f(x_1, x_2, \ldots x_n)$. The expectation of $Y = g(X_1, X_2, \ldots, X_n)$ is defined as

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n)\, \mathrm{d}x_1\, \mathrm{d}x_2 \ldots \mathrm{d}x_n. \tag{1.24}$$

[14] A degenerate random variable has probability 1 on a single point, and so is not meaningfully random.

It is straight forward to see that rule that the expectation of the sum is the sum of the expectation carries over to multivariate random variables, and so

$$\mathrm{E}\left[\sum_{i=1}^{n} g_i\left(X_1, \ldots X_n\right)\right] = \sum_{i=1}^{n} \mathrm{E}\left[g_i\left(X_1, \ldots X_n\right)\right].$$

Additionally, taking $g_i\left(X_1, \ldots X_n\right) = X_i$, we have $\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{E}\left[X_i\right]$.

1.4.2 Moments

Moments are expectations of particular functions of a random variable, typically $g\left(x\right) = x^s$ for $s = 1, 2, \ldots$, and are often used to compare distributions or to estimate parameters.

Definition 1.40 (Noncentral Moment)**.** The r$^{\text{th}}$ noncentral moment of a continuous random variable X is defined

$$\mu_r' \equiv \mathrm{E}\left[X^r\right] = \int_{-\infty}^{\infty} x^r f\left(x\right) \mathrm{d}x \tag{1.25}$$

for $r = 1, 2, \ldots$.

The first non-central moment is the average, or mean, of the random variable.

Definition 1.41 (Mean)**.** The first non-central moment of a random variable X is called the mean of X and is denoted μ.

Central moments are similarly defined, only centered around the mean.

Definition 1.42 (Central Moment)**.** The r$^{\text{th}}$ central moment of a random variables X is defined

$$\mu_r \equiv \mathrm{E}\left[\left(X - \mu\right)^r\right] = \int_{-\infty}^{\infty} \left(x - \mu\right)^r f\left(x\right) \mathrm{d}x \tag{1.26}$$

for $r = 2, 3 \ldots$.

Aside from the first moment, references to "moments" refer to central moments. Moments may not exist if a distribution is sufficiently heavy-tailed. However, if the r^{th} moment exists, then any moment of lower order must also exist.

Theorem 1.11 (Lesser Moment Existence)**.** *If μ_r' exists for some r, then μ_s' exists for $s \leq r$. Moreover, for any r, μ_r' exists if and only if μ_r exists.*

Central moments are used to describe a distribution since they are invariant to changes in the mean. The second central moment is known as the variance.

Definition 1.43 (Variance)**.** The second central moment of a random variable X, $\mathrm{E}\left[\left(X - \mu\right)^2\right]$ is called the variance and is denoted σ^2 or equivalently $\mathrm{V}\left[X\right]$.

The variance operator ($\mathrm{V}[\cdot]$) also has a number of useful properties.

- If c is a constant, then $\mathrm{V}[c] = 0$.
- If c is a constant, then $\mathrm{V}[cX] = c^2\mathrm{V}[X]$.
- If a is a constant, then $\mathrm{V}[a + X] = \mathrm{V}[X]$.
- The variance of the sum is the sum of the variances plus twice all of the covariances[a],

$$\mathrm{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{V}[X_i] + 2\sum_{j=1}^{n}\sum_{k=j+1}^{n} \mathrm{Cov}\left[X_j, X_k\right]$$

[a] See Section 1.4.7 for more on covariances.

The variance is a measure of dispersion, although the square root of the variance, known as the standard deviation, is typically more useful.[15]

Definition 1.44 (Standard Deviation)**.** The square root of the variance is known as the standard deviations and is denoted σ or equivalently $\mathrm{std}(X)$.

The standard deviation is a more meaningful measure than the variance since its *units* are the same as the mean (and random variable). For example, suppose X is the return on the stock market next year, and that the mean of X is 8% and the standard deviation is 20% (the variance is .04). The mean and standard deviation are both measured as the percentage change in investment, and so can be directly compared, such as in the Sharpe ratio (Sharpe, 1994). Applying the properties of the expectation operator and variance operator, it is possible to define a studentized (or standardized) random variable.

Definition 1.45 (Studentization)**.** Let X be a random variable with mean μ and variance σ^2, then

$$Z = \frac{x - \mu}{\sigma} \tag{1.27}$$

is a studentized version of X (also known as standardized). Z has mean 0 and variance 1.

Standard deviation also provides a bound on the probability which can lie in the tail of a distribution, as shown in Chebyshev's inequality.

Theorem 1.12 (Chebyshev's Inequality)**.** $\Pr\left[|x - \mu| \geq k\sigma\right] \leq 1/k^2$ *for* $k > 0$.

Chebyshev's inequality is useful in a number of contexts. One of the most useful is in establishing consistency in any an estimator which has a variance that tends to 0 as the sample size diverges.

The third central moment does not have a specific name, although it is called the skewness when standardized by the scaled variance.

[15] The standard deviation is occasionally confused for the standard error. While both are square roots of variances, the standard deviation refers to deviation in a random variable while the standard error is reserved for parameter estimators.

Definition 1.46 (Skewness)**.** The third central moment, standardized by the second central moment raised to the power 3/2,

$$\frac{\mu_3}{(\sigma^2)^{\frac{3}{2}}} = \frac{\mathrm{E}\left[(X-\mathrm{E}[X])^3\right]}{\mathrm{E}\left[(X-\mathrm{E}[X])^2\right]^{\frac{3}{2}}} = \mathrm{E}\left[Z^3\right] \tag{1.28}$$

is defined as the skewness where Z is a studentized version of X.

The skewness is a general measure of asymmetry, and is 0 for symmetric distribution (assuming the third moment exists). The normalized fourth central moment is known as the kurtosis.

Definition 1.47 (Kurtosis)**.** The fourth central moment, standardized by the squared second central moment,

$$\frac{\mu_4}{(\sigma^2)^2} = \frac{\mathrm{E}\left[(X-\mathrm{E}[X])^4\right]}{\mathrm{E}\left[(X-\mathrm{E}[X])^2\right]^2} = \mathrm{E}\left[Z^4\right] \tag{1.29}$$

is defined as the kurtosis and is denoted κ where Z is a studentized version of X.

Kurtosis measures of the chance of observing a large (and absolute terms) value, and is often expressed as excess kurtosis.

Definition 1.48 (Excess Kurtosis)**.** The kurtosis of a random variable minus the kurtosis of a normal random variable, $\kappa - 3$, is known as excess kurtosis.

Random variables with a positive excess kurtosis are often referred to as heavy-tailed.

1.4.3 Related Measures

While moments are useful in describing the properties of a random variable, other measures are also commonly encountered. The median is an alternative measure of central tendency.

Definition 1.49 (Median)**.** Any number m satisfying $\Pr(X \leq m) = 0.5$ and $\Pr(X \geq m) = 0.5$ is known as the median of X.

