Properties of probability

To achieve a mathematical formulation of the properties of probability, we start with a sample space S. We regard S as the set of all possible outcomes of an experiment. For an experiment in which a coin is tossed once, we can let $S = \{\mathrm{H}, \mathrm{T}\}$. For an experiment in which a coin is tossed twice, we can let $S = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$. In assigning probabilities to the possible outcomes of the experiment, certain axioms should be satisfied.

Finite sample spaces

Suppose S has only a finite number of elements; that is, suppose the experiment has only a finite number of possible outcomes. Write $S = \{s_1, s_2, \ldots, s_N\}$. Let $\mathrm{P}(s_i)$ denote the probability assigned to s_i, that is, the probability that the experiment will result in outcome s_i. An assignment of probabilities to the outcomes should satisfy the following two requirements.

(N.1.1) (a) $0 \leq \mathrm{P}(s_i) \leq 1$ for all $i = 1, \ldots, N$.

(b) $\sum_{i=1}^{N} \mathrm{P}(s_i) = 1$.

(Given requirement (b), we could omit " ≤ 1 " from requirement (a).)

A subset of S is called an *event*. Events of interest often have concise descriptions. For example, in the sample space $S = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$ for tossing a coin twice, the event $A = \{\mathrm{HH}, \mathrm{TH}\}$ can be described as the event of getting a head on the second toss. The probability of an event is given by the sum of the probabilities of its elements.

Definition N.1.2. For an event A in a finite or countably infinite sample space,

$$\mathrm{P}(A) = \sum_{s \in A} \mathrm{P}(s).$$

Probabilities of events satisfy the following three basic properties (see Definition CB.1.2.2). These properties are called the Axioms of Probability.

(N.1.3) (a) $\mathrm{P}(A) \geq 0$ for all events A.

(b) $\mathrm{P}(S) = 1$.

(c) If A and B are disjoint events, then $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$.

By mathematical induction, property c implies

(d) If $A_1, A_2, \ldots, A_k$ are pairwise disjoint events, then $\mathrm{P}(A_1 \cup A_2 \cup \cdots \cup A_k) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \cdots + \mathrm{P}(A_k)$.

Everyone seems to accept the axioms for finite sample spaces as being perfectly sensible. But for infinite sample spaces there is some disagreement about what the axioms should be.

Countably infinite sample spaces

Suppose the number of elements in the sample space S is countably infinite. For example, consider an experiment in which a vial of water is selected from a lake and the number of microorganisms of a certain type are counted. One could figure out an upper bound on what this number could be (e.g., the estimated number of atoms in the universe), but it is convenient to let the sample space be all nonnegative integers, $S = \{0, 1, 2, \dots\}$. In general write $S = \{s_1, s_2, s_3, \dots\}$. An assignment of probabilities to the outcomes should satisfy the following two requirements.

(N.1.4) (a) $0 \le \mathrm{P}(s_i) \le 1$ for all $i = 1, 2, 3, \dots$.

(b) $\sum\limits_{i=1}^{\infty} \mathrm{P}(s_i) = 1$.

As with finite sample spaces, the term "event" simply means a subset of S and the probability of an event A is given by Definition N.1.2.

Probabilities of events in a countably infinite sample space satisfy the Axioms of Probability (N.1.3)(a, b, c, d) and also satisfy (see Definition CB.1.2.2(3)):

(N.1.5) If $A_1, A_2, A_3, \dots$ are pairwise disjoint events, then $\mathrm{P}(\bigcup\limits_{i=1}^{\infty} A_i) = \sum\limits_{i=1}^{\infty} \mathrm{P}(A_i)$.

Property (N.1.3)(d), or equivalently property (N.1.3)(c), is called the Axiom of Finite Additivity. Property (N.1.5) is called the Axiom of Countable Additivity and is a stronger property in the sense that countable additivity implies finite additivity but not vice versa.

The properties of probability in countably infinite sample spaces are a straightforward extension of the properties in finite sample spaces. But the extension to uncountably infinite sample spaces is not so straightforward.

Uncountably infinite sample spaces

Suppose the number of elements in the sample space S is uncountably infinite. For example, consider an experiment in which a fish is selected from a lake and its length is measured. The length could be any value in an interval of possible values, say from 1 centimeter to 100 centimeters. The sample space $S = [1, 100]$ contains an uncountably infinite number of elements. (In a case like this it is often convenient not to worry about lower and upper bounds and to let the sample space be $S = (0, \infty)$ or $S = (-\infty, \infty)$.)

In finite and countably infinite sample spaces it is sufficient to assign probabilities to the individual outcomes, making sure that the requirements (N.1.1) or (N.1.4) are satisfied, and then the probability of any event is determined as the sum of the probabilities of the outcomes it contains. This doesn't work for uncountably infinite sample spaces. In fact, in many cases the probability of every individual outcome is 0. For example, the uniform distribution on $(0,1)$ assigns probability $P(x)=0$ to each individual $x \in (0,1)$ and assigns probability $P((0,1))=1$ to the whole interval. Somehow the individual 0's add up to 1. Infinity is a very useful mathematical concept but it sometimes leads to strange unintuitive results.

So in an uncountably infinite sample space we must assign probabilities to events in some other way. Typically we use probability density functions (pdf's). We will review pdf's later. Probabilities defined in this way still satisfy the Axioms of Probability (N.1.3) and (N.1.5), provided we restrict the assignment of probabilities to a special collection of subsets of S. The subsets to which probabilities are assigned are called *events*. Strangely, in an uncountably infinite sample space, not all subsets of the sample space can be events. If we try to assign probabilities to all subsets of S, the Axioms cannot hold.

In an interval of real numbers, the events are all subintervals and sets that can be constructed from subintervals by taking countable unions, countable intersections, and complements.

Consequences of the axioms

From the Axioms of Probability many other properties can be derived.

Lemma N.1.6 (see C&B §1.2.2).

(a) $P(\phi)=0$ where ϕ is the empty set.

(b) $P(A^c)=1-P(A)$.

(c) $P(A)\leq 1$.

(d) If $A\subset B$, then $P(A)\subset P(B)$.

(e) $P(A\cup B)=P(A)+P(B)-P(A\cap B)$.

(f) $P(A\cup B)\leq P(A)+P(B)$.

(g) $P(A\cap B)\geq P(A)+P(B)-1$.

(h) $P(A_1\cup\cdots\cup A_k)\leq P(A_1)+\cdots+P(A_k)$.

(i) $P(A_1\cap\cdots\cap A_k)\geq P(A_1)+\cdots+P(A_k)-(k-1)$.

Parts (g) and (i) are called the Bonferroni Inequality.

Application. Suppose (L_1, U_1) is a 95% confidence interval for a parameter θ_1 and (L_2, U_2) is a 95% confidence interval for a parameter θ_2. This means that $\mathrm{P}\{L_1 \le \theta_1 \le U_1\} = 0.95$ and $\mathrm{P}\{L_2 \le \theta_2 \le U_2\} = 0.95$. By the Bonferroni Inequality (part (g)), $\mathrm{P}\{L_1 \le \theta_1 \le U_1$ and $L_2 \le \theta_2 \le U_2\} \ge 0.95 + 0.95 - 1 = 0.90$. Thus we can say that we have at least 90% confidence that both intervals contain the true values of their respective parameters. We call the two intervals 90% simultaneous confidence intervals.

Finite sample spaces with equally likely outcomes

There are situations in which all the outcomes in the sample space have the same probability of occurring. When we toss a fair coin, each of the two sides has probability $1/2$ of being the side facing up. When we roll a balanced die, each of the six sides has probability $1/6$ of being the side facing up. When we draw a card at random from a 52-card deck, each card has probability $1/52$ of being drawn.

Write $S = \{s_1, s_2, \dots, s_N\}$. Suppose $\mathrm{P}(s_i) = 1/N$ for all $i = 1, 2, \dots, N$.
For any event $A \subset S$,

$$\mathrm{P}(A) = \sum_{s \in A} \mathrm{P}(s) = \sum_{s \in A} \frac{1}{N} = \frac{\#\text{ elements in } A}{N} .$$

= proportion of elts of S that are in A

Therefore calculation of a probability is a matter of counting the number of elements in a set.

Theorem CB.1.2.4 (Fundamental Theorem of Counting). If a job consists of k separate tasks and the i-th task can be done in n_i ways ($i = 1, \dots, k$), then the entire job can be done in $n_1 n_2 \cdots n_k$ ways.

Lemma N.1.7. Consider a set having n elements.

(a) The number of subsets of the set that have k elements is $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ $\frac{n(n-1)(n-2)\cdots(n-k+2)(n-k+1)}{k(k-1)(k-2)\cdots\cdots\cdots(2)(1)}$.

(b) The number of sequences of distinct elements from the set that have length k is $n(n-1)(n-2)\cdots(n-k+2)(n-k+1) = \frac{n!}{(n-k)!}$.

(c) The number of sequences of elements, not necessarily distinct, from the set that have length k is n^k.

Conditional probability and independence

Definition CB.1.3.1. The conditional probability of A given B is $\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$

Conditional probability and independence

Definition CB.1.3.1. The *conditional probability* of A given B is $\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$.

To interpret $\mathrm{P}(A)$ we can image repeating the experiment an infinite number of times and observing the sequence of outcomes. Then $\mathrm{P}(A)$ would be the proportion of observed outcomes that are in A. The conditional probability $\mathrm{P}(A|B)$ would be the proportion of observed outcomes that are in A if we restrict our attention only to outcomes that are in B.

If someone performs the experiment but we aren't told anything about the outcome, our degree of belief that the outcome is in A is $\mathrm{P}(A)$. If someone tells us that the outcome is in B, our degree of belief that the outcome is in A is $\mathrm{P}(A|B)$.

For a fixed event B, the conditional probabilities $\mathrm{P}(A|B)$ satisfy the Axioms of Probability (Definition CB.1.2.2). So conditional probabilities can be regarded as probabilities in their own right. In particular, they satisfy all the properties of probability. For example, $\mathrm{P}(A^c|B) = 1 - \mathrm{P}(A|B)$ and $\mathrm{P}(A_1 \cup A_2|B) = \mathrm{P}(A_1|B) + \mathrm{P}(A_2|B) - \mathrm{P}(A_1 \cap A_2|B)$.

From the definition we see that if we know the values of $\mathrm{P}(B)$ and $\mathrm{P}(A \cap B)$, then we can take the ratio and get $\mathrm{P}(A|B)$. Also note that if we know the values of $\mathrm{P}(B)$ and $\mathrm{P}(A|B)$, then we can get $\mathrm{P}(A \cap B)$ as

(*) $$\mathrm{P}(A \cap B) = \mathrm{P}(B)\mathrm{P}(A|B).$$

Similarly, if we know the values of $\mathrm{P}(A)$ and $\mathrm{P}(B|A)$, then we can get $\mathrm{P}(A \cap B)$ as

(N.1.8) $$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B|A).$$

From (*) and (N.1.8) we see that $\mathrm{P}(B)\mathrm{P}(A|B) = \mathrm{P}(A)\mathrm{P}(B|A)$ and hence:

Theorem N.1.9 (Bayes' Theorem). $\mathrm{P}(A|B) = \mathrm{P}(B|A)\frac{\mathrm{P}(A)}{\mathrm{P}(B)}$.

Bayes' Theorem allows us to calculate $\mathrm{P}(A|B)$ if we know the values of $\mathrm{P}(B|A)$, $\mathrm{P}(A)$ and $\mathrm{P}(B)$. For finding the value of $\mathrm{P}(B)$, the following theorem can be useful.

Theorem N.1.10 (Theorem of Total Probability). Suppose S is partitioned as $S = A_1 \cup \cdots \cup A_k$ where the A_i are pairwise disjoint. Then

$$\mathrm{P}(B) = \mathrm{P}(A_1)\mathrm{P}(B|A_1) + \cdots + \mathrm{P}(A_k)\mathrm{P}(B|A_k).$$

Combining the two theorems, we see:

Corollary N.1.11 (Bayes' Rule, C&B p. 21). Suppose S is partitioned as $S = A_1 \cup \cdots \cup A_k$ where the A_i are pairwise disjoint. Then

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1)+\cdots+P(A_k)P(B|A_k)} .$$

Note that for the case $k=2$, Theorem N.1.10 says $P(B) = P(A)P(B|A) + P(A^c)P(B|A^c)$ and Corollary N.1.11 says $P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A)+P(A^c)P(B|A^c)}$.

Definition N.1.12. Events A and B are *independent* if $P(A|B) = P(A)$.

Thus, when the event A is independent of the event B, if someone performs the experiment and tells us that the outcome is in B, our degree of belief that the outcome is in A is still $P(A)$, the same as if we didn't know about B.

There are several equivalent ways to define independence. Using the definition of conditional probability, one can show that:

Lemma N.1.~~12~~13. $P(A|B) = P(A)$ iff $P(A \cap B) = P(A)P(B)$ iff $P(B|A) = P(B)$ iff $P(A|B) = P(A|B^c)$ iff $P(B|A) = P(B|A^c)$.

The second condition in the lemma is the definition of independence in the textbook (Definition CB.1.3.2).

Definition N.1.14. Events $A_1, \ldots, A_k$ are *mutually independent* if, ~~given~~ for any one of these events, A_i and any collection of the other events, $A_{j_1}, \ldots, A_{j_m}$ $(1 \le m \le k-1)$, ~~then~~ the equality $P(A_i|\, A_{j_1} \cap \cdots \cap A_{j_m}) = P(A_i)$ holds.

This is equivalent to Definition CB.1.3.3, which says that, ~~given~~ for any collection of these events, $A_{i_1}, \ldots, A_{i_m}$ $(1 \le m \le k)$, ~~then~~ the equality $P(A_{i_1} \cap \cdots \cap A_{i_m}) = P(A_{i_1})\cdots P(A_{i_m})$ holds.

Random variables

The concept of a random variable can be regarded as equivalent to the concept of probability for sets of real numbers. So there is the same sort of vagueness about what a random variable "really is" that there is about what probability "really is". Roughly speaking, a random variable is a real number that has been randomly generated.

Let X be a random variable. Since it is randomly generated, its behavior must be described in terms of probability. The *probability distribution* (or simply *distribution*) of X is a specification of the probabilities $P(X \in A)$ for all events A in the real line $(-\infty, \infty)$ (that is, sets of real numbers that can be constructed by taking finite or countably infinite unions, intersections and complements of intervals — see Example CB.1.2.2). Of course these probabilities must be specified in such a way as to satisfy the three Axioms of Probability:

1. $P(X \in A) \ge 0$ for all events A.
2. $P(X \in (-\infty, \infty)) = 1$.

3. If $A_1, A_2, \ldots$ are pairwise disjoint events, then $\mathrm{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$.

A random variable X is called *discrete* (and its distribution is called discrete) if there is a finite or countably infinite set C such that $\mathrm{P}(X \in C) = 1$. Examples of some common discrete distributions may be found in C&B on pp. 624-625. For the Binomial distribution we can take $C = \{0, 1, \ldots, n\}$. For the Poisson distribution we can take $C = \{0, 1, 2, \ldots\}$.

A random variable X is called *continuous* (and its distribution is called continuous) if $\mathrm{P}(X = x) = 0$ for all real numbers x. Examples of some common continuous distributions may be found in C&B on pp. 626-629.

By definition, the distribution of a random variable X consists of the values of $\mathrm{P}(X \in A)$ for all events A. But to specify all these probabilities it is not necessary to specify them all explicitly.

Probability mass and density functions

♦ (See Definition CB.1.6.1 and Theorem CB.1.6.1). For a discrete distribution, it suffices to give its *probability mass function* (pmf),

$$f_X(x) = \mathrm{P}(X = x) \text{ for all } x \in (-\infty, \infty)$$

(or for all $x \in C$ with the understanding that $\mathrm{P}(X = x) = 0$ for all $x \notin C$). Given the pmf, the distribution of X is determined by $\mathrm{P}(X \in A) = \sum_{x \in A} f_X(x)$.

For Axiom 1 to hold we need (a) $f_X(x) \geq 0$ for all x.
For Axiom 2 to hold we need (b) $\sum_{x \in C} f_X(x) = 1$.
Axiom 3 automatically follows from the properties of summation.
In order for a function $f(x)$ to be a valid pmf, it is necessary and sufficient for it to satisfy conditions (a) and (b) (with $f(x) = 0$ for all $x \notin C$).

Example. Suppose $X \sim \text{Binomial}(n, p)$. See pp. 624 and 89-92 in C&B.
This is a discrete random variable with pmf $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, \ldots, n$.
The number n is a nonnegative integer and $0 \leq p \leq 1$.
Let's verify that $f(x)$ satisfies properties (a) and (b) for a pmf. It is easy to see that $f(x) \geq 0$ for all x. For (b) we can use the Binomial Theorem (Theorem CB.3.1.1):
$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}$. Apply it with $a = p$ and $b = 1 - p$, noting that
$(p + (1-p))^n = 1^n = 1$. ||

♦ (See Definition CB.1.6.2 and Theorem CB.1.6.1). For a continuous distribution, it usually (see below) suffices to give its *probability density function* (pdf), say $f_X(x)$. Given the pdf, the distribution of X is determined by $\mathrm{P}(X \in A) = \int\limits_A f_X(x)\mathrm{d}x$.

For Axiom 1 to hold we need (a) $f_X(x) \geq 0$ for all x.

For Axiom 2 to hold we need (b) $\int\limits_{-\infty}^{\infty} f_X(x)\mathrm{d}x = 1$.

Axiom 3 automatically follows from the properties of integration.

In order for a function $f(x)$ to be a valid pdf, it is necessary and sufficient for it to satisfy conditions (a) and (b).

Example. Suppose $X \sim \text{Normal}(\mu, \sigma^2)$. See pp. 628 and 103-107 in C&B.

This is a continuous random variable with pdf $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for $-\infty < x < \infty$.

The parameter μ may be any real number and σ may be any positive number.

