Discrete Mathematics and Probability Theory

Computer Science 70, Spring 2016

Sinho Chewi

Contents

Chapter 1

Discrete Random Variables: Expectation, and Distributions

We discuss random variables and see how they can be used to model common situations. We will see that the expectation of a random variable is a useful property of the distribution that satisfies an important property: linearity. We also introduce common discrete probability distributions.

1.1 Random Variables: Review

Recall that a **random variable** is a function $X : \Omega \to \mathbb{R}$ that assigns a real number to every outcome ω in the probability space. We typically denote them by capital letters.

We define addition of random variables in the following way: the random variable $X + Y$ is the random variable that maps ω to $X(\omega) + Y(\omega)$. Similarly, the random variable XY is the random variable that maps ω to $X(\omega)Y(\omega)$. More generally, let $f : \mathbb{R}^n \to \mathbb{R}$ be any function. Then $f(X_1, \ldots, X_n)$ is defined to be the random variable that maps ω to $f(X_1(\omega), \ldots, X_n(\omega))$.

We say that two random variables are **independent** if

$$\forall x, y \in \mathbb{R} \ \Pr(X = x, Y = y) = \Pr(X = x) \Pr(Y = y) \tag{1.1}$$

The **distribution** of a random variable is the set of possible values of the random variable, along with their respective probabilities. Typically, the distribution of a random variable is specified by giving a formula for $\Pr(X = k)$. We use the symbol $\sim$ to denote that a random variable has a known distribution, e.g. $X \sim \text{Bin}(n, p)$ means that X follows the $\text{Bin}(n, p)$ distribution.

Equivalently, we can describe a probability distribution by its **cumulative distribution function**, or its **c.d.f.** function. The c.d.f. is usually specified as a formula for $\Pr(X \leq k)$. The c.d.f. contains just as much information as the original distribution. To see this fact, observe that we can recover the probability distribution function (also known as the p.d.f.) from the c.d.f. by the following formula

$$\boxed{\Pr(X = k) = \Pr(X \leq k) - \Pr(X \leq k - 1)} \tag{1.2}$$

(assuming X takes on integer values).

The **joint distribution** of two random variables X and Y is the probability $\Pr(X = j, Y = k)$ for all possible pairs of values (j, k). The joint distribution must satisfy the normalization condition

$$\boxed{\sum_j \sum_k \Pr(X = j, Y = k) = 1} \tag{1.3}$$

We can recover the distribution of X separately (known as the **marginal distribution** of X) by summing over all possible values of Y:

$$\boxed{\Pr(X = j) = \sum_k \Pr(X = j, Y = k)} \tag{1.4}$$

Notice the utility of independence: if X and Y are independent, then we can write their joint probability as a product of their marginal probabilities ($\Pr(X = j, Y = k) = \Pr(X = j)\Pr(Y = k)$), which immensely simplifies calculations. The results easily generalize to multiple random variables.

1.2 Expectation

Knowing the full probability distribution gives us a lot of information, but sometimes it is helpful to have a summary of the distribution. The **expectation** or **expected value** is the average value of a random variable. Two equivalent equations for the expectation are given below:

$$\boxed{E(X) = \sum_{\omega\in\Omega} X(\omega)\Pr(\omega) = \sum_k k\Pr(X = k)} \tag{1.5}$$

The interpretation of the expected value is as follows: pick N outcomes, $\omega_1, \ldots, \omega_N$ from a probability distribution (we call this N trials of an experiment). For each trial, record the value of $X(\omega_i)$. Then

$$\frac{X(\omega_1) + \cdots + X(\omega_N)}{N} \to E(X)$$

as $N \to \infty$. Therefore, $E(X)$ is the *long-run average* of an experiment in which you measure the value of X.

Often, the expectation values are easier to work with than the full probability distributions because they satisfy nice properties. In particular, they satisfy **linearity**: suppose X, Y are random variables, $a \in \mathbb{R}$ is a constant, and c is the constant random variable (i.e. $\forall\omega \in \Omega\, c(\omega) = c$). Then:

1. $E(X + Y) = E(X) + E(Y)$
2. $E(aX) = aE(X)$
3. $E(X + c) = E(X) + c$

We will use these properties repeatedly to solve complicated problems.

In the previous section, we noted that if X is a random variable and $f : \mathbb{R} \to \mathbb{R}$ is a function, then $f(X)$ is a random variable. The expectation of $f(X)$ is defined as follows:

$$\boxed{E(f(X)) = \sum_{\omega\in\Omega} f(X(\omega))\Pr(\omega) = \sum_k f(k)\Pr(X = k)} \tag{1.6}$$

The definition can be extended easily to functions of multiple random variables using the joint distribution:

$$\boxed{E(f(X_1, \ldots, X_n)) = \sum_{x_1,\ldots,x_n} f(x_1, \ldots, x_n)\Pr(X_1 = x_1, \ldots, X_n = x_n)} \tag{1.7}$$

Next, we prove an important fact about the expectation of independent random variables.