The median measures the point where 50% of the distribution lies on either side (it may not be unique), and is just a particular quantile. The median has a few advantages over the mean, and in particular, it is less affected by outliers (e.g. the difference between mean and median income) and it always exists (the mean doesn't exist for very heavy-tailed distributions).

The interquartile range uses quartiles[16] to provide an alternative measure of dispersion than standard deviation.

Definition 1.50 (Interquartile Range)**.** The value $q_{.75} - q_{.25}$ is known as the interquartile range.

The mode complements the mean and median as a measure of central tendency. A mode is a local maximum of a density.

[16] Other *tiles* include terciles (3), quartiles (4), quintiles (5), deciles (10) and percentiles (100). In all cases the bin ends are $[(i-1/m), i/m]$ where m is the number of bins and $i = 1, 2, \ldots, m$.

Definition 1.51 (Mode)**.** Let X be a random variable with density function $f(x)$. A point c where $f(x)$ attains a maximum is known as a mode.

Distributions can be unimodal or multimodal.

Definition 1.52 (Unimodal Distribution)**.** Any random variable which has a single, unique mode is called unimodal.

Note that modes in a multimodal distribution do not necessarily have to have equal probability.

Definition 1.53 (Multimodal Distribution)**.** Any random variable which as more than one mode is called multimodal.

Figure 1.9 shows a number of distributions. The distributions depicted in the top panels are all unimodal. The distributions in the bottom pane are mixtures of normals, meaning that with probability p random variables come from one normal, and with probability $1-p$ they are drawn from the other. Both mixtures of normals are multimodal.

1.4.4 Multivariate Moments

Other moment definitions are only meaningful when studying 2 or more random variables (or an n-dimensional random variable). When applied to a vector or matrix, the expectations operator applies element-by-element. For example, if X is an n-dimensional random variable,

$$\mathrm{E}[X] = \mathrm{E}\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \\ \vdots \\ \mathrm{E}[X_n] \end{bmatrix}. \tag{1.30}$$

Covariance is a measure which captures the tendency of two variables to move together in a linear sense.

Definition 1.54 (Covariance)**.** The covariance between two random variables X and Y is defined

$$\mathrm{Cov}[X, Y] = \sigma_{XY} = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]. \tag{1.31}$$

Covariance can be alternatively defined using the joint product moment and the product of the means.

Theorem 1.13 (Alternative Covariance)**.** *The covariance between two random variables X and Y can be equivalently defined*

$$\sigma_{XY} = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]. \tag{1.32}$$

Inverting the covariance expression shows that no covariance is sufficient to ensure that the expectation of a product is the product of the expectations.

Theorem 1.14 (Zero Covariance and Expectation of Product)**.** *If X and Y have $\sigma_{XY} = 0$, then* $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.

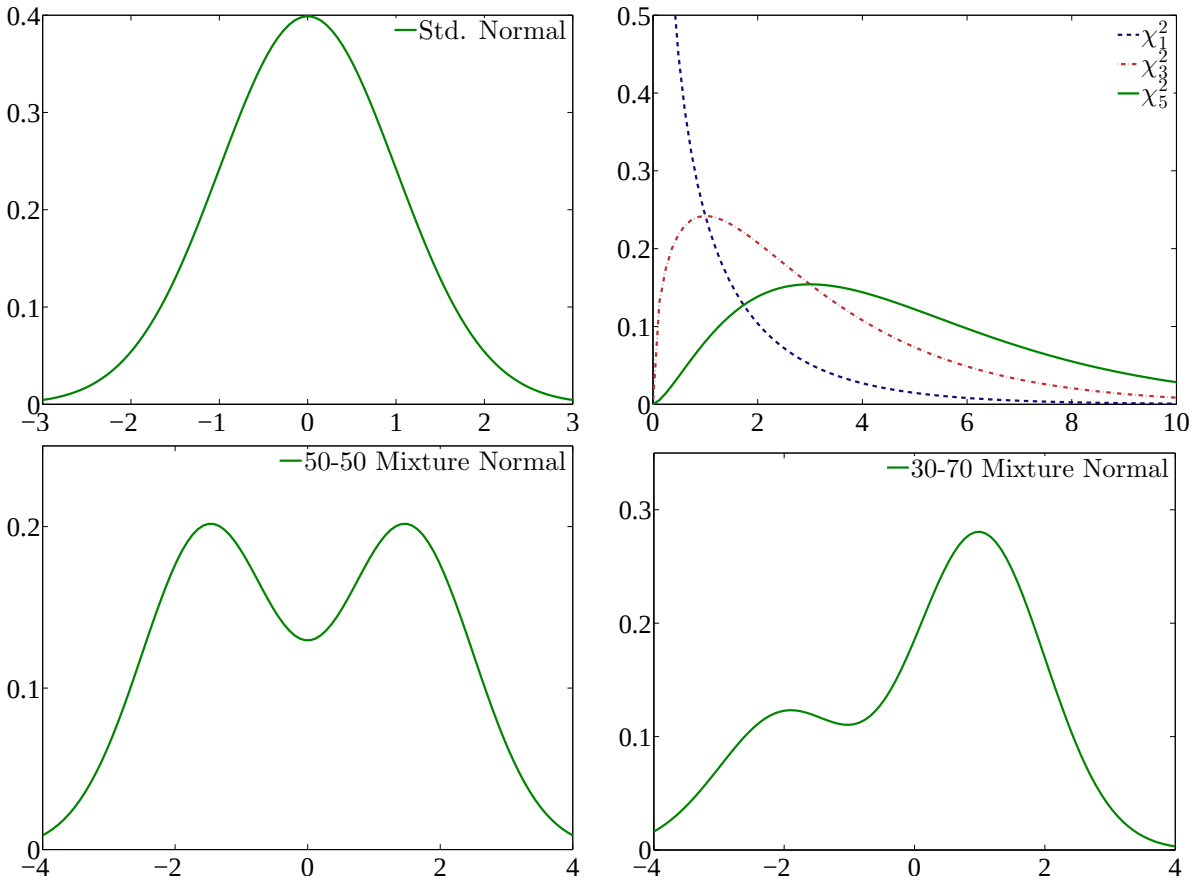


Figure 1.9: These four figures show two unimodal (upper panels) and two multimodal (lower panels) distributions. The upper left is a standard normal density. The upper right shows three χ^2 densities for $\nu = 1, 3$ and 5. The lower panels contain mixture distributions of 2 normals – the left is a 50-50 mixture of $N(-1, 1)$ and $N(1, 1)$ and the right is a 30-70 mixture of $N(-2, 1)$ and $N(1, 1)$.

The previous result follows directly from the definition of covariance since $\sigma_{XY} = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$. In financial economics, this result is often applied to products of random variables so that the mean of the product can be directly determined by knowledge of the mean of each variable and the covariance between the two. For example, when studying consumption based asset pricing, it is common to encounter terms involving the expected value of consumption growth times the pricing kernel (or stochastic discount factor) – in many cases the full joint distribution of the two is intractable although the mean and covariance of the two random variables can be determined.