Let's verify that $f(x)$ satisfies properties (a) and (b) for a pdf. It is easy to see that $f(x) \geq 0$ for all x. Verifying (b) is not so easy.

$$\int\limits_{-\infty}^{\infty} f(x)\mathrm{d}x = \int\limits_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma^2}} \int\limits_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x .$$

Now change the variable of integration from x to $z = (x-\mu)/\sigma$. Note that $x = \sigma z + \mu$ and $\mathrm{d}x = \sigma\mathrm{d}z$. Hence the integral equals

$$\frac{1}{\sqrt{2\pi\sigma^2}} \int\limits_{-\infty}^{\infty} e^{-\frac{z^2}{2}} \sigma\mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} e^{-\frac{z^2}{2}} \mathrm{d}z .$$

Note that the integrand $e^{-\frac{z^2}{2}}$ is symmetric about 0, because z^2 is symmetric about 0. So the integral over the negative half of the real line is equal to the integral over the positive half of the real line. Therefore

$$\int\limits_{-\infty}^{\infty} f(x)\mathrm{d}x = \frac{2}{\sqrt{2\pi}} \int\limits_{0}^{\infty} e^{-\frac{z^2}{2}} \mathrm{d}z .$$

Change the variable of integration from z to $w = z^2/2$. Note that $z = \sqrt{2w}$ and $\mathrm{d}z = \frac{1}{\sqrt{2w}}\mathrm{d}w$. Hence the integral equals

$$\frac{2}{\sqrt{2\pi}} \int\limits_{0}^{\infty} \frac{e^{-w}}{\sqrt{2w}} \mathrm{d}w = \frac{1}{\sqrt{\pi}} \int\limits_{0}^{\infty} \frac{e^{-w}}{\sqrt{w}} \mathrm{d}w .$$

This integral still looks difficult. But it is a well-known integral; it is an instance of the gamma function:

$$\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}\mathrm{d}t\,,$$

which is ~~well-defined~~ finite for all $a > 0$. See C&B p. 100. When $a = n$ is a positive integer, then $\Gamma(n) = (n-1)!$. For $a = \frac{1}{2}$, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. But for most values of a, the integral must be calculated numerically. Note that $e^{-w}/\sqrt{w} = w^{\frac{1}{2}-1}e^{-w}$, and so we have ←.

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \frac{1}{\sqrt{\pi}}\Gamma(\tfrac{1}{2}) = \frac{1}{\sqrt{\pi}}\sqrt{\pi} = 1\,. \;\|$$

A distribution is called *absolutely continuous* if it has a pdf. Almost all continuous distributions are absolutely continuous but there are a few strange exceptions. All the distributions that we will deal with are absolutely continuous and so it will be convenient for us to use the word "continuous" to mean "absolutely continuous".

Cumulative distribution functions

Another way to specify a distribution of a random variable X, besides giving its pmf or pdf, is to give its *cumulative distribution function* (cdf),

$$F_X(x) = \mathrm{P}(X \le x) \text{ for all } x \in (-\infty, \infty)\,.$$

For the Axioms of Probability to hold we require that (a) $F_X(x) \to 0$ as $x \to -\infty$ and $F_X(x) \to 1$ as $x \to \infty$, (b) $F_X(x)$ is a nondecreasing function, and (c) $F_X(x)$ is right-continuous. In order for a function $F(x)$ to be a valid cdf, it is necessary and sufficient for it to satisfy conditions (a), (b) and (c).

For a discrete distribution, the pmf can be obtained from the cdf as $f_X(x) = F_X(x) - F_X(x-\epsilon)$ for a sufficiently small positive value of ϵ.

For a continuous distribution, the pdf can be obtained from the cdf as $f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x}F_X(x)$.

CHAPTER 2 – Expectations and transformations

Expected values

Definition CB.2.2.1. (a) The *expected value* (or *expectation* or *mean*) of a function $g(X)$ of a discrete random variable is

$$\mathrm{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x)$$

where $\mathcal{X}$ is a set such that $\mathrm{P}(X \in \mathcal{X}) = 1$. The sum always exists if $\mathcal{X}$ is finite or if $g(x)$ is bounded, but if $\mathcal{X}$ is countably infinite and $g(x)$ is unbounded, then the sum may not exist, in which case the expected value does not exist.

(b) The *expected value* of a function $g(X)$ of a continuous random variable is

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x .$$

The integral always exists if $g(x)$ is bounded, but if $g(x)$ is unbounded, then the integral may not exist, in which case the expected value does not exist.

Example. Suppose $X \sim \text{Cauchy}(0,1)$. See pp. 626 and 109-110. Its mean $\mathrm{E}(X)$ does not exist — see Example CB.2.2.3. ||

Example. Suppose $X \sim \text{Binomial}(n, p)$. Let's calculate its mean.
$\mathrm{E}(X) = \sum_{x=0}^{n} x f_X(x) = \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=1}^{n} x \binom{n}{x} p^x (1-p)^{n-x}$.
Recall that $\binom{n}{x} = \frac{n!}{x!(n-x)!}$. Hence $x \binom{n}{x} = \frac{n!}{(x-1)!(n-x)!} = n \binom{n-1}{x-1}$.
Now we have $\mathrm{E}(X) = \sum_{x=1}^{n} n \binom{n-1}{x-1} p^x (1-p)^{n-x} = np \sum_{x=1}^{n} \binom{n-1}{x-1} p^{x-1} (1-p)^{\overset{(n-1)-(x-1)}{n-x}}$
$= np \sum_{y=0}^{m} \binom{m}{y} p^y (1-p)^{m-y}$ where $m = n-1$ and $y = x-1$. Note that $\binom{m}{y} p^y (1-p)^{m-y}$ is the pmf of the Binomial(m, p) distribution on $\{0, 1, \ldots, m\}$, and so $\sum_{y=0}^{m} \binom{m}{y} p^y (1-p)^{m-y} = 1$. Therefore, $\mathrm{E}(X) = np \cdot 1 = np$. ||

Example. Suppose $X \sim \text{Normal}(\mu, \sigma^2)$. Let's calculate its mean.

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x \, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x .$$

Again it helps to change the variable of integration from x to $z = (x-\mu)/\sigma$. This yields

$$\mathrm{E}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\sigma z + \mu) e^{-\frac{z^2}{2}} \sigma \mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sigma z + \mu) e^{-\frac{z^2}{2}} \mathrm{d}z$$

$$= \sigma \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} \mathrm{d}z + \mu \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} \mathrm{d}z$$

$$= \sigma \frac{1}{\sqrt{2\pi}} A + \mu B \text{, where } A = \int_{-\infty}^{\infty} z e^{-\frac{z^2}{2}} \mathrm{d}z \text{ and } B = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \mathrm{d}z .$$

Note that $\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ is the pdf of the Normal(0, 1) distribution, and so $B = 1$. Note that the integrand $h(z) = z e^{-\frac{z^2}{2}}$ in A is an odd function, i.e., it satisfies $h(-z) = -h(z)$ for all $z > 0$. For an odd function, $\int_{-\infty}^{\infty} h(z)\mathrm{d}z = \int_{-\infty}^{0} h(z)\mathrm{d}z + \int_{0}^{\infty} h(z)\mathrm{d}z = -\int_{0}^{\infty} h(z)\mathrm{d}z + \int_{0}^{\infty} h(z)\mathrm{d}z$, which is 0 <u>provided</u> that $\int_{0}^{\infty} h(z)\mathrm{d}z$ exists. Now $\int_{0}^{\infty} z e^{-\frac{z^2}{2}} \mathrm{d}z = \left[-e^{-\frac{z^2}{2}} \right]_{z=0}^{z=\infty} = -0 + 1 = 1 < \infty$. Hence $A = 0$ and $\mathrm{E}(X) = \sigma \frac{1}{\sqrt{2\pi}} \cdot 0 + \mu \cdot 1 = \mu$. The symbol μ is commonly used to denote the mean of a distribution. We have shown that the parameter μ in the Normal(μ, σ^2) distribution is consistent with this common usage. ||

Some useful properties of expectation are listed in the following lemma.

<u>Lemma N.2.1</u> (see Theorem CB.2.2.1). Let X be a random variable, c a constant, and $g(X)$ and $h(X)$ functions of X whose expectations exist.

(a) $\mathrm{E}(c) = c$.
(b) $\mathrm{E}(X + c) = \mathrm{E}(X) + c$.
(c) $\mathrm{E}(cX) = c\mathrm{E}(X)$.
(d) $\mathrm{E}[g(X) + h(X)] = \mathrm{E}[g(X)] + \mathrm{E}[h(X)]$.
(e) If $g(X) \leq h(X)$, then $\mathrm{E}[g(X)] \leq \mathrm{E}[h(X)]$.

<u>Moments</u>

<u>Definition CB.2.3.1</u>. (a) The k-th *moment* of X is $\mu'_k = \mathrm{E}(X^k)$.
(b) Let $\mu = \mu'_1 = \mathrm{E}(X)$. The k-th *central moment* of X is $\mu_k = \mathrm{E}[(X - \mu)^k]$.

The *mean* of X is $\mu'_1 = \mathrm{E}(X)$, often denoted by μ or μ_X. It measures the "location" of the distribution of X.
The *variance* of X is $\mu_2 = \mathrm{E}[(X - \mu)^2] = \mathrm{Var}(X)$, often denoted by σ^2 or σ^2_X. It measures the "spread" or "variation" in the distribution of X.
The *skewness* of X is $\mu_3/(\mu_2)^{3/2}$. It measures the "asymmetry" or "lopsidedness" of the distribution of X.
The *kurtosis* of X is $\mu_4/(\mu_2)^2$. It measures the "peakedness" or "heavy-tailedness" of the distribution of X.

The following two formulas are sometimes helpful for calculating the variance of a random variable.

Lemma N.2.2. Let $\mu = \mathrm{E}(X)$.

(a) $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2$.

(b) $\mathrm{Var}(X) = \mathrm{E}[X(X-1)] + \mu - \mu^2$.

Proof of (a): By definition, $\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2]$. Expand $(X-\mu)^2 = X^2 - 2\mu X + \mu^2$. Using Lemma N.2.1, we see $\mathrm{E}(X^2 - 2\mu X + \mu^2) = \mathrm{E}(X^2) + \mathrm{E}(-2\mu X) + \mathrm{E}(\mu^2) = \mathrm{E}(X^2) - 2\mu \mathrm{E}(X) + \mu^2 = \mathrm{E}(X^2) - 2\mu^2 + \mu^2 = \mathrm{E}(X^2) - \mu^2$. □

Example. Suppose $X \sim \mathrm{Binomial}(n, p)$. Let's calculate its variance.
Instead of directly calculating $\mathrm{E}[(X-\mu)^2] = \sum\limits_{x=0}^{n} (x-\mu)^2 f_X(x)$, it is easier to use Lemma N.2.2(b) and to first calculate $\mathrm{E}[X(X-1)] = \sum\limits_{x=0}^{n} x(x-1) f_X(x) = \sum\limits_{x=2}^{n} x(x-1)\binom{n}{x} p^x (1-p)^{n-x}$. Note that $x(x-1)\binom{n}{x} = \frac{n!}{(x-2)!(n-x)!} = n(n-1)\binom{n-2}{x-2}$. Now we have $\mathrm{E}[X(X-1)] = n(n-1)p^2 \sum\limits_{x=2}^{n} \binom{n-2}{x-2} p^{x-2}(1-p)^{n-x}$ $= n(n-1)p^2 \sum\limits_{y=0}^{m} \binom{m}{y} p^y (1-p)^{m-y}$ where $m = n-2$ and $y = x-2$. Since $\binom{m}{y} p^y (1-p)^{m-y}$ is the pmf of the Binomial(m, p) distribution, the sum is equal to 1. Therefore, $\mathrm{E}[X(X-1)] = n(n-1)p^2$. Recall that the mean of X is $\mu = np$. Now $\mathrm{Var}(X) = n(n-1)p^2 + np - (np)^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1-p)$. ||

Example. Suppose $X \sim \mathrm{Normal}(\mu, \sigma^2)$. Let's calculate its variance.
$\mathrm{Var}(X) = \mathrm{E}[(X-\mu)^2] = \int\limits_{-\infty}^{\infty} (x-\mu)^2 f_X(x) dx = \int\limits_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$ $= \frac{1}{\sqrt{2\pi\sigma^2}} \int\limits_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$. Again it helps to change the variable of integration from x to $z = (x-\mu)/\sigma$. This yields $\mathrm{Var}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \int\limits_{-\infty}^{\infty} \sigma^2 z^2 e^{-\frac{z^2}{2}} \sigma dz = \frac{\sigma^2}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz$.

One way to evaluate this integral is to use the method of integration by parts. Another way, which we will use here, is to change the variable of integration to $w = z^2/2$. Then $\int\limits_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}} dz = 2\int\limits_{0}^{\infty} z^2 e^{-\frac{z^2}{2}} dz = 2\int\limits_{0}^{\infty} 2we^{-w}/\sqrt{2w}\, dw = 2\sqrt{2}\int\limits_{0}^{\infty} \sqrt{w} e^{-w} dw$. Note that $\int\limits_{0}^{\infty} w^{\frac{3}{2}-1} e^{-w} dw = \Gamma(\frac{3}{2}) = \frac{1}{2}\Gamma(\frac{1}{2}) = \frac{1}{2}\sqrt{\pi}$. Now $\mathrm{Var}(X) = \frac{\sigma^2}{\sqrt{2\pi}} 2\sqrt{2}\,\frac{1}{2}\sqrt{\pi} = \sigma^2$.

The symbol σ^2 is commonly used to denote the variance of a distribution. We have shown that the parameter σ^2 in the Normal(μ, σ^2) distribution is consistent with this common usage. ||

Some properties of variance are listed in the following lemma.

Lemma N.2.3 (see Theorem CB.2.3.1). Let X be a random variable, c a constant, and $g(X)$ and $h(X)$ functions of X whose variances are finite.

(a) $\text{Var}(c) = 0$.

(b) $\text{Var}(X + c) = \text{Var}(X)$.

(c) $\text{Var}(cX) = c^2\text{Var}(X)$.

(d) $\text{Var}[g(X) + h(X)] = \text{Var}[g(X)] + \text{Var}[h(X)] + 2\text{Cov}[g(X), h(X)]$.

(e) $\text{Var}(X) > 0$ except when $\text{P}(X = c) = 1$ for some constant c.

Moment-generating functions

Definition CB.2.3.1. The *moment-generating function* (mgf) of a random variable X (or of its distribution) is $M_X(t) = \text{E}(e^{tX})$, provided this expectation exists for all t in an open interval containing 0.

The usefulness of moment-generating functions is shown in the following three theorems. The first theorem shows that, true to its name, the mgf of a distribution can be used to generate the moments.

Theorem CB.2.3.2. The k-th moment of X can be obtained by taking the k-th derivative of the mgf and evaluating it at $t = 0$. That is, $\text{E}(X^k) = \frac{d^k}{dt^k} M_X(t)\Big|_{t=0}$.

`Partial justification`: Let us try to convince ourselves that the equation $\text{E}(X) = \frac{d}{dt} M_X(t)\Big|_{t=0}$ makes sense. Starting from the definition $M_X(t) = \text{E}(e^{tX})$, we have $\frac{d}{dt} M_X(t) = \frac{d}{dt}\text{E}(e^{tX})$. Expectation is defined in terms of a sum (for a discrete distribution) or an integral (for a continuous distribution). In "most" cases it is legitimate to differentiate a sum or integral under the summation or integral sign (see Section 2.4 in C&B). Thus $\frac{d}{dt}\text{E}(e^{tX}) = \text{E}(\frac{d}{dt}e^{tX})$. Now $\frac{d}{dt}e^{tX} = Xe^{tX}$, so $\frac{d}{dt}M_X(t) = \text{E}(Xe^{tX})$ and $\frac{d}{dt}M_X(t)\Big|_{t=0} = \text{E}(Xe^0) = \text{E}(X)$. □

Example. Suppose $X \sim \text{Binomial}(n, p)$. (i) Its mgf can be obtained as follows.

$$M_X(t) = \text{E}(e^{tX}) = \sum_{x=0}^{n} e^{tx} f_X(x) = \sum_{x=0}^{n} e^{tx}\binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n}\binom{n}{x}(e^t p)^x (1-p)^{n-x}$$

$= \text{(by the Binomial Theorem) } [e^t p + (1-p)]^n = (pe^t + 1 - p)^n$.

(ii) We can use the mgf to obtain the mean.

First, $\frac{\mathrm{d}}{\mathrm{d}t}M_X(t) = \frac{\mathrm{d}}{\mathrm{d}t}(pe^t+1-p)^n = n(pe^t+1-p)^{n-1}pe^t = np(pe^t+1-p)^{n-1}e^t$.
Then, $\mathrm{E}(X) = \frac{\mathrm{d}}{\mathrm{d}t}M_X(t)\Big|_{t=0} = np(pe^0+1-p)^{n-1}e^0 = np$.