Theorem 1.1 (Expectation of Independent Random Variables)**.** *Let X and Y be independent random variables. Then the random variable XY satisfies*

$$E(XY) = E(X)E(Y) \tag{1.8}$$

Proof.

$$\begin{aligned} E(XY) &= \sum_{x,y} xy \Pr(X = x, Y = y) \\ &= \sum_{x,y} xy \Pr(X = x) \Pr(Y = y) \\ &= \left(\sum_{x} x \Pr(X = x) \right) \left(\sum_{y} y \Pr(Y = y) \right) \\ &= E(X)E(Y) \end{aligned}$$

Observe that the definition of independent random variables was used in line 2 of the proof. It is crucial to remember that *the theorem does not hold true when X and Y are not independent!* □

1.2.1 Tail Sum Formula

Next, we derive an important formula for computing the expectation of a probability distribution.

Theorem 1.2 (Tail Sum Formula)**.** *Let X be a random variable that only takes on values in $\mathbb{N}$. Then*

$$E(X) = \sum_{k=1}^{\infty} \Pr(X \geq k)$$

Proof. We manipulate the formula for the expectation:

$$\begin{aligned} E(X) &= \sum_{x=1}^{\infty} x \Pr(X = x) \\ &= \sum_{x=1}^{\infty} \sum_{k=1}^{x} \Pr(X = x) \\ &= \sum_{k=1}^{\infty} \sum_{x=k}^{\infty} \Pr(X = x) \\ &= \sum_{k=1}^{\infty} \Pr(X \geq k) \end{aligned}$$

□

The formula is known as the *tail sum formula* because we compute the expectation by summing over the tail probabilities of the distribution.

1.3 Important Probability Distributions

We will now give many important examples of probability distributions and their expectations.

1.3.1 Uniform Distribution

As a first example of probability distributions, we will consider the uniform distribution over the set $\{1, \ldots, n\}$, typically denoted as $\text{Unif}\{1, \ldots, n\}$. The meaning of *uniform* is that each element of the set is equally likely to be chosen; therefore, the probability distribution is

$$\Pr(X = k) = \frac{1}{n}, \quad k \in \{1, \ldots, n\}$$

The expectation of the uniform distribution is calculated fairly easily from the definition:

$$\begin{aligned} E(X) &= \sum_{k=1}^{n} k \cdot \frac{1}{n} \\ &= \frac{1}{n} \sum_{k=1}^{n} k \\ &= \frac{1}{n} \cdot \frac{n(n+1)}{2} \\ &= \frac{n+1}{2} \end{aligned}$$

where to evaluate the sum, we have used the triangular number identity (easily proven using induction):

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2} \tag{1.9}$$

1.3.2 Bernoulli Distribution

The Bernoulli distribution with parameter p, denoted $\text{Ber}(p)$, is a very simple distribution that describes the result of performing one experiment which succeeds with probability p. Define the probability space $\Omega = \{\text{Success}, \text{Failure}\}$ with $\Pr(\text{Success}) = p$ and $\Pr(\text{Failure}) = 1 - p$. Then,

$$X(\omega) = \begin{cases} 0, & \omega = \text{Failure} \\ 1, & \omega = \text{Success} \end{cases}$$

The distribution of X is

$$\Pr(X = k) = \begin{cases} 1 - p, & k = 0 \\ p, & k = 1 \\ 0, & \text{otherwise} \end{cases}$$

The expectation of the $\text{Ber}(p)$ distribution is

$$E(X) = 0 \cdot \Pr(X = 0) + 1 \cdot \Pr(X = 1) = 0 \cdot (1 - p) + 1 \cdot p = p$$

A quick example: the number of heads in one fair coin flip follows the $\text{Ber}(1/2)$ distribution.

1.3.3 Indicator Random Variables

Let $A \subseteq \Omega$ be an event. We define the indicator of A, 1_A, to be the random variable

$$1_A(\omega) = \begin{cases} 0, & \omega \notin A \\ 1, & \omega \in A \end{cases}$$

Observe that 1_A follows the $\text{Ber}(p)$ distribution where $p = \Pr(A)$.