The Cauchy-Schwarz inequality is a version of the triangle inequality and states that the expectation of the squared product is less than the product of the squares.

Theorem 1.15 (Cauchy-Schwarz Inequality)**.** $\mathrm{E}\left[(XY)^2\right] \leq \mathrm{E}\left[X^2\right]\mathrm{E}\left[Y^2\right]$.

Example 1.32. When X is an n-dimensional random variable, it is useful to assemble the variances and covariances into a covariance matrix.

Definition 1.55 (Covariance Matrix)**.** The covariance matrix of an n-dimensional random variable X is defined

$$\mathrm{Cov}[X] = \mathbf{\Sigma} = \mathrm{E}\left[(X - \mathrm{E}[X])(X - \mathrm{E}[X])'\right] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \vdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \dots & \sigma_n^2 \end{bmatrix}$$

where the i$^{\text{th}}$ diagonal element contains the variance of X_i (σ_i^2) and the element in position (i, j) contains the covariance between X_i and X_j $\left(\sigma_{ij}\right)$.

When X is composed of two sub-vectors, a block form of the covariance matrix is often convenient.

Definition 1.56 (Block Covariance Matrix)**.** Suppose X_1 is an n_1-dimensional random variable and X_2 is an n_2-dimensional random variable. The block covariance matrix of $X = \left[X_1' \; X_2'\right]'$ is

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{12}' & \mathbf{\Sigma}_{22} \end{bmatrix} \tag{1.33}$$

where $\mathbf{\Sigma}_{11}$ is the n_1 by n_1 covariance of X_1, $\mathbf{\Sigma}_{22}$ is the n_2 by n_2 covariance of X_2 and $\mathbf{\Sigma}_{12}$ is the n_1 by n_2 covariance matrix between X_1 and X_2 and element (i, j) equal to $\mathrm{Cov}\left[X_{1,i}, X_{2,j}\right]$.

A standardized version of covariance is often used to produce a scale-free measure.

Definition 1.57 (Correlation)**.** The correlation between two random variables X and Y is defined

$$\mathrm{Corr}[X, Y] = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}. \tag{1.34}$$

Additionally, the correlation is always in the interval $[-1, 1]$, which follows from the Cauchy-Schwarz inequality.

Theorem 1.16. *If X and Y are independent random variables, then $\rho_{XY} = 0$ as long as σ_X^2 and σ_Y^2 exist.*

It is important to note that the converse of this statement is not true – that is, a lack of correlation does not imply that two variables are independent. In general, a correlation of 0 only implies independence when the variables are multivariate normal.

Example 1.33. Suppose X and Y have $\rho_{XY} = 0$, then X and Y are not necessarily independent. Suppose X is a discrete uniform random variable taking values in $\{-1, 0, 1\}$ and $Y = X^2$, so that $\sigma_X^2 = 2/3$, $\sigma_Y^2 = 2/9$ and $\sigma_{XY} = 0$. While X and Y are uncorrelated, the are clearly not independent, since when the random variable Y takes the value 1, X must be 0.

The corresponding correlation matrix can be assembled. Note that a correlation matrix has 1s on the diagonal and values bounded by $[-1, 1]$ on the off-diagonal positions.

Definition 1.58 (Correlation Matrix)**.** The correlation matrix of an n-dimensional random variable X is defined

$$(\Sigma \odot \mathbf{I}_n)^{-\frac{1}{2}} \Sigma (\Sigma \odot \mathbf{I}_n)^{-\frac{1}{2}} \tag{1.35}$$

where the i, j^{th} element has the form $\sigma_{X_i X_j} / \left(\sigma_{X_i} \sigma_{X_j}\right)$ when $i \neq j$ and 1 when $i = j$.

1.4.5 Conditional Expectations

Conditional expectations are similar to other forms of expectations only using conditional densities in place of joint or marginal densities. Conditional expectations essentially treat one of the variables (in a bivariate random variable) as constant.

Definition 1.59 (Bivariate Conditional Expectation)**.** Let X be a continuous bivariate random variable comprised of X_1 and X_2. The conditional expectation of X_1 given X_2

$$\mathrm{E}\left[g(X_1) | X_2 = x_2\right] = \int_{-\infty}^{\infty} g(x_1) f\left(x_1 | x_2\right) \, \mathrm{d}x_1 \tag{1.36}$$

where $f\left(x_1 | x_2\right)$ is the conditional probability density function of X_1 given X_2.[17]

In many cases, it is useful to avoid specifying a specific value for X_2 in which case $\mathrm{E}\left[X_1 | X_1\right]$ will be used. Note that $\mathrm{E}\left[X_1 | X_2\right]$ will typically be a function of the random variable X_2.

Example 1.34. Suppose X is a bivariate normal distribution with components X_1 and X_2, $\boldsymbol{\mu} = [\mu_1 \; \mu_2]'$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix},$$

then $\mathrm{E}\left[X_1 | X_2 = x_2\right] = \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}(x_2 - \mu_2)$. This follows from the conditional density of a bivariate random variable.

The law of iterated expectations uses conditional expectations to show that the conditioning does not affect the final result of taking expectations – in other words, the order of taking expectations does not matter.

[17] A conditional expectation can also be defined in a natural way for functions of X_1 given $X_2 \in B$ where $\Pr(X_2 \in B) > 0$.

Theorem 1.17 (Bivariate Law of Iterated Expectations)**.** *Let X be a continuous bivariate random variable comprised of X_1 and X_2. Then* $\mathrm{E}\left[\mathrm{E}\left[g\left(X_1\right)|X_2\right]\right] = \mathrm{E}\left[g\left(X_1\right)\right]$.

The law of iterated expectations follows from basic properties of an integral since the order of integration does not matter as long as all integrals are taken.

Example 1.35. Suppose X is a bivariate normal distribution with components X_1 and X_2, $\boldsymbol{\mu} = [\mu_1\ \mu_2]'$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix},$$

then $\mathrm{E}[X_1] = \mu_1$ and

$$\begin{aligned} \mathrm{E}\left[\mathrm{E}\left[X_1|X_2\right]\right] &= \mathrm{E}\left[\mu_1 + \frac{\sigma_{12}}{\sigma_2^2}\left(X_2 - \mu_2\right)\right] \\ &= \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}\left(\mathrm{E}\left[X_2\right] - \mu_2\right) \\ &= \mu_1 + \frac{\sigma_{12}}{\sigma_2^2}\left(\mu_2 - \mu_2\right) \\ &= \mu_1. \end{aligned}$$

When using conditional expectations, any random variable conditioned on behaves "as-if" non-random (in the conditional expectation), and so $\mathrm{E}\left[\mathrm{E}\left[X_1X_2|X_2\right]\right] = \mathrm{E}\left[X_2\mathrm{E}\left[X_1|X_2\right]\right]$. This is a very useful tool when combined with the law of iterated expectations when $\mathrm{E}\left[X_1|X_2\right]$ is a known function of X_2.