(iii) We can differentiate again to obtain the variance.
First, $\frac{\mathrm{d}^2}{\mathrm{d}t^2}M_X(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left[np(pe^t+1-p)^{n-1}e^t\right]$
$= np\{(n-1)(pe^t+1-p)^{n-2}pe^{2t} + (pe^t+1-p)^{n-1}e^t\}$
$= np(pe^t+1-p)^{n-2}e^t\{(n-1)pe^t + (pe^t+1-p)\}$
$= np(pe^t+1-p)^{n-2}e^t(npe^t+1-p)$.
Then $\mathrm{E}(X^2) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}M_X(t)\Big|_{t=0} = np(pe^0+1-p)^{n-2}e^0(npe^0+1-p) = np(np+1-p)$.
Now $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = np(np+1-p) - (np)^2 = np(1-p)$. ||

<u>Example</u>. Suppose $X \sim \text{Normal}(\mu,\sigma^2)$. (i) Its mgf is:
$M_X(t) = \mathrm{E}(e^{tX}) = \int\limits_{-\infty}^{\infty} e^{tx}f_X(x)\mathrm{d}x = \int\limits_{-\infty}^{\infty} e^{tx}\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x = \frac{1}{\sqrt{2\pi\sigma^2}}\int\limits_{-\infty}^{\infty} e^{tx-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x$.
The trick to evaluating this integral is to "complete the square" in the exponent. That is, write
$tx - \frac{1}{2\sigma^2}(x-\mu)^2 = -\frac{1}{2\sigma^2}(-2\sigma^2tx + x^2 - 2\mu x + \mu^2) = -\frac{1}{2\sigma^2}[x^2 - 2(\sigma^2t+\mu)x + \mu^2]$
$= -\frac{1}{2\sigma^2}[\{x-(\sigma^2t+\mu)\}^2 - (\sigma^2t+\mu)^2 + \mu^2] = \cdots = -\frac{1}{2\sigma^2}(x-\mu^*)^2 + (\frac{1}{2}\sigma^2t^2+\mu t)$
where $\mu^* = \sigma^2t+\mu$. Now $M_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}}\int\limits_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}(x-\mu^*)^2+(\frac{1}{2}\sigma^2t^2+\mu t)}\,\mathrm{d}x$
$= e^{\frac{1}{2}\sigma^2t^2+\mu t}\int\limits_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(x-\mu^*)^2}\,\mathrm{d}x$. Note that the integrand is the pdf of the Normal(μ^*,σ^2) distribution, and so the integral is 1. Therefore $M_X(t) = e^{\mu t+\frac{1}{2}\sigma^2t^2}$.
(ii) The mgf can be used to calculate the mean. First, $\frac{\mathrm{d}}{\mathrm{d}t}M_X(t) = \frac{\mathrm{d}}{\mathrm{d}t}(e^{\mu t+\frac{1}{2}\sigma^2t^2})$
$= (\mu+\sigma^2t)e^{\mu t+\frac{1}{2}\sigma^2t^2}$, and then $\mathrm{E}(X) = \frac{\mathrm{d}}{\mathrm{d}t}M_X(t)\Big|_{t=0} = (\mu+\sigma^2 0)e^{\mu 0+\frac{1}{2}\sigma^2 0^2} = \mu$.
(iii) To get the variance, we first calculate $\frac{\mathrm{d}^2}{\mathrm{d}t^2}M_X(t) = \frac{\mathrm{d}}{\mathrm{d}t}\left[(\mu+\sigma^2t)e^{\mu t+\frac{1}{2}\sigma^2t^2}\right]$
$= \sigma^2e^{\mu t+\frac{1}{2}\sigma^2t^2} + (\mu+\sigma^2t)^2e^{\mu t+\frac{1}{2}\sigma^2t^2}$. Then $\mathrm{E}(X^2) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}M_X(t)\Big|_{t=0}$
$= \sigma^2e^{\mu 0+\frac{1}{2}\sigma^2 0^2} + (\mu+\sigma^2 0)^2e^{\mu 0+\frac{1}{2}\sigma^2 0^2} = \sigma^2+\mu^2$. So $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$
$= (\sigma^2+\mu^2) - \mu^2 = \sigma^2$. ||

The next theorem, which will be useful in Chapter 4, says that a distribution is completely characterized by its mgf (if it exists).

<u>Theorem CB.2.3.3</u>. If two distributions have the same mgf, the two distributions must be equal. That is, if $M_X(t) = M_Y(t)$ for all t in an open interval containing 0, then $\mathrm{P}(X \in A) = \mathrm{P}(Y \in A)$ for all events A. ←

By applying the preceding theorem together with the following theorem, mgf's can be used to find the limiting distribution of a sequence of distributions.

Theorem CB.2.3.4. If a sequence of mgf's converges to an mgf, then the sequence of distributions converges "in distribution" to the distribution determined by the limiting mgf. That is, if $M_{X_i}(t) \longrightarrow M_X(t)$ as $i \longrightarrow \infty$ for all t in an open interval containing 0, then $\mathrm{P}(X_i \in A) \longrightarrow \mathrm{P}(X \in A)$ as $i \longrightarrow \infty$ for "almost" all events A.

If X has a continuous distribution, then the probabilities converge for all events A.
If X has a discrete distribution, then for technical reasons, convergence is not required for intervals with endpoints having positive probability, such as $A = (-\infty, a]$ when $\mathrm{P}(X = a) > 0$.
This theorem will be used in Chapter 5 to prove the Central Limit Theorem.

Example. You may know from other courses that for large n a binomial distribution can be approximated by a normal distribution. A mathematical justification of this approximation can be obtained from Theorem CB.2.3.4. Let $X_n \sim \mathrm{Binomial}(n, p)$. Standardize it to obtain $Z_n = (X_n - \mu_n)/\sigma_n$ where $\mu_n = \mathrm{E}(X_n) = np$ and $\sigma_n = \mathrm{SD}(X_n) = \sqrt{np(1-p)}$. Let $Z \sim \mathrm{Normal}(0,1)$. It can be shown (but it's not easy) that $M_{Z_n}(t) \longrightarrow M_Z(t)$ as $n \longrightarrow \infty$ for all t. ||

The distribution of a function of a random variable

If X is a random variable and $g(x)$ is a function, then $Y = g(X)$ is also a random variable. The distribution of X determines the distribution of Y. If we know the distribution of X, this means that we can obtain the value of $\mathrm{P}(X \in A)$ for any event A. We can also obtain the value of $\mathrm{P}(Y \in A)$ for any event A. Let $B = \{x : g(x) \in A\}$. For any function $g(x)$ that occurs in the standard theory of statistics, the set B is also an event. Note that $\mathrm{P}(Y \in A) = \mathrm{P}(X \in B)$.

Discrete distributions

Proposition N.2.4. Suppose X has a discrete distribution with pmf $f_X(x)$. Suppose $Y = g(X)$. Then Y is a discrete random variable with pmf $f_Y(y) = \sum_{x \in B_y} f_X(x)$ where $B_y = \{x : g(x) = y\}$.

This is simply saying that $\mathrm{P}(Y = y) = \mathrm{P}[g(X) = y] = \mathrm{P}(X \in B_y) = \sum_{x \in B_y} \mathrm{P}(X = x)$.

We can restrict attention to a set $\mathcal{X}$ such that $\mathrm{P}(X \in \mathcal{X}) = 1$. If we let $\mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}$, then $\mathrm{P}(Y \in \mathcal{Y}) = 1$.
The expected value of $Y = g(X)$ can be computed in two ways:
(1) $\mathrm{E}(Y) = \sum_{y \in \mathcal{Y}} y f_Y(y)$ or (2) $\mathrm{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x) f_X(x)$.

By substituting the expression for $f_Y(y)$ in Proposition N.2.4 into (1), one can verify that (1) and (2) yield the same answer. Given the pmf of X, unless you are interested in the pmf of $Y = g(X)$ for some other reason, it is typically easier to calculate $\mathrm{E}[g(X)]$ using (2) rather than (1).

Example. Suppose X is a discrete random variable with pmf given below:

x	-2	-1	0	1	2	3
$f_X(x)$	0.02	0.16	0.25	0.30	0.21	0.06

Let $Y = X^2$. Let $\mathcal{X} = \{-2, -1, 0, 1, 2, 3\}$. Then $\mathcal{Y} = \{0, 1, 4, 9\}$.
Applying Proposition N.2.4, we get:

$B_0 = \{x \in \mathcal{X} : x^2 = 0\} = \{0\}$, $f_Y(0) = f_X(0) = 0.25$;
$B_1 = \{x : x^2 = 1\} = \{-1, 1\}$, $f_Y(1) = f_X(-1) + f_X(1) = 0.16 + 0.30 = 0.46$;
$B_4 = \{x : x^2 = 4\} = \{-2, 2\}$, $f_Y(4) = f_X(-2) + f_X(2) = 0.02 + 0.21 = 0.23$;
$B_9 = \{x : x^2 = 9\} = \{3\}$, $f_Y(9) = f_X(3) = 0.06$.

The value of $\mathrm{E}(Y) = \mathrm{E}(X^2)$ can be calculated as either
(1) $\mathrm{E}(Y) = 0(0.25) + 1(0.46) + 4(0.23) + 9(0.06) = 1.92$, or
(2) $\mathrm{E}(X^2) = (-2)^2(0.02) + (-1)^2(0.16) + 0^2(0.25) + 1^2(0.30) + 2^2(0.21) + 3^2(0.06)$
$= 1.92$. ||

Continuous distributions

To deal with a function of a continuous random variable, we should keep track of the support of the distribution. The *support* of a distribution with pmf or pdf $f_X(x)$ is the set $\mathcal{X} = \{x : f_X(x) > 0\}$.

Theorem CB.2.1.1. Suppose X has cdf $F_X(x)$ and support $\mathcal{X} = (a, b)$ where a could ~~can~~ be $-\infty$ and b could ~~can~~ be ∞. Suppose $Y = g(X)$ where $g(x)$ is a strictly increasing function on (a, b). Then the cdf of Y is $\in$.

$$F_Y(y) = \begin{cases} 0 & \text{for } y \le g(a) \\ F_X(g^{-1}(y)) & \text{for } g(a) < y < g(b) \\ 1 & \text{for } y \ge g(b) \end{cases}.$$

This is simply saying that $\mathrm{P}(Y \le y) = \mathrm{P}[g(X) \le y] = \mathrm{P}[X \le g^{-1}(y)]$.

When $g(x)$ is a one-to-one function (which is true when it is strictly increasing) the notation $g^{-1}(y)$ is used for the unique value of x such that $g(x) = y$. That is, if the equation $g(x) = y$ is solved for x, then the solution constitutes the equation $x = g^{-1}(y)$.

Theorem CB.2.1.2. Suppose X has a continuous distribution with pdf $f_X(x)$ and support $\mathcal{X} = (a, b)$. ~~Suppose $Y = g(X)$ where $g(x)$ is a strictly increasing function on (a, b).~~

(a) Suppose $Y = g(X)$ where $g(x)$ is a strictly increasing function on $(a\,,b)$ whose inverse function is continuously differentiable. Then Y has a continuous distribution with pdf

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \cdot \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) & \text{for } g(a) < y < g(b) \\ 0 & \text{otherwise} \end{cases} .$$

(b) (postponed until later)

`Justification`: Recall that $f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y)$. Using the preceding theorem we see $\frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y))$. Now use the chain rule of differentiation: $\frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)\Big|_{x=g^{-1}(y)} \cdot \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y)$. Note that $\frac{\mathrm{d}}{\mathrm{d}x} F_X(x) = f_X(x)$. □

This is an important formula. It may be easier to remember as

$$f_Y(y) = f_X(x)\frac{\mathrm{d}x}{\mathrm{d}y} \quad \text{where } x = g^{-1}(y) .$$

One way to remember to write $\frac{\mathrm{d}x}{\mathrm{d}y}$ rather than $\frac{\mathrm{d}y}{\mathrm{d}x}$ is to remember that $\int f_Y(y)\mathrm{d}y = \int f_X(x)\frac{\mathrm{d}x}{\mathrm{d}y}\mathrm{d}y = \int f_X(x)\mathrm{d}x = 1$, where the $\mathrm{d}y$'s "cancel".

Example. Suppose $X \sim \text{Uniform}(0\,,1)$. Its pdf is $f_X(x) = 1$ for $0 < x < 1$ (and $= 0$ otherwise). Let $Y = X^2$. The pdf of Y is given by $f_Y(y) = f_X(x)\frac{\mathrm{d}x}{\mathrm{d}y}$. Solve $y = x^2$ for $x = \sqrt{y}$. Now $\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{\mathrm{d}}{\mathrm{d}y}\sqrt{y} = \frac{\mathrm{d}}{\mathrm{d}y} y^{\frac{1}{2}} = \frac{1}{2} y^{-\frac{1}{2}} = \frac{1}{2\sqrt{y}}$ and so $f_Y(y) = 1 \cdot \frac{1}{2\sqrt{y}} = \frac{1}{2\sqrt{y}}$ for $0^2 < y < 1^2$, that is, for $0 < y < 1$ (and $= 0$ otherwise). ||

The expected value of $Y = g(X)$ can be computed in two ways: (1) $\mathrm{E}(Y) = \int y f_Y(y)\mathrm{d}y$ or (2) $\mathrm{E}[g(X)] = \int g(x) f_X(x)\mathrm{d}x$. Formula (1) is what you get from (2) by changing the variable of integration from x to $y = g(x)$. Given the pdf of X, unless you are interested in the pdf of $Y = g(X)$ for some other reason, it is often easier to calculate $\mathrm{E}[g(X)]$ using (2) rather than (1).

Note:
Y may not h a cont. dist ⟸

Example. As in the preceding example, suppose $X \sim \text{Uniform}(0\,,1)$. Its pdf is $f_X(x) = 1$ for $0 < x < 1$. To calculate the expectation of X^2, we can proceed as in (2):

$$\mathrm{E}(X^2) = \int x^2 f_X(x)\mathrm{d}x = \int_0^1 x^2\mathrm{d}x = \frac{x^3}{3}\Bigg|_{x=0}^{x=1} = \frac{1}{3} .$$

Alternatively, we could use the pdf of $Y = X^2$ that we derived above. Recall that $f_Y(y) = \frac{1}{\sqrt{2y}}$ for $0 < y < 1$. So, as in (1),

$$\mathrm{E}(Y) = \int y f_Y(y)\mathrm{d}y = \int_0^1 y\frac{1}{2\sqrt{y}}\mathrm{d}y = \frac{1}{2}\int_0^1 \sqrt{y}\,\mathrm{d}y = \frac{1}{2} \cdot \frac{2}{3} y^{\frac{3}{2}}\Bigg|_{y=0}^{y=1} = \frac{1}{3} . \;||$$

The following lemma can be useful for calculating the derivative $\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)$ that occurs in the formula in the theorem above.

Lemma N.2.5. If $g(x)$ has a positive derivative for $a < x < b$, then:

(a) it has an inverse function $g^{-1}(y)$ defined for $g(a) < y < g(b)$,

(b) $g^{-1}(y)$ has a positive derivative for $g(a) < y < g(b)$,

(c) $\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y) = 1\big/g'(g^{-1}(y))$ where $g'(x) = \frac{\mathrm{d}}{\mathrm{d}x}g(x)$.

For short, part (c) can be written as $\frac{\mathrm{d}x}{\mathrm{d}y} = 1\big/\frac{\mathrm{d}y}{\mathrm{d}x}$.

Theorem CB.2.1.4. If X is a continuous random variable with support $\mathcal{X} = (a\,,b)$, then:

(a) its cdf $F_X(x)$ is strictly increasing on $(a\,,b)$,

(b) $Y = F_X(X) \sim \text{Uniform}(0\,,1)$,

(c) if $U \sim \text{Uniform}(0\,,1)$, then $F_X^{-1}(U)$ has the same distribution as X.

`Justification:` (a) For $a < x < b$, that is, for x in the support, we have $\frac{\mathrm{d}}{\mathrm{d}x}F_X(x) = f_X(x) > 0$, which implies that $F_X(x)$ is strictly increasing.
(b) By Theorem CB.2.1.2, $f_Y(y) = f_X(F_X^{-1}(y)) \cdot \frac{\mathrm{d}}{\mathrm{d}y}F_X^{-1}(y)$ for $F_X(a) < y < F_X(b)$. Since $(a\,,b)$ is the support of X, $F_X(a) = 0$ and $F_X(b) = 1$. By Lemma N.2.5(c), $\frac{\mathrm{d}}{\mathrm{d}y}F_X^{-1}(y) = 1\big/F_X'(F_X^{-1}(y)) = 1\big/f_X(F_X^{-1}(y))$. Hence $f_Y(y) = f_X(F_X^{-1}(y))\big/f_X(F_X^{-1}(y))$ $= 1$ for $0 < y < 1$. This is the pdf of the Uniform$(0\,,1)$ distribution.
(c) Part (b) says that $U \overset{\mathrm{d}}{=} F_X(X)$, where the notation $V \overset{\mathrm{d}}{=} W$ means that V and W have the same distribution. If $V \overset{\mathrm{d}}{=} W$, then $g(V) \overset{\mathrm{d}}{=} g(W)$ for any function g. Thus $F_X^{-1}(U) \overset{\mathrm{d}}{=} F_X^{-1}(F_X(X)) = X$. □

Part (c) can be used for random number generation. In order to generate a random observation from a continuous distribution with cdf $F(x)$, first generate a uniform random number U and then calculate $F^{-1}(U)$.

Theorem CB.2.1.2 can be extended to other functions $g(x)$.

Theorem CB.2.1.2 (continued). Suppose X has a continuous distribution with pdf $f_X(x)$ and support $\mathcal{X} = (a\,,b)$.

(b) Suppose $Y = g(X)$ where $g(x)$ is a strictly decreasing function on $(a\,,b)$ whose inverse function is continuously differentiable. Then Y has a continuous distribution with pdf

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \cdot \left|\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right| & \text{for } g(a) < y < g(b) \\ 0 & \text{otherwise} \end{cases}.$$

Theorem CB.2.1.3. Suppose X has a continuous distribution with pdf $f_X(x)$ and support $\mathcal{X} = (a\,,b)$. Consider $Y = g(X)$. Suppose the support can be split into subintervals, $(a\,,b) = (c_1\,,c_2) \cup (c_2\,,c_3) \;\cup \cdots \cup (c_{k-1}\,,c_k) \cup (c_k\,,c_{k+1})$, where $c_1 = a$ and $c_{k+1} = b$, such that on each subinterval $g(x)$ is either strictly increasing or strictly decreasing. (Note that for convenience we are ignoring the points $c_2\,, c_3\,, \ldots,$ c_k, which is permissible because a finite set of points has probability 0 for a continuous distribution.) Let $g_i^{-1}(y)$ denote the inverse function of $g(x)$ on the interval $(c_i\,,c_{i+1})$. Then Y has a continuous distribution with pdf

$$f_Y(y) = \textstyle\sum_i f_X(g_i^{-1}(y)) \cdot \left|\frac{\mathrm{d}}{\mathrm{d}y} g_i^{-1}(y)\right| ,$$

where the summation is over those indices i for which $g_i^{-1}(y)$ is well-defined, that is, for which there exists some $x \in (c_i\,,c_{i+1})$ with $g(x) = y$. (If there are no such indices i, then the sum is taken to be 0.)