Example 1.36. Suppose X is a bivariate normal distribution with components X_1 and X_2, $\boldsymbol{\mu} = \mathbf{0}$ and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix},$$

then

$$\begin{aligned} \mathrm{E}\left[X_1X_2\right] &= \mathrm{E}\left[\mathrm{E}\left[X_1X_2|X_2\right]\right] \\ &= \mathrm{E}\left[X_2\mathrm{E}\left[X_1|X_2\right]\right] \\ &= \mathrm{E}\left[X_2\left(\frac{\sigma_{12}}{\sigma_2^2}X_2\right)\right] \\ &= \frac{\sigma_{12}}{\sigma_2^2}\mathrm{E}\left[X_2^2\right] \\ &= \frac{\sigma_{12}}{\sigma_2^2}\left(\sigma_2^2\right) \\ &= \sigma_{12}. \end{aligned}$$

One particularly useful application of conditional expectations occurs when the conditional expectation is known and constant, so that $\mathrm{E}\left[X_1|X_2\right] = c$.

Example 1.37. Suppose X is a bivariate random variable composed of X_1 and X_2 and that $\mathrm{E}\left[X_1|X_2\right] = c$. Then $\mathrm{E}[X_1] = c$ since

$$\begin{aligned}\mathrm{E}[X_1] &= \mathrm{E}\left[\mathrm{E}\left[X_1|X_2\right]\right] \\ &= \mathrm{E}[c] \\ &= c.\end{aligned}$$

Conditional expectations can be taken for general n-dimensional random variables, and the law of iterated expectations holds as well.

Definition 1.60 (Conditional Expectation)**.** Let X be a n-dimensional random variable and and partition the first j ($1 \leq j < n$) elements of X into X_1, and the remainder into X_2 so that $X = \left[X_1' \; X_2'\right]'$. The conditional expectation of $g(X_1)$ given $X_2 = \mathbf{x}_2$

$$\mathrm{E}\left[g(X_1)|X_2 = \mathbf{x}_2\right] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} g\left(x_1, \ldots, x_j\right) f\left(x_1, \ldots, x_j|\mathbf{x}_2\right) \mathrm{d}x_j \ldots \mathrm{d}x_1 \tag{1.37}$$

where $f\left(x_1, \ldots, x_j|\mathbf{x}_2\right)$ is the conditional probability density function of X_1 given $X_2 = \mathbf{x}_2$.

The law of iterated expectations also holds for arbitrary partitions as well.

Theorem 1.18 (Law of Iterated Expectations)**.** *Let X be a n-dimensional random variable and and partition the first j ($1 \leq j < n$) elements of X into X_1, and the remainder into X_2 so that $X = \left[X_1' \; X_2'\right]'$. Then* $\mathrm{E}\left[\mathrm{E}\left[g(X_1)|X_2\right]\right] = \mathrm{E}[g(X_1)]$. *The law of iterated expectations is also known as the law of total expectations.*

Full multivariate conditional expectations are extremely common in time series. For example, when using daily data, there are over 30,000 observations of the Dow Jones Industrial Average available to model. Attempting to model the full joint distribution would be a formidable task. On the other hand, modeling the conditional expectation (or conditional mean) of the final observation, conditioning on those observations in the past, is far simpler.

Example 1.38. Suppose $\{X_t\}$ is a sequence of random variables where X_t comes after X_{t-j} for $j \geq 1$. The conditional conditional expectation of X_t given its past is

$$\mathrm{E}\left[X_t|X_{t-1,}X_{t-2}, \ldots\right].$$

Example 1.39. Let $\{\epsilon_t\}$ be a sequence of independent, identically distributed random variables with mean 0 and variance $\sigma^2 < \infty$. Define $X_0 = 0$ and $X_t = X_{t-1} + \epsilon_t$. X_t is a random walk, and $\mathrm{E}\left[X_t|X_{t-1}\right] = X_{t-1}$.

This leads naturally to the definition of a martingale, which is an important concept in financial economics which related to efficient markets.

Definition 1.61 (Martingale)**.** If $\mathrm{E}\left[X_{t+j}|X_{t-1}, X_{t-2} \ldots\right] = X_{t-1}$ for all $j \geq 0$ and $\mathrm{E}\left[|X_t|\right] < \infty$, both holding for all t, then $\{X_t\}$ is a martingale. Similarly, if $\mathrm{E}\left[X_{t+j} - X_{t-1}|X_{t-1}, X_{t-2} \ldots\right] = 0$ for all $j \geq 0$ and $\mathrm{E}\left[|X_t|\right] < \infty$, both holding for all t, then $\{X_t\}$ is a martingale.

1.4.6 Conditional Moments

All moments can be transformed made conditional by integrating against the conditional probability density function. For example, the (unconditional) mean becomes the conditional mean, and the variance becomes a conditional variance.

Definition 1.62 (Conditional Variance)**.** The variance of a random variable X conditional on another random variable Y is

$$\begin{aligned} \mathrm{V}\left[X|Y\right] &= \mathrm{E}\left[\left(X-\mathrm{E}\left[X|Y\right]\right)^2|Y\right] \\ &= \mathrm{E}\left[X^2|Y\right]-\mathrm{E}\left[X|Y\right]^2. \end{aligned} \tag{1.38}$$

The two definitions of conditional variance are identical to those of the (unconditional) variance where the (unconditional) expectation has been replaced by a conditional expectation. Conditioning can be used to compute higher-order moments as well.

Definition 1.63 (Conditional Moment)**.** The r^{th} central moment of a random variables X conditional on another random variable Y is defined

$$\mu_r \equiv \mathrm{E}\left[\left(X-\mathrm{E}\left[X|Y\right]\right)^r|Y\right] \tag{1.39}$$

for $r=2,3,\ldots$.

Combining the conditional expectation and the conditional variance leads to the law of total variance.

Theorem 1.19. *The variance of a random variable X can be decomposed into the variance of the conditional expectation plus the expectation of the conditional variance,*

$$\mathrm{V}[X]=\mathrm{V}\left[\mathrm{E}\left[X|Y\right]\right]+\mathrm{E}\left[\mathrm{V}\left[X|Y\right]\right]. \tag{1.40}$$

The law of total variance shows that the total variance of a variable can be decomposed into the variability of the conditional mean plus the average of the conditional variance. This is a useful decomposition for time-series.

Independence can also be defined conditionally.

Definition 1.64 (Conditional Independence)**.** Two random variables X_1 and X_2 are conditionally independent, conditional on Y, if

$$f\left(x_1,x_2|y\right)=f_1\left(x_1|y\right)f_2\left(x_2|y\right).$$

Note that random variables that are conditionally independent are not necessarily unconditionally independent.

Example 1.40. Suppose X is a trivariate normal random variable with mean $\mathbf{0}$ and covariance

$$\mathbf{\Sigma}=\begin{bmatrix}\sigma_1^2 & 0 & 0\\ 0 & \sigma_2^2 & 0\\ 0 & 0 & \sigma_3^2\end{bmatrix}$$

and define $Y_1 = x_1 + x_3$ and $Y_2 = x_2 + x_3$. Then Y_1 and Y_2 are correlated bivariate normal with mean $\mathbf{0}$ and covariance

$$\mathbf{\Sigma}_Y = \begin{bmatrix} \sigma_1^2 + \sigma_3^2 & \sigma_3^2 \\ \sigma_3^2 & \sigma_2^2 + \sigma_3^2 \end{bmatrix},$$

but the joint distribution of Y_1 and Y_2 given X_3 is bivariate normal with mean $\mathbf{0}$ and

$$\mathbf{\Sigma}_{Y|X_3} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

and so Y_1 and Y_2 are independent conditional on X_3.

Other properties of unconditionally independent random variables continue to hold for conditionally independent random variables. For example, when X_1 and X_2 are independent conditional on X_3, then the conditional covariance between X_1 and X_2 is 0 (as is the conditional correlation), and $\mathrm{E}\left[\mathrm{E}\left[X_1 X_2 | X_3\right]\right] = \mathrm{E}\left[\mathrm{E}\left[X_1|X_3\right]\mathrm{E}\left[X_2|X_3\right]\right]$ – that is, the conditional expectation of the product is the product of the conditional expectations.

1.4.7 Vector and Matrix Forms

Vector and matrix forms are particularly useful in finance since portfolios are often of interest where the underlying random variables are the individual assets and the combination vector is the vector of portfolio weights.

Theorem 1.20. *Let $Y = \sum_{i=1}^n c_i X_i$ where c_i, $i = 1, \ldots, n$ are constants. Then $\mathrm{E}[Y] = \sum_{i=1}^n c_i \mathrm{E}[X_i]$. In matrix notation, $Y = \mathbf{c}'\mathbf{x}$ where $\mathbf{c}$ is an n by 1 vector and $\mathrm{E}[Y] = \mathbf{c}'\mathrm{E}[X]$.*

The variance of the sum is the weighted sum of the variance plus all of the covariances.

Theorem 1.21. *Let $Y = \sum_{i=1}^n c_i X_i$ where c_i are constants. Then*

$$V[Y] = \sum_{i=1}^n c_i^2 V[X_i] + 2\sum_{j=1}^n \sum_{k=j+1}^n c_j c_k \mathrm{Cov}\left[X_i, X_j\right] \tag{1.41}$$

or equivalently

$$\sigma_Y^2 = \sum_{i=1}^n c_i^2 \sigma_{X_i}^2 + 2\sum_{j=1}^n \sum_{k=j+1}^n c_j c_k \sigma_{X_j X_k}.$$

This result can be equivalently expressed in vector-matrix notation.

Theorem 1.22. *Let $\mathbf{c}$ in an n by 1 vector and let X by an n-dimensional random variable with covariance $\mathbf{\Sigma}$. Define $Y = \mathbf{c}'\mathbf{x}$. The variance of Y is $\sigma_Y^2 = \mathbf{c}'\mathrm{Cov}[X]\,\mathbf{c} = \mathbf{c}'\mathbf{\Sigma}\mathbf{c}$.*

Note that the result holds when $\mathbf{c}$ is replaced by a matrix $\mathbf{C}$.

Theorem 1.23. *Let $\mathbf{C}$ be an n by m matrix and let X be an n-dimensional random variable with mean $\boldsymbol{\mu}_X$ and covariance $\mathbf{\Sigma}_X$. Define $Y = \mathbf{C}'\mathbf{x}$. The expected value of Y is $\mathrm{E}[Y] = \boldsymbol{\mu}_Y = \mathbf{C}'\mathrm{E}[X] = \mathbf{C}'\boldsymbol{\mu}_X$ and the covariance of Y is $\mathbf{\Sigma}_Y = \mathbf{C}'\mathrm{Cov}[X]\,\mathbf{C} = \mathbf{C}'\mathbf{\Sigma}_X\mathbf{C}$.*

Definition 1.65 (Multivariate Studentization)**.** Let X be an n-dimensional random variable with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, then

$$Z = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{x} - \boldsymbol{\mu}) \tag{1.42}$$

is a studentized version of X where $\boldsymbol{\Sigma}^{\frac{1}{2}}$ is a matrix square root such as the Cholesky factor or one based on the spectral decomposition of $\boldsymbol{\Sigma}$. Z has mean $\mathbf{0}$ and covariance equal to the identity matrix $\mathbf{I}_n$.

The final result for vectors relates quadratic forms of normals (inner-products) to χ^2 distributed random variables.

Theorem 1.24 (Quadratic Forms of Normals)**.** *Let X be an n-dimensional normal random variable with mean $\mathbf{0}$ and identity covariance $\mathbf{I}_n$. Then $\mathbf{x}'\mathbf{x} = \sum_{i=1}^{n} x_i^2 \sim \chi_n^2$.*

Combing this result with studentization, when X is a general n-dimensional normal random variable with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$,

$$(\mathbf{x} - \boldsymbol{\mu})' \left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)' \boldsymbol{\Sigma}^{-\frac{1}{2}} (\mathbf{x} - \boldsymbol{\mu})' = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})' \sim \chi_n^2.$$

1.4.8 Monte Carlo and Numerical Integration

Expectations of functions of continuous random variables are integrals against the underlying pdf. In some cases, these integrals are analytically tractable, although in many situations integrals cannot be analytically computed and so numerical techniques are needed to compute expected values and moments.

Monte Carlo is one method to approximate an integral. Monte Carlo utilizes simulated draws from the underlying distribution and averaging to approximate integrals.

Definition 1.66 (Monte Carlo Integration)**.** Suppose $X \sim F(\theta)$ and that it is possible to simulate a series $\{x_i\}$ from $F(\theta)$. The Monte Carlo expectation of a function $g(x)$ is defined

$$\widehat{E[g(X)]} = m^{-1} \sum_{i=1}^{m} g(x_i),$$

Moreover, as long as $\mathrm{E}\left[|g(x)|\right] < \infty$, $\lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} g(x_i) = \mathrm{E}[g(x)]$.

The intuition behind this result follows from the properties of $\{x_i\}$. Since these are i.i.d. draws from $F(\theta)$, they will, on average, tend to appear in any interval $B \in R(X)$ in proportion to the probability $\Pr(X \in B)$. In essence, the simulated values coarsely approximating the discrete approximation shown in 1.8.

While Monte Carlo integration is a general technique, there are some important limitations. First, if the function $g(x)$ takes large values in regions where $\Pr(X \in B)$ is small, it may require a very large number of draws to accurately approximate $\mathrm{E}[g(x)]$ since, by construction, there are unlikely to many points in B. In practice the behavior of $h(x) = g(x) f(x)$ plays an important role in determining the appropriate sample size.[18] Second, while Monte Carlo integration

[18] Monte Carlo integrals can also be seen as estimators, and in many cases standard inference can be used to determine the accuracy of the integral. See Chapter 1 for more details on inference and constructing confidence intervals.

is technically valid for random variables with any number of dimensions, in practice it is usually only reliable when the dimension is small (typically 3 or fewer), especially when the range is unbounded ($R(X) \in \mathbb{R}^n$). When the dimension of X is large, many simulated draws are needed to visit the corners of the (joint) pdf, and if 1,000 draws are sufficient for a unidimensional problem, 1000^n may be needed to achieve the same accuracy when X has n dimensions.