Examples. (1) Consider $Y = |X|$. The function $g(x) = |x|$ is strictly decreasing for $x \in (-\infty, 0)$ and is strictly increasing for $x \in (0, \infty)$. On the interval $(-\infty, 0)$, we have $y = -x$, $x = -y$ and $\frac{\mathrm{d}x}{\mathrm{d}y} = -1$. On the interval $(0, \infty)$, we have $y = x$, $x = y$ and $\frac{\mathrm{d}x}{\mathrm{d}y} = 1$. Thus $f_Y(y) = f_X(-y)\cdot|-1| + f_X(y)\cdot|1| = f_X(-y) + f_X(y)$.

(2) Consider $Y = X^2$. The function $g(x) = x^2$ is strictly decreasing for $x \in (-\infty, 0)$ and is strictly increasing for $x \in (0, \infty)$. On the interval $(-\infty, 0)$, we have $y = x^2$, $x = -\sqrt{y}$ and $\frac{\mathrm{d}x}{\mathrm{d}y} = -\frac{1}{2\sqrt{y}}$. On the interval $(0, \infty)$, we have $y = x^2$, $x = \sqrt{y}$ and $\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{1}{2\sqrt{y}}$. Thus $f_Y(y) = f_X(-\sqrt{y})\cdot\left|-\frac{1}{2\sqrt{y}}\right| + f_X(\sqrt{y})\cdot\left|\frac{1}{2\sqrt{y}}\right| = \frac{1}{2\sqrt{y}}\left[f_X(-\sqrt{y}) + f_X(\sqrt{y})\right]$.

(3) Consider $Y = X^2$ when $X \sim \text{Normal}(0, 1)$. The pdf of X is $f_X(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}$. From (2) we know $f_Y(y) = \frac{1}{2\sqrt{y}}\left[f_X(-\sqrt{y}) + f_X(\sqrt{y})\right] = \frac{1}{2\sqrt{y}}\left[2\frac{1}{\sqrt{2\pi}}\, e^{-\frac{y}{2}}\right] = \frac{1}{\sqrt{2\pi y}}\, e^{-\frac{y}{2}}$ for $y > 0$ (and $= 0$ for $y \leq 0$). This is the pdf of a Chi-squared distribution with 1 degree of freedom (see p. 626). ||

CHAPTER 4 – Jointly distributed random variables

Bivariate random vectors

Suppose we observe an experiment that produces two random variables X and Y. Even though we may know everything about the distribution of X and about the distribution of Y, this does not tell us everything about the joint distribution of the two random variables. That is, we cannot figure out $\mathrm{P}(X \in A \text{ and } Y \in B)$ from the values of $\mathrm{P}(X \in A)$ and $\mathrm{P}(Y \in B)$ alone.

Example. Toss a fair coin twice. Let $V_1 =$ the number of heads on the first toss, $V_2 =$ the number of heads on the second toss, and $V_3 =$ the number of tails on the first toss. All three of these random variables have the Bernoulli($\frac{1}{2}$) distribution. Therefore, in all three of the following cases, $\mathrm{P}(X=1) = 0.5$ and $\mathrm{P}(Y=1) = 0.5$.
(i) If $X = V_1$ and $Y = V_1$, then $\mathrm{P}(X=1 \text{ and } Y=1) = 0.5$.
(ii) If $X = V_1$ and $Y = V_2$, then $\mathrm{P}(X=1 \text{ and } Y=1) = 0.25$.
(iii) If $X = V_1$ and $Y = V_3$, then $\mathrm{P}(X=1 \text{ and } Y=1) = 0$. ||

A *bivariate random vector* is, roughly speaking, a pair of numbers that has been randomly generated. Let (X,Y) be a bivariate random vector. The *joint distribution* of (X,Y) consists of the probabilities P(~~$X \in A$ and $Y \in B$~~ $(X,Y) \in D$) for all events ~~A and B.~~ D in the real plane. ←

Considered by itself, X is a random variable (and so is Y). The *marginal distribution* of X consists of the probabilities $\mathrm{P}(X \in A)$ for all events A in the real line. ←

The words "joint" and "marginal" are simply for emphasis and can be omitted; we can talk about the distribution of (X,Y) and the distribution of X without any ambiguity.

Discrete distributions

A bivariate random vector is *discrete* if there is a finite or countably infinite set C of pairs of real numbers such that $\mathrm{P}[(X,Y) \in C] = 1$. The distribution of a discrete bivariate random vector is often given by specifying its *joint probability mass function* (joint pmf),

$$f_{X,Y}(x,y) = \mathrm{P}[(X,Y) = (x,y)] = \mathrm{P}(X = x \text{ and } Y = y)$$

for all x and y. Of course for $(x,y) \notin C$, $f_{X,Y}(x,y) = 0$. Sometimes the joint pmf is denoted simply as $f(x,y)$. To be a valid joint pmf, the function must satisfy:

(a) $f(x,y) \geq 0$ for all x, y,

(b) $\sum\sum_{\text{all } x,y} f(x,y) = 1$.

For any event D in the real plane, we can calculate

$$\mathrm{P}[(X,Y)\in D] \;=\; \sum\sum_{(x,y)\,\in\, D} f_{X,Y}(x,y)\,.$$

Theorem CB.4.1.1. The marginal pmf of X can be expressed in terms of the joint pmf of (X,Y) as $f_X(x) = \sum_{\text{all } y} f_{X,Y}(x,y)$.

Justification: This theorem is simply saying that $\mathrm{P}(X=x) = \sum_{\text{all } y} \mathrm{P}(X=x \text{ and } Y=y)$. This follows from the Axiom of Additivity for probability. Note that the event $(X=x)$ is the union of the pairwise disjoint events $(X=x \text{ and } Y=y)$. □

A similar expression holds for the marginal pmf of Y.

Definition. The *expected value* of a function $g(X,Y)$ of a discrete bivariate random vector is $\mathrm{E}[g(X,Y)] = \sum\sum_{\text{all } x,\, y} g(x,y) f_{X,Y}(x,y)$.

The expected value of $Z = g(X,Y)$ can be computed in two ways:

(1) $\mathrm{E}(Z) = \sum_{\text{all } z} z f_Z(z)$ where $f_Z(z)$ is the pmf of the random variable Z,

(2) $\mathrm{E}[g(X,Y)] = \sum\sum_{\text{all } x,\, y} g(x,y) f_{X,Y}(x,y)$.

Consider a function that involves only X and not Y, say $h(X)$. Its expected value can be computed in three ways:

(1) $\mathrm{E}(Z) = \sum_{\text{all } z} z f_Z(z)$ where $Z = h(X)$,

(2) $\mathrm{E}[h(X)] = \sum_{\text{all } x} h(x) f_X(x)$,

(3) $\mathrm{E}[h(X)] = \sum\sum_{\text{all } x,\, y} h(x) f_{X,Y}(x,y)$.

In (3) we are using the definition above with $g(x,y) = h(x)$. To verify that (2) and (3) give the same result, note that $\sum\sum_{\text{all } x,\, y} h(x) f_{X,Y}(x,y) = \sum_{\text{all } x} \sum_{\text{all } y} h(x) f_{X,Y}(x,y) =$ $\sum_{\text{all } x} h(x) \sum_{\text{all } y} f_{X,Y}(x,y) =$ (by Theorem CB.4.1.1) $\sum_{\text{all } x} h(x) f_X(x)$.

Continuous distributions

We use the phrase "continuous" as short for "absolutely continuous". That is, (X,Y) is said to have a *continuous distribution* if it has a *joint probability density function* (joint pdf), which is a function such that

$$\mathrm{P}[(X,Y)\in D] \;=\; \iint_D f_{X,Y}(x,y)\,\mathrm{d}x\mathrm{d}y$$

for all events D in the real plane. Sometimes the joint pdf is denoted simply as $f(x,y)$.

To be a valid joint pdf, the function must satisfy:

(a) $f(x,y) \geq 0$ for all x,y,

(b) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = 1$.

Lemma (see CB(4.1.2)). The marginal pdf of X can be expressed in terms of the joint pdf of (X,Y) as $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}y$.

Justification: The marginal pdf of X is defined to be a function such that $\mathrm{P}(X \in A) = \int_A f_X(x)\mathrm{d}x$ for all events A in the real line. So we check that $\int_A \left\{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}y\right\}\mathrm{d}x = \iint_D f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y$ where $D = \{(x,y) : x \in A,\ y \in (-\infty,\infty)\}$. By the definition of the joint pdf, this last integral equals $\mathrm{P}(X \in A \text{ and } Y \in (-\infty,\infty)) = \mathrm{P}(X \in A)$. □

A similar expression holds for the marginal pdf of Y.

Definition. The *expected value* of a function $g(X,Y)$ of a continuous bivariate random vector is $\mathrm{E}[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y$.

The expected value of $Z = g(X,Y)$ can be computed in two ways:

(1) as $\mathrm{E}(Z)$ in terms of the distribution of Z,
which might be continuous or discrete or a combination
of a continuous component and a discrete component,

(2) as $\mathrm{E}[g(X,Y)]$ in the definition above.

Consider a function that involves only X and not Y, say $h(X)$. Its expected value can be computed in three ways:

(1) as $\mathrm{E}(Z)$ in terms of the distribution of Z,

(2) as $\mathrm{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x)\mathrm{d}x$, [illegible])

(3) as $\mathrm{E}[h(X)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x) f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y$.

In (3) we are using the definition above with $g(x,y) = h(x)$. To verify that (2) and (3) give the same result, note that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x) f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{\infty} h(x)\left\{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}y\right\}\mathrm{d}x =$ (by CB(4.1.2)) $\int_{-\infty}^{\infty} h(x) f_X(x)\mathrm{d}x$.

Example. Suppose a point is randomly picked within a disk of radius R. The disk can be expressed as $D = \{ (x, y) : \sqrt{x^2+y^2} < R \}$. What is meant by "randomly picked"? This can be mathematically formulated by giving the bivariate random vector (X, Y) a uniform distribution on the disk, that is,

$$f_{X,Y}(x, y) = \begin{cases} c & \text{if } \sqrt{x^2+y^2} < R \\ 0 & \text{otherwise} \end{cases}.$$

a. Find the correct value of c in order to have a valid pdf.
The pdf must satisfy $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y = 1$. Now $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}x\mathrm{d}y = \iint_D c\,\mathrm{d}x\mathrm{d}y = c\iint_D \mathrm{d}x\mathrm{d}y = c \times (\text{area of } D) = c\pi R^2$. So set $c = 1/\pi R^2$.

b. Find the marginal distribution of X.

$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}y = \int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \frac{1}{\pi R^2}\mathrm{d}y = \frac{1}{\pi R^2}\left[(\sqrt{R^2-x^2}) - (-\sqrt{R^2-x^2})\right] = \frac{2}{\pi R^2}\sqrt{R^2-x^2}$ for $-R < x < R$. The limits of integration come from the condition $\sqrt{x^2+y^2} < R$, which implies $-\sqrt{R^2-x^2} < y < \sqrt{R^2-x^2}$.

c. Find $\mathrm{E}(X)$.

$\mathrm{E}(X) = \int_{-\infty}^{\infty} x f_X(x)\mathrm{d}x = \int_{-R}^{R} x\frac{2}{\pi R^2}\sqrt{R^2-x^2}\,\mathrm{d}x = 0$ because of the following lemma and corollary.

Lemma N.4.1. If $g(x)$ is an odd function (that is, if $g(-x) = -g(x)$ for all x) and if $\int_{-\infty}^{\infty} |g(x)|\mathrm{d}x < \infty$, then $\int_{-\infty}^{\infty} g(x)\mathrm{d}x = 0$.

The finiteness of $\int_{-\infty}^{\infty} |g(x)|\mathrm{d}x$ is needed to ensure that $\int_{-\infty}^{\infty} g(x)\mathrm{d}x$ exists.

Corollary N.4.2. If a pdf $f_X(x)$ is an even function (that is, if $f_X(-x) = f_X(x)$ for all x) and if $\mathrm{E}(|X|) < \infty$, then $E(X) = 0$.

The corollary follows from the lemma because $x f_X(x)$ is an odd function if $f_X(x)$ is even. Recall that the Cauchy$(0, 1)$ distribution has a pdf that is an even function, but its mean is not 0 because the mean does not exist (because $\mathrm{E}(|X|) = \infty$).

In part c of our example, note that the pdf of X is an even function. Also note that $\mathrm{E}(|X|) = \int_{-R}^{R} |x|\frac{2}{\pi R^2}\sqrt{R^2-x^2}\,\mathrm{d}x \le \int_{-R}^{R} R\frac{2}{\pi R^2}\sqrt{R^2-0^2}\,\mathrm{d}x = \frac{4R}{\pi} < \infty$.

d. Find $\mathrm{Var}(X)$.

$$\mathrm{Var}(X) = \mathrm{E}(X^2) = \int_{-R}^{R} x^2 \frac{2}{\pi R^2}\sqrt{R^2-x^2}\,\mathrm{d}x = \frac{2}{\pi R^2}\int_{-R}^{R} x^2\sqrt{R^2-x^2}\,\mathrm{d}x\,.$$

Lemma N.4.3. If $g(x)$ is an even function (that is, if $g(-x) = g(x)$ for all x) and if $\int_{-\infty}^{\infty}|g(x)|\mathrm{d}x < \infty$, then $\int_{-\infty}^{\infty}g(x)\mathrm{d}x = 2\int_{0}^{\infty}g(x)\mathrm{d}x$.

Therefore, $\mathrm{Var}(X) = \frac{4}{\pi R^2}\int_0^R x^2\sqrt{R^2-x^2}\,\mathrm{d}x$. Change the variable of integration to $u = x/R$ and then to $t = u^2$ to obtain $\mathrm{Var}(X) = \frac{4R^2}{\pi}\int_0^1 u^2\sqrt{1-u^2}\,\mathrm{d}u =$ $\frac{2R^2}{\pi}\int_0^1 \sqrt{t}\sqrt{1-t}\,\mathrm{d}t$. We recognize $\sqrt{t}\sqrt{1-t} = t^{\frac{1}{2}}(1-t)^{\frac{1}{2}}$ as a kernel of a Beta$(\frac{3}{2}, \frac{3}{2})$ pdf. (A function $h(x)$ is said to be a *kernel* of a pdf $f_X(x)$ if $f_X(x) = ch(x)$ for some constant c not depending on x.) Since a pdf integrates to 1, the form of the Beta(α, β) pdf on p. CB.626 tells us that $\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}\mathrm{d}x = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$. Now $\mathrm{Var}(X) = \frac{2R^2}{\pi}\frac{\Gamma(\frac{3}{2})\Gamma(\frac{3}{2})}{\Gamma(3)} =$ $\frac{2R^2}{\pi}\frac{\frac{1}{2}\sqrt{\pi}\frac{1}{2}\sqrt{\pi}}{2} = \frac{R^2}{4}$. Thus $\mathrm{SD}(X) = \frac{R}{2}$.

e. An alternative way to calculate $\mathrm{Var}(X)$ is to use the joint pdf of (X, Y) rather than the marginal pdf of X that is used in part d.

$$\mathrm{Var}(X) = \mathrm{E}(X^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^2 f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = \iint_D x^2\frac{1}{\pi R^2}\mathrm{d}x\mathrm{d}y = \frac{1}{\pi R^2}\iint_D x^2\mathrm{d}x\mathrm{d}y\,.$$

If we do the double integral as an iterated integral in the order $\int\{\int x^2\mathrm{d}y\}\mathrm{d}x =$ $\int x^2\{\int \mathrm{d}y\}\mathrm{d}x$, we are doing the same procedure as in part d, since $\int\frac{1}{\pi R^2}\mathrm{d}y$, with the appropriate limits of integration, gives us $f_X(x)$. An alternative procedure is to compute an iterated integral in the order $\int\{\int x^2\mathrm{d}x\}\mathrm{d}y$. Now $\mathrm{Var}(X) = \frac{1}{\pi R^2}\int_{-R}^{R}\{\int_{-\sqrt{R^2-y^2}}^{\sqrt{R^2-y^2}} x^2\mathrm{d}x\}\mathrm{d}y =$ $\frac{1}{\pi R^2}\int_{-R}^{R}\frac{2}{3}(R^2-y^2)^{\frac{3}{2}}\mathrm{d}y$. By noting that the integrand is an even function and by changing the variable of integration from y to $t = y/R$ and then to $w = t^2$, one obtains $\mathrm{Var}(X) =$ $\frac{2R^2}{3\pi}\int_0^1 w^{-\frac{1}{2}}(1-w)^{\frac{3}{2}}\mathrm{d}y$. Here we recognize the kernel of a Beta$(\frac{1}{2}, \frac{5}{2})$ pdf, which leads us to

$$\mathrm{Var}(X) = \frac{2R^2}{3\pi}\frac{\Gamma(\frac{1}{2})\Gamma(\frac{5}{2})}{\Gamma(3)} = \frac{R^2}{4}\,.\ \|$$

The properties of expectation for functions of a bivariate random vector are essentially the same as for functions of a single random variable.

Lemma N.4.4 (see p. CB.131). Let (X, Y) be a bivariate random vector, c a constant, and $g(X, Y)$ and $h(X, Y)$ real-valued functions of (X, Y) whose expectations exist.

(a) $\mathrm{E}[g(X,Y)+c] = \mathrm{E}[g(X,Y)]+c$.

(b) $\mathrm{E}[cg(X,Y)] = c\mathrm{E}[g(X,Y)]$.

(c) $\mathrm{E}[g(X,Y)+h(X,Y)] = \mathrm{E}[g(X,Y)]+\mathrm{E}[h(X,Y)]$.

(d) If $g(X,Y) \leq h(X,Y)$, then $\mathrm{E}[g(X,Y)] \leq \mathrm{E}[h(X,Y)]$.