Alternatively the function to be integrated can be approximated using a polygon with an easy-to-compute area, such as the rectangles approximating the normal pdf shown in figure 1.8. The quality of the approximation will depend on the resolution of the grid used. Suppose u and l are the upper and lower bounds of the integral, respectively, and that the region can be split into m intervals $l = b_0 < b_1 < \ldots < b_{m-1} < b_m = u$. Then the integral of a function $h(\cdot)$ is

$$\int_l^u h(x)\,\mathrm{d}x = \sum_{i=1}^{m} \int_{b_{i-1}}^{b_i} h(x)\,\mathrm{d}x.$$

In practice, l and u may be infinite, in which case some cut-off point is required. In general, the cut-off should be chosen to that they vast majority of the probability lies between l and u ($\int_l^u f(x)\,\mathrm{d}x \approx 1$).

This decomposition is combined with an area for approximating the area under h between b_{i-1} and b_i. The simplest is the rectangle method, which uses a rectangle with a height equal to the value of the function at the mid-point.

Definition 1.67 (Rectangle Method)**.** The rectangle rule approximates the area under the curve with a rectangle and is given by

$$\int_l^u h(x)\,\mathrm{d}x \approx h\left(\frac{u+l}{2}\right)(u-l).$$

The rectangle rule would be exact if the function was piece-wise flat. The trapezoid rule improves the approximation by replacing the function at the midpoint with the average value of the function and would be exact for any piece-wise linear function (including piece-wise flat functions).

Definition 1.68 (Trapezoid Method)**.** The trapezoid rule approximates the area under the curve with a trapezoid and is given by

$$\int_l^u h(x)\,\mathrm{d}x \approx \frac{h(u)+h(l)}{2}(u-l).$$

The final method is known as Simpson's rule which is based on using a quadratic approximation to the underlying function. It is exact when the underlying function is piece-wise linear or quadratic.

Definition 1.69 (Simpson's Rule)**.** Simpson's Rule uses an approximation that would be exact if they underlying function were quadratic, and is given by

$$\int_l^u h(x)\,\mathrm{d}x \approx \frac{u-l}{6}\left(h(u)+4h\left(\frac{u+l}{2}\right)+h(l)\right).$$

Example 1.41. Consider the problem of computing the expected payoff of an option. The payoff of a call option is given by

$$c = \max(s_1 - k, 0)$$

where k is the strike price, $s_1$is the stock price at expiration and s_0 is the current stock price. Suppose returns are normally distributed with mean $\mu = .08$ and standard deviation $\sigma = .20$. In this problem, $g(r) = (s_0 \exp(r) - k) I_{[s_0 \exp(r)>k]}$ where $I_{[\cdot]}$ and a binary indicator function which takes the value 1 when the argument is true, and

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

Combined, the function the be integrated is

$$\begin{aligned}\int_{-\infty}^{\infty} h(r)\,\mathrm{d}r &= \int_{-\infty}^{\infty} g(r) f(r)\,\mathrm{d}r \\ &= \int_{-\infty}^{\infty} (s_0 \exp(r) - k) I_{[s_0 \exp(r)>k]} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(r-\mu)^2}{2\sigma^2}\right)\mathrm{d}r\end{aligned}$$

$s_0 = k = 50$ was used in all results.

All four methods were applied to the problem. The number of bins and the range of integration was varied for the analytical approximations. The number of bins ranged across {10,20, 50, 1000} and the integration range spanned $\{\pm 3\sigma, \pm 4\sigma, \pm 6\sigma, \pm 10\sigma\}$ and the bins were uniformly spaced along the integration range. Monte Carlo integration was also applied with $m \in \{100, 1000\}$.

All thing equal, increasing the number of bins increases the accuracy of the approximation. In this example, 50 appears to be sufficient. However, having a range which is too small produces values which differ from the correct value of 7.33. The sophistication of the method also improves the accuracy, especially when the number of nodes is small. The Monte Carlo results are also close, on average. However, the standard deviation is large, about 5%, even when 1000 draws are used, so that large errors would be commonly encountered and so many more points are needed to ensure that the integral is always accurate.

Shorter Problems

Problem 1.1. Suppose

$$\begin{bmatrix} X \\ U \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_U^2 \end{bmatrix}\right)$$

and $Y = 2X + U$. What is $\mathrm{E}[Y]$ and $\mathrm{V}[Y]$?

Problem 1.2. Show $\mathrm{Cov}[aX + bY, cX + dY] = ac\mathrm{V}[X] + bd\mathrm{V}[Y] + (ad + bc)\mathrm{Cov}[X, Y]$.

Problem 1.3. Show that the two forms of the covariance,

$$\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] \text{ and } \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])]$$

are equivalent when X and Y are continuous random variables.

Problem 1.4. Suppose $\{X_i\}$ is a sequence of random variables where $\mathrm{V}[X_i] = \sigma^2$ for all i, $\mathrm{Cov}[X_i, X_{i-1}] = \theta$ and $\mathrm{Cov}\left[X_i, X_{i-j}\right] = 0$ for $j > 1$. What is $\mathrm{V}\left[\bar{X}\right]$ where $\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$?

Problem 1.5. Suppose $Y = \beta X + \epsilon$ where $X \sim N\left(\mu_X, \sigma_X^2\right)$, $\epsilon \sim N\left(0, \sigma^2\right)$ and X and ϵ are independent. What is $\mathrm{Corr}[X, Y]$?

Problem 1.6. Prove that $\mathrm{E}[a + bX] = a + b\mathrm{E}[X]$ when X is a continuous random variable.

Problem 1.7. Prove that $\mathrm{V}[a + bX] = b^2\mathrm{V}[X]$ when X is a continuous random variable.

Problem 1.8. Prove that $\mathrm{Cov}[a + bX, c + dY] = bd\mathrm{Cov}[X, Y]$ when X and Y are a continuous random variables.

Problem 1.9. Prove that $\mathrm{V}[a + bX + cY] = b^2\mathrm{V}[X] + c^2\mathrm{V}[Y] + 2bc\mathrm{Cov}[X, Y]$ when X and Y are a continuous random variables.

Problem 1.10. Suppose $\{X_i\}$ is an i.i.d. sequence of random variables. Show that

$$\mathrm{V}\left[\bar{X}\right] = \mathrm{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = n^{-1}\sigma^2$$

where σ^2 is $\mathrm{V}[X_1]$.

Problem 1.11. Prove that $\mathrm{Corr}[a + bX, c + dY] = \mathrm{Corr}[X, Y]$.