The distribution of (X, Y) is completely determined by its *joint cumulative distribution function* (cdf), $F_{X,Y}(x, y) = \mathrm{P}(X \leq x \text{ and } Y \leq y)$ for all $-\infty < x < \infty$ and $-\infty < y < \infty$.

If (X, Y) has a discrete joint distribution, then the joint pmf can be obtained from the joint cdf as $f_{X,Y}(x,y) = F_{X,Y}(x,y) - F_{X,Y}(x-\epsilon, y) - F_{X,Y}(x, y-\epsilon) + F_{X,Y}(x-\epsilon, y-\epsilon)$.

If (X, Y) has a continuous joint distribution, then the jcint pdf can be obtained from the joint cdf as $f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y)$.

Conditional distributions

Discrete distributions

Recall that if we know $\mathrm{P}(A)$ and $\mathrm{P}(B|A)$, then we can obtain $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B|A)$. Suppose (X, Y) has a discrete joint distribution. If we know $\mathrm{P}(X = x)$ and $\mathrm{P}(Y = y \,|\, X = x)$, then we can obtain $\mathrm{P}(X = x \text{ and } Y = y) = \mathrm{P}(X = x)\mathrm{P}(Y = y \,|\, X = x)$. That is, in pmf notation, $f_{X,Y}(x,y) = f_X(x)\, f_{Y|X}(y \,|\, x)$. Thus we define the *conditional pmf* of Y given X to be

$$f_{Y|X}(y \,|\, x) = \mathrm{P}(Y = y \,|\, X = x)$$

if the conditional probability is well-defined, i.e., if $f_X(x) = \mathrm{P}(X = x) > 0$.

For fixed x, the conditional pmf is a valid pmf for y, because (a) $f_{Y|X}(y \,|\, x) \geq 0$ for all y and (b) $\sum_{\text{all } y} f_{Y|X}(y \,|\, x) = \sum_{\text{all } y} \mathrm{P}(Y = y \,|\, X = x) = \sum_{\text{all } y} \mathrm{P}(X = x \text{ and } Y = y)/\mathrm{P}(X = x) = \mathrm{P}(X = x)/\mathrm{P}(X = x) = 1$.

The *conditional expectation* of $g(Y)$ given that $X = x$ is simply the expectation of Y under its conditional distribution given $X = x$:

$$\mathrm{E}[\,g(Y) \,|\, X = x] = \sum_{\text{all } y} g(y) f_{Y|X}(y \,|\, x).$$

Write $h(x) = \mathrm{E}[\,g(Y) \,|\, X = x]$. This is a function of x. Plugging in the random variable X, we obtain the random variable $h(X) = \mathrm{E}[\,g(Y) \,|\, X\,]$ (this notation is used instead of the funny-looking $\mathrm{E}[\,g(Y) \,|\, X = X\,]$). It can be shown that $\mathrm{E}[h(X)] = \mathrm{E}[\,g(Y)]$.

Theorem N.4.5 (see Theorem CB.4.4.1). $\mathrm{E}[g(Y)] = \mathrm{E}[\mathrm{E}[g(Y) \mid X]]$.

This provides a two-step procedure for calculating $\mathrm{E}[g(Y)]$ that is sometimes easier than calculating it directly from the marginal distribution of Y.

Example. A salesman makes an average of 1.6 sales per day. The average amount of a sale is \$150. On a randomly selected day, what is the expected value of the salesman's total amount of sales for that day? To formulate this problem in mathematical terms, let $T =$ the total amount of sales on that day. We want to calculate $\mathrm{E}(T)$. We see that the value of T depends on $N =$ the number of sales on that day. We know that $\mathrm{E}(N) = 1.6$ and $\mathrm{E}(T \mid N) = 150\,N$. Using the theorem above, $\mathrm{E}(T) = \mathrm{E}[\mathrm{E}(T \mid N)] = \mathrm{E}[150\,N] = 150\,\mathrm{E}(N) = 150(1.6) = 240$. ||

Justification of the theorem: $\mathrm{E}[\mathrm{E}[g(Y) \mid X]] = \sum_{\text{all } x} \left[\sum_{\text{all } y} g(y) f_{Y|X}(y \mid x) \right] f_X(x)$

$= \sum\sum_{\text{all } x,y} g(y) f_{X,Y}(x, y) = \mathrm{E}[g(Y)]$, because $f_{Y|X}(y \mid x) f_X(x) = f_{X,Y}(x, y)$. □

Continuous distributions

Suppose (X, Y) has a continuous joint distribution. The *conditional pdf* of Y given that $X = x$ is

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

provided that $f_X(x) > 0$. Sometimes $f_{Y|X}(y \mid x)$ and $f_X(x)$ are given first and then the joint pdf is obtained as $f_{X,Y}(x, y) = f_X(x) f_{Y|X}(y \mid x)$.

Example. In regression analysis it is often assumed that $Y \mid X = x \sim \mathrm{Normal}(\alpha + \beta x, \sigma^2)$. (In this case we are often not concerned about $f_X(x)$ and $f_{X,Y}(x, y)$.) ||

For fixed x, the conditional pdf is a valid pdf for y, because (a) $f_{Y|X}(y \mid x) \geq 0$ for all y and (b) $\int_{-\infty}^{\infty} f_{Y|X}(y \mid x)\mathrm{d}y = \int_{-\infty}^{\infty} \frac{f_{X,Y}(x, y)}{f_X(x)} \mathrm{d}y = \frac{1}{f_X(x)} \int_{-\infty}^{\infty} f_{X,Y}(x, y)\mathrm{d}y = \frac{1}{f_X(x)} f_X(x) = 1$.

The *conditional expectation* of $g(Y)$ given that $X = x$ is the expectation of Y under its conditional distribution given $X = x$:

$$\mathrm{E}[g(Y) \mid X = x] = \int_{-\infty}^{\infty} g(y) f_{Y|X}(y \mid x)\mathrm{d}y.$$

Theorem N.4.5 is true for all bivariate random vectors, continuous or discrete or mixed, for which the expectations exist.

Conditional mean and variance

The *conditional mean* of Y given that $X = x$ is $\mathrm{E}(Y \mid X = x)$, which is the mean of Y under its conditional distribution given $X = x$.
The *conditional variance* of Y given that $X = x$ is the variance of Y under its conditional distribution given $X = x$, that is,

$$\begin{aligned}\mathrm{Var}(Y \mid X = x) &= \mathrm{E}[(Y - \mu_{Y|x})^2 \mid X = x] \quad \text{where } \mu_{Y|x} = \mathrm{E}(Y \mid X = x) \\ &= \mathrm{E}(Y^2 \mid X = x) - \mu_{Y|x}^2 \\ &= \mathrm{E}(Y^2 \mid X = x) - [\mathrm{E}(Y \mid X = x)]^2 .\end{aligned}$$

According to Theorem N.4.5, the unconditional mean of Y can be obtained from the conditional mean as $\mathrm{E}(Y) = \mathrm{E}[\mathrm{E}(Y \mid X)]$. There is a similar formula for the variance:

Theorem CB.4.4.2. $\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}[\mathrm{E}(Y \mid X)]$.

$\mathrm{Var}(Y \mid X = x)$ measures the variation of Y around $\mathrm{E}(Y \mid X = x)$ for a fixed value of x. So the variation of Y can be decomposed into the component $\mathrm{E}[\mathrm{Var}(Y \mid X)]$, which measures the expected variation of Y around $\mathrm{E}(Y \mid X)$, and the component $\mathrm{Var}[\mathrm{E}(Y \mid X)]$, which measures the variation of $\mathrm{E}(Y \mid X)$.

Example (cont'd). Recall the example about the salesman in which $\mathrm{E}(N) = 1.6$ and $\mathrm{E}(T \mid N) = 150\,N$. We found that $\mathrm{E}(T) = 240$. Let us try to calculate $\mathrm{Var}(T)$. To do this we need more information. The standard deviation of the number of sales in a day is 1.1, that is $\mathrm{Var}(N) = (1.1)^2 = 1.21$. The standard deviation of the amount of a sale is \$75. Let A_1, ..., A_N be the amounts of the N sales during a day. Then $T = A_1 + \cdots + A_N$. Assume the amounts A_i are independent of one another. Then (see Lemma CB.5.2.1)
$\mathrm{Var}(T \mid N = n) = \mathrm{Var}(A_1 + \cdots + A_n) = n(75)^2 = 5625\,n$. So $\mathrm{Var}(T \mid N) = 5625\,N$.
Now the theorem above can be applied to yield $\mathrm{Var}(T) = \mathrm{E}[\mathrm{Var}(T \mid N)] + \mathrm{Var}[\mathrm{E}(T \mid N)] =$ $\mathrm{E}[5625\,N] + \mathrm{Var}[150\,N] = 5625\,\mathrm{E}(N) + (150)^2\mathrm{Var}(N) = 5625(1.6) + 22{,}500(1.21)$ $= 36{,}225$. So $\mathrm{SD}(T) = 190$. ||

Independence

Recall that A and B are independent events if $\mathrm{P}(A \mid B) = \mathrm{P}(A)$.
Let (X, Y) be a bivariate random vector.

Definition N.4.6. X and Y are *independent* if $\mathrm{P}(X \in A \mid Y \in B) = \mathrm{P}(X \in A)$ for all events A and B for which the conditional probability is well-defined (that is, $\mathrm{P}(Y \in B) > 0$).

Lemma N.1.13 implies the following lemma.

Lemma N.4.7. The following statements are equivalent:

(i) X and Y are independent.

(ii) $\mathrm{P}(X \in A \text{ and } Y \in B) = \mathrm{P}(X \in A)\mathrm{P}(Y \in B)$ for all events A and B.

(iii) $\mathrm{P}(Y \in B \,|\, X \in A) = \mathrm{P}(Y \in B)$ for all events A and B for which $\mathrm{P}(X \in A) > 0$.

Now suppose $(X\,, Y)$ have a joint pmf or pdf $f_{X,Y}(x\,,y)$.

Lemma N.4.8. The following statements are equivalent:

(i) X and Y are independent.

(ii) $f_{X|Y}(x\,|\,y) = f_X(x)$ for all x and y for which $f_Y(y) > 0$.

(iii) $f_{X,Y}(x\,,y) = f_X(x)f_Y(y)$ for all x and y.

(iv) $f_{Y|X}(y\,|\,x) = f_Y(y)$ for all x and y for which $f_X(x) > 0$.

Justification: Statement (ii) follows from statement (i), i.e., from Definition N.4.6, by letting $A = \{x\}$ and $B = \{y\}$ (in the discrete case). Recall that $f_{X,Y}(x\,,y) = f_X(x)f_{Y|X}(y\,|\,x) = f_Y(y)f_{X|Y}(x\,|\,y)$. This implies the equivalence of statements (ii), (iii) and (iv). To complete the justification, let us show that statement (iii) implies statement (ii) of Lemma N.4.7. In the continuous case, we have $\mathrm{P}(X \in A \text{ and } Y \in B) = \iint\limits_D f_{X,Y}(x\,,y)\mathrm{d}x\mathrm{d}y$ where $D = \{(x\,,y) : x \in A \text{ and } y \in B\} = A \times B$. Assuming (iii), this equals $\iint\limits_{A \times B} f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y = \int\limits_A f_X(x)\mathrm{d}x \int\limits_B f_Y(y)\mathrm{d}y = \mathrm{P}(X \in A)\mathrm{P}(Y \in B)$. □

(Strictly speaking, in the continuous case, the word "all" in Lemma N.4.8 should really be "almost all" — due to the technical fact that the pdf of a continuous random variable can be changed on a finite or countably infinite number of points without changing the distribution.)

Example. Let X and Y be two independent random observations from a population distributed as Normal$(\mu\,, \sigma^2)$. The joint pdf of $(X\,, Y)$ is $f_{X,Y}(x\,,y) = f_X(x)f_Y(y) =$ $\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \frac{1}{2\pi\sigma^2}\, e^{-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(y-\mu)^2}{2\sigma^2}} = \frac{1}{2\pi\sigma^2}\, e^{-\frac{(x-\mu)^2 + (y-\mu)^2}{2\sigma^2}}$. ||

Lemma CB.4.2.1. Let $(X\,, Y)$ be a bivariate random vector with joint pmf or pdf $f_{X,Y}(x\,,y)$. Then X and Y are independent if and only if there are functions $g(x)$ and $h(y)$ such that $f_{X,Y}(x\,,y) = g(x)h(y)$ for all x and y.

Justification: If X and Y are independent, then Lemma N.4.8 says that $f_{X,Y}(x\,,y) = g(x)h(y)$ with $g(x) = f_X(x)$ and $h(y) = f_Y(y)$. Conversely, suppose $f_{X,Y}(x\,,y) = g(x)h(y)$ where it is not necessarily true that $g(x) = f_X(x)$ and $h(y) = f_Y(y)$. Then $g(x)$ is a kernel of $f_X(x)$, because (in the continuous case) $f_X(x) =$

$\int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}y = \int_{-\infty}^{\infty} g(x)h(y)\mathrm{d}y = g(x)\int_{-\infty}^{\infty} h(y)\mathrm{d}y = c\,g(x)$ where $c = \int_{-\infty}^{\infty} h(y)\mathrm{d}y$. Similarly, $f_Y(y) = b\,h(y)$ for a constant b. Hence $f_{X,Y}(x,y) = g(x)h(y) = f_X(x)f_Y(y)/cb$. Since $f_{X,Y}(x,y)$ and $f_X(x)f_Y(y)$ both integrate to 1, we must have $cb = 1$. □

It is important to remember that the definition of $f_{X,Y}(x,y)$ includes the ranges of x and y. To keep track of this, it often helps to write the ranges in terms of indicator functions.

Example. Consider the joint pdf $f(x,y) = 8xy$ for $0 < x < y < 1$. We can write $8xy = g(x)h(y)$ where $g(x) = 8x$ and $h(y) = y$. But X and Y are not independent, because, for instance, $\mathrm{P}(X > 0.5 \,|\, Y < 0.5) = 0 \neq 0.5625 = \mathrm{P}(X > 0.5)$. The pdf is not simply $f(x,y) = 8xy$. This is the value of the pdf only for x and y satisfying $0 < x < y < 1$. It is safer to write $f(x,y) = 8xy\,I_{(0,y)}(x)I_{(x,1)}(y)$. (See p. CB.114 for the indicator function notation.) This makes it clearer that we cannot factor the joint pdf into the form $g(x)h(y)$. ||

More generally, X and Y cannot be independent if, in the support of (X,Y), the range of x depends on the value of y. That is, for X and Y to be independent, $\{x : f_{X,Y}(x,y) > 0\}$ must be the same for all y. Also, $\{y : f_{X,Y}(x,y) > 0\}$ must be the same for all x.

Theorem CB.4.2.1. If X and Y are independent, then $\mathrm{E}[g(X)h(Y)] = \mathrm{E}[g(X)]\mathrm{E}[h(Y)]$.

Justification: In the continuous case, $\mathrm{E}[g(X)h(Y)] =$
$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)\mathrm{d}x\mathrm{d}y =$
$\int_{-\infty}^{\infty} g(x)f_X(x)\mathrm{d}x \int_{-\infty}^{\infty} h(y)f_Y(y)\mathrm{d}y = \mathrm{E}[g(X)]\mathrm{E}[h(Y)]$. □

Corollary. Suppose X and Y are independent with means μ and ν and variances σ^2 and τ^2. Then $\mathrm{E}(XY) = \mu\nu$ and $\mathrm{Var}(XY) = \sigma^2\tau^2 + \sigma^2\nu^2 + \mu^2\tau^2$.

Justification: $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y) = \mu\nu$ and $\mathrm{Var}(XY) = \mathrm{E}(X^2Y^2) - [\mathrm{E}(XY)]^2$
$= \mathrm{E}(X^2)\mathrm{E}(Y^2) - \mu^2\nu^2 = (\sigma^2 + \mu^2)(\tau^2 + \nu^2) - \mu^2\nu^2 = \sigma^2\tau^2 + \sigma^2\nu^2 + \mu^2\tau^2$. □

Theorem CB.4.2.~~1~~2. If X and Y are independent, then $M_{X+Y}(t) = M_X(t)M_Y(t)$.

Justification: $M_{X+Y}(t) = \mathrm{E}[e^{t(X+Y)}] = \mathrm{E}[e^{tX+tY}] = \mathrm{E}[e^{tX}e^{tY}]$. Now apply the preceding theorem to get $M_{X+Y}(t) = \mathrm{E}[e^{tX}]\mathrm{E}[e^{tY}] = M_X(t)M_Y(t)$. □

Example (Theorem CB.4.3.1). Suppose $X \sim \mathrm{Poisson}(\lambda)$ and $Y \sim \mathrm{Poisson}(\mu)$ are independent. On p. 625 we see that the mgf of X is $M_X(t) = e^{\lambda(e^t-1)}$. By the theorem, $M_{X+Y}(t) = M_X(t)M_Y(t) = e^{\lambda(e^t-1)}e^{\mu(e^t-1)} = e^{\lambda(e^t-1)+\mu(e^t-1)} = e^{(\lambda+\mu)(e^t-1)}$, which is the mgf of a $\mathrm{Poisson}(\lambda+\mu)$ distribution. By the uniqueness of mgf's (Theorem CB.2.3.3), $X + Y \sim \mathrm{Poisson}(\lambda+\mu)$. ||

Example (Theorem CB.4.2.3). Suppose $X \sim \text{Normal}(\mu, \sigma^2)$ and $Y \sim \text{Normal}(\nu, \tau^2)$ are independent. On p. 628 we see that the mgf of X is $M_X(t) = e^{\mu t + \sigma^2 t^2/2}$. By the theorem, $M_{X+Y}(t) = M_X(t) M_Y(t) = e^{\mu t + \sigma^2 t^2/2} e^{\nu t + \tau^2 t^2/2} = e^{\mu t + \sigma^2 t^2/2 + \nu t + \tau^2 t^2/2}$
$= e^{(\mu+\nu)t + (\sigma^2+\tau^2)t^2/2}$, which is the mgf of a Normal$(\mu+\nu, \sigma^2+\tau^2)$ distribution. By the uniqueness of mgf's, $X + Y \sim \text{Normal}(\mu+\nu, \sigma^2+\tau^2)$. ||

Theorem CB.4.3.2. If X and Y are independent, then $U = g(X)$ and $V = h(Y)$ are independent.

Justification: Use Lemma N.4.7. Consider $\text{P}(U \in A \text{ and } V \in B) =$ $\text{P}(g(X) \in A \text{ and } h(Y) \in B) = \text{P}(X \in C \text{ and } Y \in D)$ where $C = \{x : g(x) \in A\}$ and $D = \{y : h(y) \in B\}$. Now appeal to the independence of X and Y to conclude that $\text{P}(U \in A \text{ and } V \in B) = \text{P}(X \in C)\text{P}(Y \in D) = \text{P}(U \in A)\text{P}(V \in B)$. □

The distribution of a function of a bivariate random vector

If (X, Y) is a bivariate random vector and $g(x, y)$ is a real-valued function, then $U = g(X, Y)$ is a random variable. Given the distribution of (X, Y) we would like to find the distribution of U.

Discrete distributions

Suppose (X, Y) has joint pmf $f_{X,Y}(x, y)$. The pmf of $U = g(X, Y)$ is

$$f_U(u) = \text{P}(U = u) = \text{P}[g(X, Y) = u].$$

Rather than deal with X and Y simultaneously, it is often easier to deal with them one at a time, by using the Theorem of Total Probability.

$$\begin{aligned} \text{P}[g(X, Y) = u] &= \sum_{\text{all } x} \text{P}(X = x)\text{P}[g(X, Y) = u \mid X = x] \\ &= \sum_{\text{all } x} \text{P}(X = x)\text{P}[g(x, Y) = u \mid X = x]. \end{aligned}$$

Lemma N.4.9. If X and Y are independent discrete random variables and $U = g(X, Y)$, then the pmf of U is $f_U(u) = \sum_{\text{all } x} \text{P}(X = x)\text{P}[g(x, Y) = u]$.

Example. Suppose $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Binomial}(m, p)$ are independent. Let us find the pmf of $U = X + Y$. Using the lemma above, we obtain

$$\begin{aligned} f_U(u) &= \sum_{\text{all } x} \text{P}(X = x)\text{P}(x + Y = u) = \sum_{\text{all } x} \text{P}(X = x)\text{P}(Y = u - x) \\ &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \binom{m}{u-x} p^{u-x} (1-p)^{m-(u-x)} \end{aligned}$$

$$= \sum_{x=0}^{n} \binom{n}{x} \binom{m}{u-x} p^u (1-p)^{n+m-u} = \binom{n+m}{u} p^u (1-p)^{n+m-u},$$

because $\sum_{x=0}^{n} \binom{n}{x} \binom{m}{u-x} = \binom{n+m}{u}$, because to choose u things from among a set of $n+m$ things, one can divide the set into two subsets of sizes n and m and then choose x things ($x \leq n$) from the subset of size n and choose the remaining $u-x$ things from the subset of size m. We recognize the pmf of U as that of the Binomial$(n+m, p)$ distribution. ||

Continuous distributions

Suppose (X, Y) has joint pdf $f_{X,Y}(x, y)$. The pdf of $U = g(X, Y)$ can be found through the three steps below, provided that the function $g(x, y)$ is "well-behaved" enough so that steps 1 and 2 are possible.

Step 1. Write $U = g_1(X, Y)$. Find (if possible) $V = g_2(X, Y)$ such that X and Y can be solved for in terms of U and V:

$$U = g_1(X, Y) \qquad X = h_1(U, V)$$
$$V = g_2(X, Y) \qquad Y = h_2(U, V).$$

Step 2. If the functions $h_1(u, v)$ and $h_2(u, v)$ are differentiable, then the joint pdf of (U, V) can be obtained as

(CB.4.3.2) $\quad f_{U,V}(u, v) = f_{X,Y}(x, y)\,|J|$

where J is the Jacobian matrix $J = \frac{\partial(x,y)}{\partial(u,v)} = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}$ and $|J|$ is the absolute value of the determinant of the matrix J. Thus $|J| = |\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial y}{\partial u}\frac{\partial x}{\partial v}|$. (Note that this is the bivariate analog of the univariate Theorem CB.2.1.2, which states that if $U = g(X)$, if X can be solved for in terms of U as $X = h(U)$, and if $h(u)$ is differentiable, then $f_U(u) = f_X(x)\,|\frac{dx}{du}|$.)

Step 3. The marginal pdf of U is obtained as $f_U(u) = \int_{-\infty}^{\infty} f_{U,V}(u, v) dv$.

Example. Suppose X and Y are two independent Normal$(0, 1)$ random variables. Let us find the pdf of $U = X/Y$.

Step 1. Let $V = Y$. (Many other choices for V are possible, such as $V = X$ or $1/X$ or $1/Y$ or $X + Y$. One can try to choose V to make the calculation of the Jacobian easy or one can choose V to be another random variable besides U that is of interest.) Then

$$U = X/Y \qquad X = UV$$
$$V = Y \qquad Y = V.$$

Step 2. $\dfrac{\partial x}{\partial u} = v \qquad \dfrac{\partial x}{\partial v} = u$

$$\frac{\partial y}{\partial u} = 0 \qquad \frac{\partial y}{\partial v} = 1 \quad \text{so} \quad J = \begin{bmatrix} v & u \\ 0 & 1 \end{bmatrix} \quad \text{so} \quad |J| = |v|\,.$$

$$f_{U,V}(u\,,v) = f_{X,Y}(x\,,y)\,\big|J\big| = f_{X,Y}(uv\,,v)\,|v| = f_X(uv)f_Y(v)\,|v|$$

$$= \frac{1}{\sqrt{2\pi}}\,e^{-\frac{(uv)^2}{2}} \cdot \frac{1}{\sqrt{2\pi}}\,e^{-\frac{v^2}{2}}\,|v| = \frac{|v|}{2\pi}\,e^{-\frac{1}{2}(u^2v^2+v^2)} = \frac{|v|}{2\pi}\,e^{-\frac{1}{2}(u^2+1)v^2}$$

We should always be careful about the domain of any pdf formula. The formula above holds for all x and y, and so it holds for all u and v.

Step 3. $f_U(u) = \int_{-\infty}^{\infty} f_{U,V}(u\,,v)\mathrm{d}v = \int_{-\infty}^{\infty} \frac{|v|}{2\pi}\,e^{-\frac{1}{2}(u^2+1)v^2}\mathrm{d}v =$ (because the integrand is an even function) $2\int_0^{\infty} \frac{|v|}{2\pi}\,e^{-\frac{1}{2}(u^2+1)v^2}\mathrm{d}v = \frac{1}{\pi}\int_0^{\infty} v\,e^{-\frac{1}{2}(u^2+1)v^2}\mathrm{d}v$. The integrand can be recognized as the kernel of a Weibull pdf with $\gamma = 2$ and $\beta = 2/(u^2+1)$. So the integral is the reciprocal of the constant γ/β in the Weibull pdf. Thus $\int_0^{\infty} v\,e^{-\frac{1}{2}(u^2+1)v^2}\mathrm{d}v = \beta/\gamma = 1/(u^2+1)$, and so $f_U(u) = \frac{1}{\pi(1+u^2)}$, which is seen to be the pdf of the Cauchy$(0\,,1)$ distribution. Therefore, the ratio of two independent Normal$(0\,,1)$ random variables is a Cauchy$(0\,,1)$ random variable. ||

Example (continued). Continue to let X and Y be two independent Normal$(0\,,1)$ random variables. Let us find $\mathrm{P}(X/Y < 2)$. There are two ways to approach this problem.

(A) One approach is to first find the distribution of $U = X/Y$. From above we know that $U \sim$ Cauchy$(0\,,1)$. Now

$$\mathrm{P}(U < 2) = \int_{-\infty}^{2} f_U(u)\mathrm{d}u = \int_{-\infty}^{2} \frac{1}{\pi(1+u^2)}\mathrm{d}u$$

$$= \frac{1}{\pi}\arctan(u)\Bigg|_{u=-\infty}^{u=2} = \frac{1}{\pi}\big[\arctan(2) - \arctan(-\infty)\big]$$

$$= \frac{1}{\pi}\big[1.107 - (-\tfrac{\pi}{2})\big] = 0.8524\,.$$

(B) Another approach is to calculate the probability directly from the joint distribution of $(X\,,Y)$. $\mathrm{P}(X/Y < 2) = \mathrm{P}[(X < 2Y \text{ and } Y > 0) \text{ or } (X > 2Y \text{ and } Y < 0)] = \mathrm{P}(X < 2Y \text{ and } Y > 0) + \mathrm{P}(X > 2Y \text{ and } Y < 0)$. Let $\phi(x)$ denote the pdf of the Normal$(0\,,1)$ distribution and let $\Phi(x)$ denote its cdf. Now

$$\mathrm{P}(X < 2Y \text{ and } Y > 0) = \iint_D f_{X,Y}(x\,,y)\mathrm{d}x\mathrm{d}y$$

where $D = \{(x\,,y) : x < 2y \text{ and } y > 0\}$. Thus

$$\mathrm{P}(X < 2Y \text{ and } Y > 0) = \iint_D \phi(x)\phi(y)\mathrm{d}x\mathrm{d}y = \int_0^{\infty}\Big[\int_{-\infty}^{2y}\phi(x)\mathrm{d}x\Big]\phi(y)\mathrm{d}y$$

$$= \int_0^{\infty}\Phi(2y)\phi(y)\mathrm{d}y\,.$$

It can be shown that $\mathrm{P}(X > 2Y \text{ and } Y < 0)$ has the same value, so $\mathrm{P}(X/Y < 2) = 2\int_0^\infty \Phi(2y)\phi(y)dy$. One can integrate this numerically. Using MATLAB, this integral can be calculated by forming the function

```
        function g = g(y)
        g = normcdf(2*y).*normpdf(y);
```

and then calculate the numerical integral

```
        2*quad('g',0,b)
```

for $\mathsf{b} = 1,2,3,4,5$, noting convergence to 0.8524 for $\mathsf{b} = 4,5$. ||

Example. Suppose X and Y are two independent random variables with distributions Gamma$(\alpha, 1)$ and Gamma$(\rho, 1)$ respectively. Let us find the pdf of $U = X/(X+Y)$.

Step 1. Let $V = X+Y$. Then we have

$$\begin{array}{ll} U = X/(X+Y) & X = UV \\ V = X+Y & Y = (1-U)V. \end{array}$$

Step 2. $\frac{\partial x}{\partial u} = v \qquad \frac{\partial x}{\partial v} = u$

$$\frac{\partial y}{\partial u} = -v \qquad \frac{\partial y}{\partial v} = 1-u, \text{ so } J = \begin{bmatrix} v & u \\ -v & 1-u \end{bmatrix}, \text{ so } |J| = |v|.$$

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(x,y)\,|J| = f_{X,Y}(uv,(1-u)v)\,|v| \\ &= f_X(uv)f_Y((1-u)v)\,|v| = \frac{1}{\Gamma(\alpha)}(uv)^{\alpha-1}e^{-uv}\cdot\frac{1}{\Gamma(\rho)}((1-u)v)^{\rho-1}e^{-(1-u)v}\,v \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\rho)}u^{\alpha-1}(1-u)^{\rho-1}v^{\alpha-1+\rho-1+1}e^{-uv-(1-u)v} \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\rho)}u^{\alpha-1}(1-u)^{\rho-1}v^{\alpha+\rho-1}e^{-v}. \end{aligned}$$

The domain of this pdf formula is $0 < x < \infty$ and $0 < y < \infty$, hence $0 < uv < \infty$ and $0 < (1-u)v < \infty$. Adding these two inequalities we obtain $0 < v < \infty$. Dividing the two preceding inequalities by v, we obtain $0 < u < \infty$ and $0 < 1-u < \infty$, which is equivalent to $0 < u < 1$.

Alternative step 3. Note that

$$f_{U,V}(u,v) = Cg(u)h(v),\ 0 < u < 1,\ 0 < v < \infty,$$

where C is a constant, $g(u)$ is a kernel of the Beta(α,ρ) pdf and $h(v)$ is a kernel of the Gamma$(\alpha+\rho,1)$ pdf. Without doing any integration, we can conclude that $X/(X+Y) \sim$ Beta(α,ρ), $X+Y \sim$ Gamma$(\alpha+\rho,1)$, and that the two random variables are independent. ||

Hierarchical models

Lemma N.4.10. If $X \,|\, Y = y$ has pmf $f_{X|Y}(x \,|\, y)$ and Y has pmf $f_Y(y)$, then X has pmf

$$f_X(x) = \sum_{\text{all } y} f_{X|Y}(x \,|\, y) f_Y(y) .$$

Justification: This follows from Theorem CB.4.1.1 and the fact that $f_{X,Y}(x\,,y) = f_{X|Y}(x \,|\, y) f_Y(y)$. □

Example. Suppose $X \,|\, Y = y \sim \text{Binomial}(y\,,p)$ and $Y \sim \text{Poisson}(\lambda)$. This can be used as a simple model for insect reproduction, letting Y be the number of eggs laid by an insect and letting $X \,|\, Y = y$ be the number of eggs that survive given that y eggs are laid. The pmf of the marginal distribution of X is

$$\begin{aligned} f_X(x) &= \sum_{\text{all } y} f_{X|Y}(x \,|\, y) f_Y(y) = \sum_{y=x}^{\infty} \binom{y}{x} p^x (1-p)^{y-x} \frac{e^{-\lambda}\lambda^y}{y!} \\ &= \sum_{y=x}^{\infty} \frac{y!}{x!(y-x)!} p^x (1-p)^{y-x} \frac{e^{-\lambda}\lambda^y}{y!} = \frac{p^x e^{-\lambda}}{x!} \sum_{y=x}^{\infty} \frac{1}{(y-x)!} (1-p)^{y-x} \lambda^y \\ &= \frac{p^x e^{-\lambda}}{x!} \sum_{t=0}^{\infty} \frac{1}{t!} (1-p)^t \lambda^{t+x} \quad (\text{letting } t = y - x) \\ &= \frac{p^x e^{-\lambda} \lambda^x}{x!} \sum_{t=0}^{\infty} \frac{[(1-p)\lambda]^t}{t!} = \frac{p^x e^{-\lambda}\lambda^x}{x!} e^{(1-p)\lambda} = \frac{e^{-p\lambda}(p\lambda)^x}{x!} . \end{aligned}$$

Here we have used the summation fact $\sum_{t=0}^{\infty} \frac{\lambda^t}{t!} = e^{\lambda}$, with λ replaced by $(1-p)\lambda$. This fact is a consequence of the fact that the Poisson(λ) pmf sums to 1. We recognize this pmf as a Poisson($p\lambda$) pmf. Therefore $X \sim \text{Poisson}(p\lambda)$. ||

Lemma N.4.11. If $X \,|\, Y = y$ has pdf $f_{X|Y}(x \,|\, y)$ and Y has pdf $f_Y(y)$, then X has pdf

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

Justification: This follows from CB(4.1.2) and the fact that $f_{X,Y}(x\,,y) = f_{X|Y}(x \,|\, y) f_Y(y)$. □

Example. A machine produces bolts of length X . A setting Y on the machine regulates the average length of the bolts. The setting is subject to error. Assuming $X \,|\, Y = y \sim \text{Normal}(y\,, b^2)$ and $Y \sim \text{Normal}(a\,, c^2)$, let us find the marginal distribution of X .
Its pdf is

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X|Y}(x \,|\, y) f_Y(y) dy \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi b^2}}\, e^{-\frac{(x-y)^2}{2b^2}} \cdot \frac{1}{\sqrt{2\pi c^2}}\, e^{-\frac{(y-a)^2}{2c^2}} dy \\ &= \frac{1}{2\pi bc} \int_{-\infty}^{\infty} e^{-\frac{1}{2b^2}(x-y)^2 - \frac{1}{2c^2}(y-a)^2} dy \end{aligned}$$

Through messy algebraic manipulation, the exponent can be rewritten as

$$-\frac{1}{2b^2}(x-y)^2-\frac{1}{2c^2}(y-a)^2 \;=\; -\frac{1}{2\sigma^{*2}}(y-\mu^*)^2-\frac{1}{2(b^2+c^2)}(x-a)^2$$

where μ^* is some quantity not involving y (other than this, it doesn't matter exactly what μ^* is) and $\sigma^{*2}=b^2c^2/(b^2+c^2)$. Now

$$f_X(x) \;=\; \frac{1}{2\pi bc}\,e^{-\frac{(x-a)^2}{2(b^2+c^2)}}\int_{-\infty}^{\infty} e^{-\frac{(y-\mu^*)^2}{2\sigma^{*2}}}\,dy$$

$$=\; \frac{1}{2\pi bc}\,e^{-\frac{(x-a)^2}{2(b^2+c^2)}}\sqrt{2\pi\sigma^{*2}}$$

$$=\; \frac{1}{\sqrt{2\pi(b^2+c^2)}}\,e^{-\frac{(x-a)^2}{2(b^2+c^2)}}\,,$$

which is the pdf of the Normal$(a\,,b^2+c^2)$ distribution. ||

Lemma N.4.12. If $X\,|\,Y=y$ is discrete with pmf $f_{X|Y}(x\,|\,y)$ and Y is continuous with pdf $f_Y(y)$, then X has pmf $f_X(x)=\int_{-\infty}^{\infty} f_{X|Y}(x\,|\,y)f_Y(y)dy$.

To formally prove this lemma, one should first define what is meant by $f_{X|Y}(x\,|\,y)$ when X is discrete and Y is continuous. But skipping the formal definitions, we can say that by analogy with the preceding two lemmas, this result seems correct.