Problem 1.12. Suppose $\{X_i\}$ is a sequence of random variables where, for all i, $\mathrm{V}[X_i] = \sigma^2$, $\mathrm{Cov}[X_i, X_{i-1}] = \theta$ and $\mathrm{Cov}\left[X_i, X_{i-j}\right] = 0$ for $j > 1$.. What is $\mathrm{V}\left[\bar{X}\right]$?

Problem 1.13. Prove that $\mathrm{E}\left[a + bX|Y\right] = a + b\mathrm{E}\left[X|Y\right]$ when X and Y are continuous random variables.

Problem 1.14. Suppose that $\mathrm{E}\left[X|Y\right] = Y^2$ where Y is normally distributed with mean μ and variance σ^2. What is $\mathrm{E}[a + bX]$?

Problem 1.15. Suppose $\mathrm{E}\left[X|Y = y\right] = a + by$ and $\mathrm{V}\left[X|Y = y\right] = c + dy^2$ where Y is normally distributed with mean μ and variance σ^2. What is $\mathrm{V}[X]$?

Problem 1.16. Show that the law of total variance holds for a $\mathrm{V}[X_1]$ when X is a bivariate normal with mean $\boldsymbol{\mu} = [\mu_1\ \mu_2]'$ and covariance

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

Longer Exercises

Exercise 1.1. Sixty percent (60%) of all traders hired by a large financial firm are rated as performing satisfactorily or better in their first-year review. Of these, 90% earned a first in financial econometrics. Of the traders who were rated as unsatisfactory, only 20% earned a first in financial econometrics.

1. What is the probability that a trader is rated as satisfactory or better given they received a first in financial econometrics?

2. What is the probability that a trader is rated as unsatisfactory given they received a first in financial econometrics?

3. Is financial econometrics a useful indicator of trader performance? Why or why not?

Exercise 1.2. Large financial firms use automated screening to detect rogue trades – those that exceed risk limits. One of your colleagues has introduced a new statistical test using the trading data that, given that a trader has exceeded her risk limit, detects this with probability 98%. It also only indicates false positives – that is non-rogue trades that are flagged as rogue – 1% of the time.

1. Assuming 99% of trades are legitimate, what is the probability that a detected trade is rogue? Explain the intuition behind this result.

2. Is this a useful test? Why or why not?

3. How low would the false positive rate have to be to have a 98% chance that a detected trade was actually rogue?

Exercise 1.3. Your corporate finance professor uses a few jokes to add levity to his lectures. Each week he tells 3 different jokes. However, he is also very busy, and so forgets week to week which jokes were used.

1. Assuming he has 12 jokes, what is the probability of 1 repeat across 2 consecutive weeks?

2. What is the probability of hearing 2 of the same jokes in consecutive weeks?

3. What is the probability that all 3 jokes are the same?

4. Assuming the term is 8 weeks long, and they your professor has 96 jokes, what is the probability that there is no repetition during the term? Note that he remembers the jokes he gives in a particular lecture, only forgets across lectures.

5. How many jokes would your professor need to know to have a 99% chance of not repeating any in the term?

Exercise 1.4. A hedge fund company manages three distinct funds. In any given month, the probability that the return is positive is shown in the following table:

$$\begin{array}{ll} \Pr(r_{1,t} > 0) = .55 & \Pr(r_{1,t} > 0 \cup r_{2,t} > 0) = .82 \\ \Pr(r_{2,t} > 0) = .60 & \Pr(r_{1,t} > 0 \cup r_{3,t} > 0) = .7525 \\ \Pr(r_{3,t} > 0) = .45 & \Pr(r_{2,t} > 0 \cup r_{3,t} > 0) = .78 \\ \Pr\left(r_{2,t} > 0 \cap r_{3,t} > 0 | r_{1,t} > 0\right) = .20 & \end{array}$$

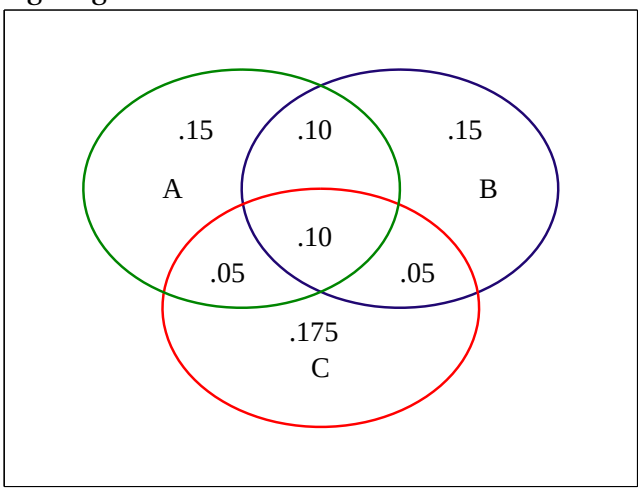
1. Are the events of "positive returns" pairwise independent?

2. Are the events of "positive returns" independent?

3. What is the probability that funds 1 and 2 have positive returns, given that fund 3 has a positive return?

4. What is the probability that at least one fund will have a positive return in any given month?

Exercise 1.5. Suppose the probabilities of three events, A, B and C are as depicted in the following diagram:



1. Are the three events pairwise independent?

2. Are the three events independent?

3. What is $\Pr(A \cap B)$?

4. What is $\Pr\left(A \cap B | C\right)$?

5. What is $\Pr\left(C | A \cap B\right)$?

6. What is $\Pr\left(C | A \cup B\right)$?

Exercise 1.6. At a small high-frequency hedge fund, two competing algorithms produce trades. Algorithm α produces 80 trades per second and 5% lose money. Algorithm β produces 20 trades per second but only 1% lose money. Given the last trade lost money, what is the probability it was produced by algorithm β?

Exercise 1.7. Suppose $f(x, y) = 2 - x - y$ where $x \in [0, 1]$ and $y \in [0, 1]$.

1. What is $\Pr(X > .75 \cap Y > .75)$?

2. What is $\Pr(X + Y > 1.5)$?

3. Show formally whether X and Y are independent.

4. What is $\Pr\left(Y < .5 | X = x\right)$?

Exercise 1.8. Suppose $f(x, y) = x\,y$ for $x \in [0, 1]$ and $y \in [0, 2]$.

1. What is the joint cdf?
2. What is $\Pr(X < 0.5 \cap Y < 1)$?
3. What is the marginal cdf of X? What is $\Pr(X < 0.5)$?
4. What is the marginal density of X?
5. Are X and Y independent?

Exercise 1.9. Suppose $F(x) = 1 - p^{x+1}$ for $x \in [0, 1, 2, \ldots]$ and $p \in (0, 1)$.

1. Find the pmf.
2. Verify that the pmf is valid.
3. What is $\Pr(X \leq 8)$ if $p = .75$?
4. What is $\Pr(X \leq 1)$ given $X \leq 8$?

Exercise 1.10. A firm producing widgets has a production function $q(L) = L^{0.5}$ where L is the amount of labor. Sales prices fluctuate randomly and can be \$10 (20%), \$20 (50%) or \$30 (30%). Labor prices also vary and can be \$1 (40%), 2 (30%) or 3 (30%). The firm always maximizes profits after seeing both sales prices and labor prices.