Example. Suppose $X\,|\,Y=y\sim$ Poisson(y) and $Y\sim$ Exponential(β). The marginal pmf of X is $f_X(x)=\int_{-\infty}^{\infty} f_{X|Y}(x\,|\,y)f_Y(y)dy \;=\; \int_0^{\infty}\frac{e^{-y}y^x}{x!}\cdot\frac{1}{\beta}e^{-y/\beta}\,dy \;=$ $\frac{1}{\beta(x!)}\int_0^{\infty} y^x e^{-y(1+\frac{1}{\beta})}dy$. The last integrand is a kernel of the Gamma$(x+1\,,\frac{\beta}{1+\beta})$ pdf, so we obtain $f_X(x)\;=\;\frac{1}{\beta(x!)}\Gamma(x+1)\left(\frac{\beta}{1+\beta}\right)^{x+1} \;=\; \frac{\beta^x}{(1+\beta)^{x+1}} \;=\; \frac{1}{1+\beta}\left(\frac{\beta}{1+\beta}\right)^x$, which is the pmf of the Negative binomial$(p=\frac{1}{1+\beta}\,,r=1)$ distribution. This is almost the same as the Geometric distribution: if $X\sim$ Negative binomial$(p\,,1)$, then $X+1\sim$ Geometric(p). ||

Covariance and correlation

Suppose X and Y are jointly distributed random variables with means μ_X and μ_Y. Their covariance is defined to be $\text{Cov}(X\,,Y)=\text{E}[(X-\mu_X)(Y-\mu_Y)]$.

Theorem CB.4.5.1. $\text{Cov}(X\,,Y)=\text{E}(XY)-\mu_X\mu_Y$.

This formula can be justified through algebraic manipulation of $\text{E}[(X-\mu_X)(Y-\mu_Y)]$ (see p. CB.161). When $X=Y$, note that $\text{Cov}(X\,,X)=\text{E}[(X-\mu_X)^2]=\text{Var}(X)$. In this special case the formula in the preceding theorem is the same as formula CB(2.3.1): $\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu_X^2$.

Some properties of covariance are listed in the following lemma.

Lemma N.4.13 (see Theorem CB.4.5.3). Let X, Y and Z be jointly distributed random variables and let c be a constant.

(a) $\mathrm{Cov}(c, Y) = 0$.

(b) $\mathrm{Cov}(X + Z, Y) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(Z, Y)$.

(c) $\mathrm{Cov}(cX, Y) = c\,\mathrm{Cov}(X, Y)$.

(d) $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$.

(e) $\mathrm{Cov}(Y, X) = \mathrm{Cov}(X, Y)$.

`Partial justification`: These properties all follow from the definitions of covariance and variance and from the properties of expectation.

For (a), recall that $\mathrm{E}(c) = c$, so that $\mathrm{Cov}(c, Y) = \mathrm{E}[(c - c)(Y - \mu_Y)] = \mathrm{E}(0) = 0$.

For (c), recall that $\mathrm{E}(cX) = c\,\mathrm{E}(X)$, so that $\mathrm{Cov}(cX, Y) = \mathrm{E}[(cX - c\mu_X)(Y - \mu_Y)] = \mathrm{E}[c(X - \mu_X)(Y - \mu_Y)] = c\,\mathrm{E}[(X - \mu_X)(Y - \mu_Y)] = c\,\mathrm{Cov}(X, Y)$.

For (d), $\mathrm{Var}(X + Y) = \mathrm{E}[(X + Y - \mu_X - \mu_Y)^2] = \mathrm{E}[((X - \mu_X) + (Y - \mu_Y))^2] = \mathrm{E}[(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)] = \mathrm{E}[(X - \mu_X)^2] + \mathrm{E}[(Y - \mu_Y)^2] + 2\mathrm{E}[(X - \mu_X)(Y - \mu_Y)] = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y)$. □

Theorem CB.4.5.2. If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$.

`Justification`: $\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] =$ (by independence) $\mathrm{E}(X - \mu_X)\mathrm{E}(Y - \mu_Y) = 0$, because $\mathrm{E}(X - \mu_X) = \mathrm{E}(X) - \mu_X = \mu_X - \mu_X = 0$. □

The *correlation* (or *correlation coefficient*) of X and Y is defined to be $\rho_{XY} = \mathrm{Cov}(X, Y)/\sigma_X \sigma_Y$.

Suppose that Y tends to be above average when X is above average and that Y tends to be below average when X is below average. We can express this by saying that $X - \mu_X$ and $Y - \mu_Y$ tend to have the same sign, or by saying that $(X - \mu_X)(Y - \mu_Y)$ tends to be positive. Then $\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)] > 0$ and $\rho_{XY} > 0$, and we say that X and Y are positively correlated.

Theorem CB.4.5.4. (a) $-1 \le \rho_{XY} \le 1$.

(b) If $\rho_{XY} = 1$, then $Y = aX + b$ (w.p. 1) where a and b are constants with $a > 0$.

(c) If $\rho_{XY} = -1$, then $Y = aX + b$ (w.p. 1) where a and b are constants with $a < 0$.

(The abbreviation “w.p. 1” means “with probability 1”.)

Justification: A variance is always nonnegative, and it is zero if and only if the random variable is constant. In particular, $\operatorname{Var}\left(\frac{Y}{\sigma_Y} - \rho\frac{X}{\sigma_X}\right) \geq 0$. Manipulation of this inequality using Lemma N.4.13(d) yields the theorem. □

Note that correlation really means "linear correlation". Perfect correlation of X and Y means that one is a linear function of the other.

Example. Let $X \sim \text{Normal}(0,1)$ and let $Y = X^2$. Then X and Y have perfect "quadratic correlation" but their linear correlation is 0, because $\operatorname{Cov}(X,Y) = 0$. To see this, calculate $\operatorname{Cov}(X,Y) = \mathrm{E}(XY) - \mu_X\mu_Y = \mathrm{E}(X^3) - 0 \cdot 1 = \mathrm{E}(X^3) = 0$, because $\int\limits_{-\infty}^{\infty} x^3 \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, dx$ $= 0$ by Lemma N.4.1, which is applicable because the integrand is an odd function. The integral of the absolute value is finite, that is, $\mathrm{E}(|X|^3) < \infty$, because the mgf of X exists. If the mgf exists then all moments exist.

Incidentally, this example also shows that the converse of Theorem CB.4.5.2 is false. The covariance of X and Y is 0 but they are not independent, because $\mathrm{P}(|X| < 1$ and $Y > 1) = \mathrm{P}(X^2 < 1$ and $X^2 > 1) = 0 \neq \mathrm{P}(|X| < 1) \cdot \mathrm{P}(Y > 1)$. ||

Example. Let X and Y be i.i.d. Exponential(1). The pdf is $f(x) = e^{-x}$ for $x > 0$. Let's find the correlation of $X/(X+Y)$ and X. Let $U = X/(X+Y)$ and $V = X$. We want to calculate $\operatorname{Corr}(U,V) = \operatorname{Cov}(U,V)/\sigma_U\sigma_V = [\mathrm{E}(UV) - \mu_U\mu_V]/\sigma_U\sigma_V$. Since $V = X \sim$ Exponential(1), we know $\mu_V = 1$ and $\sigma_V^2 = 1$. Note that Exponential(1) = Gamma(1, 1). We have seen in the example on p. N.34 that $U = X/(X+Y) \sim \text{Beta}(1,1)$. Note that $\text{Beta}(1,1) = \text{Uniform}(0,1)$, so $\mu_U = \frac{1}{2}$ and $\sigma_U^2 = \frac{1}{12}$. It remains to calculate $\mathrm{E}(UV)$. For this we need the joint distribution of U and V. We know how to get their joint pdf. Note that $X = V$ and $Y = (\frac{1}{U} - 1)V$. Calculate $|J| = \ldots = \frac{v}{u^2}$ and $f_{U,V}(u,v) =$ $f(v)f((\frac{1}{u} - 1)v)\frac{v}{u^2} = \ldots = \frac{v}{u^2}e^{-\frac{v}{u}}$ for $0 < u < 1$ and $0 < v < \infty$. For the ranges of u and v, note that for any fixed value of $v = x > 0$, the value of $u = x/(x+y)$ varies from 0 to 1 as y varies from ∞ to 0. So the ranges of u and v do not depend on one another.

Now $\mathrm{E}(UV) = \int\int uv\, f_{U,V}(u,v)\mathrm{d}u\mathrm{d}v = \int_0^1 \left[\int_0^\infty uv\frac{v}{u^2}e^{-\frac{v}{u}}\mathrm{d}v\right]\mathrm{d}u =$ $\int_0^1 \left[\int_0^\infty v^2 e^{-\frac{v}{u}}\mathrm{d}v\right]u^{-1}\mathrm{d}u$. The inside integrand is a kernel of the Gamma$(3,u)$ distribution. Thus $\mathrm{E}(UV) = \int_0^1 \Gamma(3)u^3u^{-1}\mathrm{d}u = 2\int_0^1 u^2\mathrm{d}u = \frac{2}{3}$. So $\operatorname{Cov}(U,V) = \frac{2}{3} - \frac{1}{2}\cdot 1 = \frac{1}{6}$ and $\operatorname{Corr}(U,V) = \frac{1}{6}/\left(\sqrt{\frac{1}{12}}\cdot\sqrt{1}\right) = 1/\sqrt{3} \approx 0.5774$. ||

Bivariate normal distributions

Suppose that $X \sim \text{Normal}(\mu_X\,,\sigma_X^2)$ and $Y \sim \text{Normal}(\mu_Y\,,\sigma_Y^2)$ and that they are independent. Then their joint pdf is

$$f_{X,Y}(x\,,y) \;=\; \frac{1}{\sqrt{2\pi\sigma_X^2}}\,e^{-\frac{(x-\mu_X)^2}{2\sigma_X^2}} \cdot \frac{1}{\sqrt{2\pi\sigma_Y^2}}\,e^{-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}}$$

$$= \frac{1}{2\pi\sigma_X\sigma_Y}\,\exp\Big\{-\tfrac{1}{2}\Big[\Big(\tfrac{x-\mu_X}{\sigma_X}\Big)^2 + \Big(\tfrac{y-\mu_Y}{\sigma_Y}\Big)^2\Big]\Big\}\,.$$

More generally, a *bivariate normal distribution* has the joint pdf $f_{X,Y}(x\,,y) \;=$

$$\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\,\exp\Big\{-\tfrac{1}{2(1-\rho^2)}\Big[\Big(\tfrac{x-\mu_X}{\sigma_X}\Big)^2 + \Big(\tfrac{y-\mu_Y}{\sigma_Y}\Big)^2 - 2\rho\Big(\tfrac{x-\mu_X}{\sigma_X}\Big)\Big(\tfrac{y-\mu_Y}{\sigma_Y}\Big)\Big]\Big\}\,.$$

Lemma N.4.14 (see pp. CB.167-168). Suppose $(X\,,Y)$ has a bivariate normal distribution with the pdf above.

(a) $X \sim \text{Normal}(\mu_X\,,\sigma_X^2)$.

(b) $Y \sim \text{Normal}(\mu_Y\,,\sigma_Y^2)$.

(c) $\text{Corr}(X\,,Y) = \rho$.

(d) $aX + bY \sim \text{Normal}(a\mu_X + b\mu_Y\,, a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\rho\sigma_X\sigma_Y)$ for any constants a and b,

(e) $Y\,|\,X = x \sim \text{Normal}\Big(\mu_Y + \rho\sigma_Y\Big(\frac{x-\mu_X}{\sigma_X}\Big), (1-\rho^2)\sigma_Y^2\Big)$.

Lemma N.4.15. Let a and b be any constants with $a \neq 0$. Then, $X \sim \text{Normal}(\mu\,,\sigma^2)$ if and only if $aX + b \sim \text{Normal}(a\mu + b\,, a^2\sigma^2)$.

This lemma can be proved using either Theorem CB.2.1.2 or Theorem CB.2.3.5.

`Partial justification of Lemma N.4.14`: By appealing to Lemma N.4.15, it suffices to consider $Z = (X - \mu_X)/\sigma_X$ and $W = (Y - \mu_Y)/\sigma_Y$ and to show (a) $Z \sim \text{Normal}(0\,,1)$, (b) $W \sim \text{Normal}(0\,,1)$, (c) $\text{Corr}(Z\,,W) = \rho$, (d) $aZ + bW = \text{Normal}(0\,, a^2 + b^2 - 2ab\rho)$, (e) $W\,|\,Z = z \sim \text{Normal}(\rho z\,, 1-\rho^2)$. Next one can show that the transformation from $(X\,,Y)$ to $(Z\,,W)$ leads to the joint pdf

$$f_{Z,W}(z\,,w) = \frac{1}{2\pi\sqrt{1-\rho^2}}\,\exp\Big\{-\tfrac{1}{2(1-\rho^2)}\Big[z^2 + w^2 - 2\rho z w\Big]\Big\}\,.$$

Part (a) can be verified by using CB(4.1.2) and the algebraic identity $z^2 + w^2 - 2\rho z w = (w - \rho z)^2 + (1-\rho^2)z^2$. Part (b) follows from a similar calculation. For part (c) one can calculate $\text{E}(ZW) = \int\!\int zw\, f_{Z,W}(z\,,w)\text{d}z\text{d}w$ by changing the variables of integration to $s = zw$ and $t = z$ (as done on p. CB.167). Part (d) can be proved using the procedure on p. N.32. To show part (e), form $f_{W|Z}(w\,|\,z) = f_{Z,W}(z\,,w)/f_Z(z)$. □

In a linear regression model for a variable Y that depends on a variable X, we consider the distribution of Y conditional on $X = x$ and assume that $\mathrm{E}(Y \mid X = x) = \alpha + \beta x$. If we assume that the joint distribution of X and Y is bivariate normal, then $\mathrm{E}(Y \mid X = x) =$ $\mu_Y + \rho\sigma_Y\left(\frac{x-\mu_X}{\sigma_X}\right) = (\mu_Y - \rho\sigma_Y\frac{\mu_X}{\sigma_X}) + \rho\sigma_Y\frac{x}{\sigma_X}$. Thus we have

$$\alpha = \mu_Y - \beta\mu_X \quad \text{and} \quad \beta = \frac{\rho\sigma_Y}{\sigma_X} = \frac{(\frac{\sigma_{XY}}{\sigma_X\sigma_Y})\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\sigma_{XY}}{\sigma_{XX}}$$

where $\sigma_{XY} = \mathrm{Cov}(X, Y)$. Compare these population quantities with the least-squares estimates calculated from the sample:

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x} \quad \text{and} \quad \widehat{\beta} = \frac{S_{xy}}{S_{xx}}$$

where $S_{xy} = \sum(x_i - \overline{x})(y_i - \overline{y})$.

`Caution:` If we have a bivariate random vector (X, Y) and the marginal distributions of X and Y are both known to be normal, this does not imply that the random vector has a bivariate normal distribution.

Multivariate distributions

Suppose we observe an experiment that produces n random variables $X_1, X_2, \ldots, X_n$. The vector $(X_1, X_2, \ldots, X_n)$ is called a *random vector*. For example, suppose a person is randomly selected from a population and the person's height, weight, blood pressure, and cholesterol level are measured. Or suppose we randomly select four people from a population and measure their heights. These are two cases of random vectors with $n = 4$.

Discrete distributions

A random vector has a *discrete* joint distribution if there is a finite or countably infinite set C of vectors such that $\mathrm{P}[(X_1, X_2, \ldots, X_n) \in C] = 1$. Discrete distributions are often presented by the joint pmf,

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= \mathrm{P}[(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)] \\ &= \mathrm{P}[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n]. \end{aligned}$$

Probabilities can be obtained from the joint pmf as

$$\mathrm{P}[(X_1, X_2, \ldots, X_n) \in A] = \underset{(x_1, x_2, \ldots, x_n) \in A}{\sum\sum \cdots \sum} f(x_1, x_2, \ldots, x_n).$$

The marginal pmf's of X_1 and (X_1, X_2) etc. can be obtained as

$$f(x_1) = \sum_{\text{all } x_2} \cdots \sum_{\text{all } x_n} f(x_1, x_2, \ldots, x_n),$$

$$f(x_1, x_2) = \sum_{\text{all } x_3} \cdots \sum_{\text{all } x_n} f(x_1, x_2, x_3, \ldots, x_n),$$
$$\vdots$$
$$f(x_1, \ldots, x_{n-1}) = \sum_{\text{all } x_n} f(x_1, \ldots, x_{n-1}, x_n).$$

Conditional pmf's can be obtained as

$$f(x_1 \mid x_2, \ldots, x_n) = \frac{f(x_1, x_2, \ldots, x_n)}{f(x_2, \ldots, x_n)},$$
$$f(x_1, x_2 \mid x_3, \ldots, x_n) = \frac{f(x_1, x_2, x_3, \ldots, x_n)}{f(x_3, \ldots, x_n)},$$
$$\vdots$$

Expectations can be obtained as

$$\mathrm{E}[g(X_1, X_2, \ldots, X_n)] = \sum_{\text{all } x_1} \sum_{\text{all } x_2} \cdots \sum_{\text{all } x_n} g(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n).$$

Lemma N.4.4 remains true with (X, Y) replaced by $(X_1, X_2, \ldots, X_n)$. For instance,

$$\mathrm{E}[g(X_1, X_2, \ldots, X_n) + h(X_1, X_2, \ldots, X_n)]$$
$$= \mathrm{E}[g(X_1, X_2, \ldots, X_n)] + \mathrm{E}[h(X_1, X_2, \ldots, X_n)].$$

Multinomial distribution

Consider a population that is divided into k categories comprising proportions $p_1, p_2, \ldots, p_k$ of the population ($p_1 + p_2 + \cdots + p_k = 1$). Randomly select a sample of size n with replacement. Let X_j be the number of members of the sample in category j ($j = 1, \ldots, k$). Then $(X_1, X_2, \ldots, X_k) \sim \text{Multinomial}(n, p_1, p_2, \ldots, p_k)$. (Note that the length of the random vector is denoted by k here rather than n as above.)

Let us derive the joint pmf. For this we keep track of the order in which the sample was selected. Let Y_i be the category of the i-th member selected into the sample ($i = 1, \ldots, n$). Thus $(Y_1, Y_2, \ldots, Y_n)$ contains all the information that $(X_1, X_2, \ldots, X_k)$ does plus the information about the order of selection (which is not of much interest in itself but will help us in deriving the joint pmf that we are seeking).

The Y_i's have the advantage of being mutually independent of one another (we will discuss mutual independence in more detail below). The X_j's are not independent because $X_1 + X_2 + \cdots + X_k = n$.

$$\mathrm{P}(Y_1 = j_1, Y_2 = j_2, \ldots, Y_n = j_n)$$
$$= \mathrm{P}(Y_1 = j_1)\,\mathrm{P}(Y_2 = j_2) \cdots \mathrm{P}(Y_n = j_n) \quad \text{[by independence]}$$

$$= p_{j_1} p_{j_2} \cdots p_{j_n} .$$

Note that

$$\begin{aligned} X_1 &= \text{the number of } Y_i\text{'s in category } 1 , \\ X_2 &= \text{the number of } Y_i\text{'s in category } 2 , \\ &\vdots \\ X_k &= \text{the number of } Y_i\text{'s in category } k . \end{aligned}$$

Thus

$$\mathrm{P}(Y_1 = j_1 , Y_2 = j_2 , \ldots , Y_n = j_n) = p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

where x_1 is the number of j_i's equal to 1, x_2 is the number of j_i's equal to 2, and x_k is the number of j_i's equal to k. How many vectors $(j_1 , j_2 , \ldots , j_n)$ yield the same vector $(x_1 , x_2 , \ldots , x_k)$? All such vectors of j_i's can be constructed by choosing x_1 entries to be 1, x_2 entries to be 2, and so forth up to choosing x_k entries to be k. The number of ways to do this is $\binom{n}{x_1 , x_2 , \ldots , x_k} = \frac{n!}{x_1 ! \, x_2 ! \ldots x_k !}$. Therefore

$$\mathrm{P}(X_1 = x_1 , X_2 = x_2 , \ldots , X_k = x_k) = \frac{n!}{x_1 ! \, x_2 ! \ldots x_k !} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

if $x_1 + x_2 + \cdots + x_k = n$. This is Definition CB.4.6.1.

In the case $k = 2$, we have

$$\mathrm{P}(X_1 = x_1 , X_2 = x_2) = \frac{n!}{x_1 ! \, x_2 !} p_1^{x_1} p_2^{x_2}$$

which can be rewritten as

$$\mathrm{P}(X_1 = x_1) = \binom{n}{x_1} p_1^{x_1} (1 - p_1)^{n - x_1} .$$

Thus we see that $X_1 \sim \text{Binomial}(n , p_1)$. In fact this is true for general k. The n selections from the population, at random with replacement, can be regarded as n independent trials and members of category 1 can be regarded as "successes". The probability of a success in any single trial is p_1.

If we regard members of categories 1 and 2 as "successes", then we see that $X_1 + X_2 \sim \text{Binomial}(n , p_1 + p_2)$.

Lemma N.4.16. Suppose $(X_1 , X_2 , \ldots , X_k) \sim \text{Multinomial}(n , p_1 , p_2 , \ldots , p_k)$. Then

(a) $X_j \sim \text{Binomial}(n , p_j)$ for all j.

(b) $X_j + X_r \sim \text{Binomial}(n , p_j + p_r)$ for all $j \neq r$.

(c) $\text{Cov}(X_j , X_r) = - n p_j p_r$ for all $j \neq r$.

`Justification of (c):` Consider $j = 1$ and $r = 2$. A trick for deriving (c) that avoids the calculation of $\mathrm{E}(X_1X_2) = \sum\sum\limits_{x_1+x_2+\cdots+x_k=n}\cdots\sum x_1x_2 \frac{n!}{x_1!x_2!\ldots x_k!} p_1^{x_1}p_2^{x_2}\cdots p_k^{x_k}$ is the following. $\mathrm{Var}(X_1+X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + 2\,\mathrm{Cov}(X_1, X_2)$, so $\mathrm{Cov}(X_1, X_2) = \frac{1}{2}[\mathrm{Var}(X_1+X_2) - \mathrm{Var}(X_1) - \mathrm{Var}(X_2)] = \frac{1}{2}[n(p_1+p_2)(1-p_1-p_2) - np_1(1-p_1) - np_2(1-p_2) = \frac{1}{2}[-np_1p_2 - np_1p_2] = -np_1p_2$. □

The correlation of X_1 and X_2 is $\mathrm{Corr}(X_1, X_2) = -\sqrt{\frac{p_1p_2}{(1-p_1)(1-p_2)}}$.

Continuous distributions

A random vector has a *continuous* joint distribution if it has a joint pdf $f(x_1, x_2, \ldots, x_n)$ satisfying

$$\mathrm{P}[(X_1, X_2, \ldots, X_n) \in D] = \underset{D}{\int\int\cdots\int} f(x_1, x_2, \ldots, x_n)\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n$$

where $\underset{D}{\int\int\cdots\int}$ indicates that the integration is done within the limits specified by the event D.

The marginal pmf's of X_1 and (X_1, X_2) etc. can be obtained as

$$\begin{aligned}
f(x_1) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_n)\mathrm{d}x_2\cdots\mathrm{d}x_n\,, \\
f(x_1, x_2) &= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(x_1, x_2, x_3, \ldots, x_n)\mathrm{d}x_3\cdots\mathrm{d}x_n\,, \\
&\vdots \\
f(x_1, \ldots, x_{n-1}) &= \int_{-\infty}^{\infty} f(x_1, \ldots, x_{n-1}, x_n)\mathrm{d}x_n\,.
\end{aligned}$$

Conditional pmf's can be obtained as

$$\begin{aligned}
f(x_1\,|\,x_2, \ldots, x_n) &= \frac{f(x_1, x_2, \ldots, x_n)}{f(x_2, \ldots, x_n)}\,, \\
f(x_1, x_2\,|\,x_3, \ldots, x_n) &= \frac{f(x_1, x_2, x_3, \ldots, x_n)}{f(x_3, \ldots, x_n)}\,, \\
&\vdots
\end{aligned}$$

Expectations can be obtained as

$$\mathrm{E}[g(X_1, X_2, \ldots, X_n)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(x_1, x_2, \ldots, x_n)f(x_1, x_2, \ldots, x_n)\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_n\,.$$

Lemma N.4.4 remains true with (X, Y) replaced by $(X_1, X_2, \ldots, X_n)$ regardless of whether the joint distribution of $(X_1, X_2, \ldots, X_n)$ is continuous, discrete, or otherwise.

Independence

Jointly distributed random variables $X_1, X_2, \ldots, X_n$ are *mutually independent* (or simply *independent*) if

$$\mathrm{P}(X_i \in A_i \,|\, X_1 \in A_1, \ldots, X_{i-1} \in A_{i-1}, X_{i+1} \in A_{i+1}, \ldots, X_n \in A_n) = \mathrm{P}(X_i \in A_i)$$

for all events $A_1, \ldots, A_n$ and all i.

Lemma N.4.17. $X_1, X_2, \ldots, X_n$ are independent if and only if
$\mathrm{P}(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \mathrm{P}(X_1 \in A_1)\mathrm{P}(X_2 \in A_2) \cdots \mathrm{P}(X_n \in A_n)$
for all events $A_1, \ldots, A_n$.

When $n = 2$, this is Lemma N.4.7.

Lemma N.4.18. Suppose the random variables $X_1, X_2, \ldots, X_n$ have joint pmf or pdf $f(x_1, x_2, \ldots, x_n)$ and marginal pmf's or pdf's $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$. Then, the random variables are independent if and only if $f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2)\cdots f_n(x_n)$ for all $x_1, x_2, \ldots, x_n$.

When $n = 2$, this is part of Lemma N.4.8.

If $X_1, X_2, \ldots, X_n$ are independent and, moreover, have the same distribution with pmf or pdf $f(x)$, then the lemma tells us that their joint pmf or pdf is $f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n) = \prod_{i=1}^{n} f(x_i)$.

Examples. (1) Suppose $X_1, X_2, \ldots, X_n$ are independent random variables all having the Normal$(0,1)$ distribution. Their joint pdf is

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= f(x_1)f(x_2)\cdots f(x_n) \\ &= \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x_1^2}{2}}\, \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x_2^2}{2}} \cdots \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x_n^2}{2}} = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}(x_1^2 + x_2^2 + \cdots + x_n^2)} . \end{aligned}$$

(2) Suppose $X_1, X_2, \ldots, X_n$ are independent random variables all having the Bernoulli(p) distribution. Their joint pmf is

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &= f(x_1)f(x_2)\cdots f(x_n) \\ &= p^{x_1}(1-p)^{1-x_1} p^{x_2}(1-p)^{1-x_2} \cdots p^{x_n}(1-p)^{1-x_n} \\ &= p^{x_1+x_2+\cdots+x_n}(1-p)^{n-x_1-x_2-\cdots-x_n} . \quad \| \end{aligned}$$

Lemma N.4.19. Suppose $X_1, X_2, \ldots, X_n$ are independent random variables. Then
$\mathrm{E}[g_1(X_1)g_2(X_2)\cdots g_n(X_n)] = \mathrm{E}[g_1(X_1)]\mathrm{E}[g_2(X_2)]\cdots\mathrm{E}[g_n(X_n)]$.

When $n = 2$, this is Theorem CB.4.2.1.

<u>Lemma N.4.20</u>. Suppose $X_1, X_2, \ldots, X_n$ are independent random variables with mgf's $M_{X_1}(t)$, $M_{X_2}(t)$, …, $M_{X_n}(t)$. Let $Z = X_1 + X_2 + \cdots + X_n$. Then $M_Z(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t)$.

Lemma N.4.20 follows from Lemma N.4.19 by a proof similar to the proof of Theorem CB.4.2.2. The lemma implies that if $X_1, X_2, \ldots, X_n$ are independent and have the same distribution with mgf $M_X(t)$, then the mgf of $Z = X_1 + X_2 + \cdots + X_n$ is $M_Z(t) = [M_X(t)]^n$.

<u>Examples</u>. (1) Suppose $X_1, X_2, \ldots, X_n$ are independent random variables all having the Normal(μ, σ^2) distribution. (For short we say that they are i.i.d.: independent and identically distributed.) The mgf of the distribution is $M(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$. The mgf of $Z = X_1 + X_2 + \cdots + X_n$ is $[M(t)]^n = [e^{\mu t + \frac{1}{2}\sigma^2 t^2}]^n = e^{n\mu t + \frac{1}{2} n \sigma^2 t^2}$, which is the mgf of the Normal$(n\mu, n\sigma^2)$ distribution.

(2) Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli(p), which has mgf $M(t) = 1 - p + pe^t$. Then the mgf of $Z = X_1 + X_2 + \cdots + X_n$ is $[M(t)]^n = (1 - p + pe^t)^n$, which is the mgf of the Binomial(n, p) distribution.

(3) Suppose $X_1, X_2, \ldots, X_n$ are independent with $X_i \sim$ Normal(μ_i, σ_i^2). Consider $Z = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n + b$. (This example includes Example (1) as a special case.) Then the mgf of Z is $M_Z(t) = M_{a_1 X_1}(t) M_{a_2 X_2}(t) \cdots M_{a_n X_n}(t) M_b(t)$. As in Theorem CB.2.3.5, $M_{aX}(t) = M_X(at)$ and $M_b(t) = e^{bt}$. Thus $M_Z(t) = M_{X_1}(a_1 t) M_{X_2}(a_2 t) \cdots M_{X_n}(a_n t) e^{bt} = e^{\mu_1 (a_1 t) + \frac{1}{2}\sigma_1^2 (a_1 t)^2} e^{\mu_2 (a_2 t) + \frac{1}{2}\sigma_2^2 (a_2 t)^2} \cdots e^{\mu_n (a_n t) + \frac{1}{2}\sigma_n^2 (a_n t)^2} e^{bt} =$ ←.
$e^{(a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n + b)t + \frac{1}{2}(a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2)t^2}$, which is the mgf of the Normal$(a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n + b, a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2)$ distribution. ||

<u>Inequalities</u>

<u>CB(4.7.4)</u> (Cauchy-Schwarz Inequality). $|\mathrm{E}(XY)| \le \sqrt{\mathrm{E}(X^2)\mathrm{E}(Y^2)}$.

`Justification`: This inequality follows from the fact that $\mathrm{E}[\left(Y - \frac{\mathrm{E}(XY)}{\mathrm{E}(X^2)} X\right)^2] \ge 0$. □

Replacing X by $X - \mu_X$ and Y by $Y - \mu_Y$ in the inequality yields $|\mathrm{Cov}(X, Y)| \le \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$, which is equivalent to $|\rho_{XY}| \le 1$.

For the next inequality, we need a definition.

<u>Definition N.4.21</u>. A function $g(x)$ is said to be *convex* if, for any two points on its graph, the line segment joining the two points lies entirely on or above the graph.

<u>Theorem N.4.22</u> (Jensen's Inequality). For a convex function $g(x)$, $\mathrm{E}[g(X)] \ge g(\mathrm{E}[X])$.

(To be technically correct, we should specify that $g(x)$ is defined on an interval $(a\,,b)$ and that $\mathrm{P}(a < X < b) = 1$.)

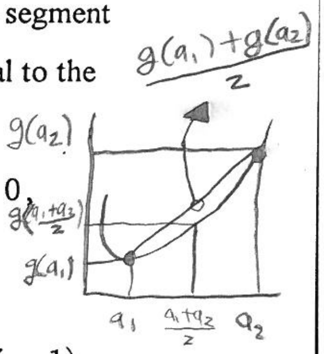
Lemma N.4.23. Suppose $g(x)$ is twice differentiable. It is convex if and only if $g''(x) \geq 0$ for all x (where $g''(x)$ denotes the second derivative of $g(x)$).

Examples. (1) $g(x) = x^2$. Check that $g'(x) = 2x$, $g''(x) = 2 \geq 0$, so the function is convex. Therefore, $\mathrm{E}[X^2] \geq (\mathrm{E}[X])^2$. This simply says that $\mathrm{Var}(X) \geq 0$.

(2) $g(x) = ax + b$. Check that $g'(x) = a$, $g''(x) = 0 \geq 0$, so the function is convex. In this case, the line segment between two points on the graph lies on, rather than above, the graph. By Jensen's Inequality, $\mathrm{E}(aX + b) \geq a\mathrm{E}(X) + b$. In this case they are actually equal.

(3) $g(x) = \frac{1}{x}$ for $x > 0$. Check that $g'(x) = -\frac{1}{x^2}$, $g''(x) = \frac{2}{x^3} \geq 0$ for $x > 0$, so the function is convex. Therefore, if $\mathrm{P}(X > 0) = 1$, then $\mathrm{E}(\frac{1}{X}) \geq \frac{1}{\mathrm{E}(X)}$. A special case occurs when X has a discrete distribution with all its probability on n equally likely points. That is, suppose $\mathrm{P}(X = a_1) = \mathrm{P}(X = a_2) = \cdots = \mathrm{P}(X = a_n) = \frac{1}{n}$. Then $\mathrm{E}(X) = a_1(\frac{1}{n}) + a_2(\frac{1}{n}) + \cdots + a_n(\frac{1}{n}) = \frac{1}{n}(a_1 + a_2 + \cdots + a_n) =$ the arithmetic mean of the a_i's. And $\mathrm{E}(\frac{1}{X}) = \frac{1}{a_1}(\frac{1}{n}) + \frac{1}{a_2}(\frac{1}{n}) + \cdots + \frac{1}{a_n}(\frac{1}{n}) = \frac{1}{n}(\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n})$. The quantity $1/\mathrm{E}(\frac{1}{X}) = n/(\frac{1}{a_1} + \frac{1}{a_2} + \cdots + \frac{1}{a_n})$ is called the harmonic mean of the a_i's. Jensen's Inequality implies that $1/\mathrm{E}(\frac{1}{X}) \leq \mathrm{E}(X)$; that is, the harmonic mean is always less than or equal to the arithmetic mean. ||

`Partial justification of Jensen's Inequality`: Consider the case when X is a discrete random variable with only two possible values, each having probability $\frac{1}{2}$. That is, $\mathrm{P}(X = a_1) = \mathrm{P}(X = a_2) = \frac{1}{2}$. Then $\mathrm{E}(X) = a_1(\frac{1}{2}) + a_2(\frac{1}{2}) = \frac{a_1 + a_2}{2}$, and $\mathrm{E}[g(X)] = g(a_1)(\frac{1}{2}) + g(a_2)(\frac{1}{2}) = \frac{g(a_1) + g(a_2)}{2}$. Jensen's Inequality says $\frac{g(a_1)+g(a_2)}{2} \geq g(\frac{a_1+a_2}{2})$. Why should this be true? Consider a graph of the function $y = g(x)$. Two points on the graph are $(a_1\,, g(a_1))$ and $(a_2\,, g(a_2))$. Above the value $x = \frac{a_1+a_2}{2}$, the line segment joining these two points has height $\frac{g(a_1)+g(a_2)}{2}$. The convexity of $g(x)$ implies that the line segment lies on or above the graph. So the height $\frac{g(a_1)+g(a_2)}{2}$ should be greater than or equal to the height of the graph above the value $x = \frac{a_1+a_2}{2}$, which is $g(\frac{a_1+a_2}{2})$. □



Theorem N.4.24 (Markov's Inequality). Suppose $\mathrm{P}(Y \geq 0) = 1$. Then, for any $r > 0$, $\mathrm{P}(Y \geq r) \leq \dfrac{\mathrm{E}(Y)}{r}$.

`Justification`: Define $Z = 1$ if $Y \geq r$ and $Z = 0$ if $Y < r$. Now, $\mathrm{E}(Y) = \mathrm{E}[\mathrm{E}(Y \mid Z)] = \mathrm{E}(Y \mid Z = 1)\mathrm{P}(Z = 1) + \mathrm{E}(Y \mid Z = 0)\mathrm{P}(Z = 0) \geq \mathrm{E}(Y \mid Z = 1)\mathrm{P}(Z = 1) = \mathrm{E}(Y \mid Y \geq r)\mathrm{P}(Y \geq r) \geq r\,\mathrm{P}(Y \geq r)$. □