1. Define the distribution of profits possible?
2. What is the probability that the firm makes at least \$100?
3. Given the firm makes a profit of \$100, what is the probability that the profit is over \$200?

Exercise 1.11. A fund manager tells you that her fund has non-linear returns as a function of the market and that his return is $r_{i,t} = 0.02 + 2r_{m,t} - 0.5r_{m,t}^2$ where $r_{i,t}$ is the return on the fund and $r_{m,t}$ is the return on the market.

1. She tells you her expectation of the market return this year is 10%, and that her fund will have an expected return of 22%. Can this be?
2. At what variance is would the expected return on the fund be negative?

Exercise 1.12. For the following densities, find the mean (if it exists), variance (if it exists), median and mode, and indicate whether the density is symmetric.

1. $f(x) = 3x^2$ for $x \in [0, 1]$
2. $f(x) = 2x^{-3}$ for $x \in [1, \infty)$
3. $f(x) = \left[\pi\left(1 + x^2\right)\right]^{-1}$ for $x \in (-\infty, \infty)$
4. $f(x) = \begin{pmatrix} 4 \\ x \end{pmatrix} .2^x .8^{4-x}$ for $x \in \{0, 1, 2, 3, 4\}$

Exercise 1.13. The daily price of a stock has an average value of £2. Then then $\Pr(X > 10) < .2$ where X denotes the price of the stock. True or false?

Exercise 1.14. An investor can invest in stocks or bonds which have expected returns and covariances as

$$\boldsymbol{\mu} = \begin{bmatrix} .10 \\ .03 \end{bmatrix}, \ \boldsymbol{\Sigma} = \begin{bmatrix} .04 & -.003 \\ -.003 & .0009 \end{bmatrix}$$

where stocks are the first component.

1. Suppose the investor has £1,000 to invest and splits the investment evenly. What is the expected return, standard deviation, variance and Sharpe Ratio (μ/σ) for the investment?

2. Now suppose the investor seeks to maximize her expected utility where her utility is defined is defined in terms of her portfolio return, $U(r) = \mathrm{E}[r] - .01\mathrm{V}[r]$. How much should she invest in each asset?

Exercise 1.15. Suppose $f(x) = (1-p)^x p$ for $x \in (0, 1, \ldots)$ and $p \in (0, 1]$. Show that a random variable from the distribution is "memoryless" in the sense that $\Pr\left(X \geq s + r | X \geq r\right) = \Pr(X \geq s)$. In other words, the probability of surviving s or more periods is the same whether starting at 0 or after having survived r periods.

Exercise 1.16. Your Economics professor offers to play a game with you. You pay £1,000 to play and your Economics professor will flip a fair coin and pay you 2^x where x is the number of tries required for the coin to show heads.

1. What is the pmf of X?

2. What is the expected payout from this game?

Exercise 1.17. Consider the roll of a fair pair of dice where a roll of a 7 or 11 pays $2x$ and anything else pays $-x$ where x is the amount bet. Is this game fair?

Exercise 1.18. Suppose the joint density function of X and Y is given by $f(x, y) = \frac{1}{2}x \exp(-x y)$ where $x \in [3, 5]$ and $y \in (0, \infty)$.

1. Give the form of $\mathrm{E}\left[Y | X = x\right]$.

2. Graph the conditional expectation curve.

Exercise 1.19. Suppose a fund manager has $10,000 of yours under management and tells you that the expected value of your portfolio in two years time is $30,000 and that with probability 75% your investment will be worth at least $40,000 in two years time.

1. Do you believe her?

2. Next, suppose she tells you that the standard deviation of your portfolio value is 2,000. Assuming this is true (as is the expected value), what is the most you can say about the probability your portfolio value falls between $20,000 and $40,000 in two years time?

Exercise 1.20. Suppose the joint probability density function of two random variables is given by $f(x) = \frac{2}{5}(3x + 2y)$ where $x \in [0,1]$ and $y \in [0,1]$.

1. What is the marginal probability density function of X?
2. What is $\mathrm{E}\left[X|Y = y\right]$? Are X and Y independent? (Hint: What must the form of $\mathrm{E}\left[X|Y\right]$ be when they are independent?)

Exercise 1.21. Let Y be distributed χ^2_{15}.

1. What is $\Pr(y > 27.488)$?
2. What is $\Pr(6.262 \leq y \leq 27.488)$?
3. Find C where $\Pr(y \geq c) = \alpha$ for $\alpha \in \{0.01, 0.05, 0.01\}$.
   Next, Suppose Z is distributed χ^2_5 and is independent of Y .
4. Find C where $\Pr(y + z \geq c) = \alpha$ for $\alpha \in \{0.01, 0.05, 0.01\}$.

Exercise 1.22. Suppose X is a bivariate random variable with parameters

$$\boldsymbol{\mu} = \begin{bmatrix} 5 \\ 8 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix}.$$

1. What is $\mathrm{E}\left[X_1|X_2\right]$?
2. What is $\mathrm{V}\left[X_1|X_2\right]$?
3. Show (numerically) that the law of total variance holds for X_2.

Exercise 1.23. Suppose $y \sim N(5, 36)$ and $x \sim N(4, 25)$ where X and Y are independent.

1. What is $\Pr(y > 10)$?
2. What is $\Pr(-10 < y < 10)$?
3. What is $\Pr(x - y > 0)$?
4. Find C where $\Pr(x - y > C) = \alpha$ for $\alpha \in \{0.10, 0.05, 0.01\}$?

Rectangle Method

Bins	$\pm 3\sigma$	$\pm 4\sigma$	$\pm 6\sigma$	$\pm 10\sigma$
10	7.19	7.43	7.58	8.50
20	7.13	7.35	7.39	7.50
50	7.12	7.33	7.34	7.36
1000	7.11	7.32	7.33	7.33

Trapezoid Method

Bins	$\pm 3\sigma$	$\pm 4\sigma$	$\pm 6\sigma$	$\pm 10\sigma$
10	6.96	7.11	6.86	5.53
20	7.08	7.27	7.22	7.01
50	7.11	7.31	7.31	7.28
1000	7.11	7.32	7.33	7.33

Simpson's Rule

Bins	$\pm 3\sigma$	$\pm 4\sigma$	$\pm 6\sigma$	$\pm 10\sigma$
10	7.11	7.32	7.34	7.51
20	7.11	7.32	7.33	7.34
50	7.11	7.32	7.33	7.33
1000	7.11	7.32	7.33	7.33

Monte Carlo

Draws (m)	100	1000
Mean	7.34	7.33
Std. Dev.	0.88	0.28

Table 1.1: Computed values for the expected payout for an option, where the correct value is 7.33 The top three panels use approximations to the function which have simple to compute areas. The bottom panel shows the average and standard deviation from a Monte Carlo integration where the number of points varies and 10,000 simulations were used.