An important property of indicator random variables (and Bernoulli random variables) is that $X = X^2 = X^k$ for any $k \geq 1$. To see why this is true, note that X can only take on values in the set $\{0, 1\}$. Since $0^2 = 0$ and $1^2 = 1$, then $\forall \omega \in \Omega \, X(\omega) = X^2(\omega)$. By induction, we can prove that $X = X^k$ for $k \geq 1$. We will use this property when we discuss the variance of probability distributions.

The expectation of the indicator random variable is

$$\boxed{E(1_A) = \Pr(A)} \tag{1.10}$$

(because it is a Bernoulli random variable with $p = \Pr(A)$).

1.3.4 Binomial Distribution

The binomial distribution with parameters n and p, abbreviated $\text{Bin}(n, p)$, describes the number of successes when we conduct n independent trials, where each trial has a probability p of success. The binomial distribution is found by the following argument: the probability of having a series of trials with k successes (and therefore $n - k$ failures) is $p^k(1-p)^{n-k}$. We need to multiply this expression by the number of ways to achieve k successes in n trials, which is $\binom{n}{k}$. Hence

$$\Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0, \ldots, n\}$$

Prove for yourself that the probabilities sum to 1, i.e.

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1 \tag{1.11}$$

Let us proceed to compute the expectation of this distribution. According to the formula,

$$E(X) = \sum_k k \Pr(X = k) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

This is quite a difficult sum to calculate! (Try it yourself and see if you can make any progress. For details on how one might compute this sum, see the bonus section at the end.) To make our work simpler, we will instead make a connection between the binomial distribution and the Bernoulli distribution we defined earlier. Let X_i be the indicator random variable for the event that trial i is a success. (In the language of the previous section, $X_i = 1_{A_i}$, where A_i is the event that trial i is a success.) The key insight lies in observing

$$X = X_1 + \cdots + X_n$$

Essentially, each indicator variable X_i is 1 or 0 depending on whether trial i is a success, so if we sum up all of the indicator variables, then we obtain the total number of successes in all n trials. Therefore, compute

$$\begin{aligned} E(X) &= E(X_1 + \cdots + X_n) \\ &= E(X_1) + \cdots + E(X_n) \end{aligned}$$

Notice that in the last line, we used linearity of expectation. Now we can finally see why linearity of expectation is so powerful: combined with indicator variables, it allows us to break up the expectation of a complicated random variable into the sum of the expectations of simple random variables. Indeed, the variables X_i are very simple. Using our result from the previous section on indicator random variables,

$$E(X_i) = \Pr(\text{trial } i \text{ is a success}) = p$$

Each term in the sum is simply p, and there are n such terms, so therefore

$$E(X) = np$$

The result should make intuitive sense: if you are conducting n trials, and the probability of success is p, then you expect a fraction p of the trials to be successes, which is saying that you expect np total successes. The expectation matches our intuition.

By the way, the random variables X_i are an example of i.i.d. random variables, which is a term that comes up very frequently (so we might as well define it now): i.i.d. stands for *independent and identically distributed*. Indeed, since each trial is independent of each other by assumption, the variables X_i are independent, although we did not need this fact to compute the expectation. Linearity of expectation is powerful: it holds even when the variables are not independent! Also, the X_i variables were identically distributed, which means they all had the same probability distribution: $X_i \sim \text{Ber}(p)$.

A strategy now emerges for tackling complicated expected value questions: when computing $E(X)$, try to see if you can break down X into the sum of indicator random variables. Then, computing the expectation becomes much easier because you can take advantage of linearity.

1.3.5 Geometric Distribution

The geometric distribution with parameter p, abbreviated Geom(p), describes the number of trials required to obtain the first success, assuming that each trial has a probability of success p. If it takes exactly k trials to obtain the first success, there were first $k-1$ failures (each with probability $1-p$) and one success (with probability p). Hence, the distribution is

$$\Pr(X = k) = (1-p)^{k-1}p, \quad k > 0$$

Prove for yourself that the probabilities sum to 1, i.e.

$$\sum_{k=1}^{\infty}(1-p)^{k-1}p = 1 \tag{1.12}$$

When working with the geometric distribution, it is often easier to work with the tail probabilities $\Pr(X > k)$. In order for $X > k$ to hold, there must be at least k failures; hence,

$$\Pr(X > k) = (1-p)^k$$

Note that the tail probability is related to the c.d.f. in the following way: $\Pr(X > k) = 1 - \Pr(X \leq k)$.

The clever way to find the expectation of the geometric distribution uses a method known as the renewal method. $E(X)$ is the expected number of trials until the first success. Suppose we carry out the first trial, and one of two outcomes occurs. With probability p, we obtain a success and we are done (it only took 1 trial until success). With probability $1-p$, we obtain a failure, and we are right back where we started. In the latter case, how many trials do we expect until our first success? The answer is $1 + E(X)$: we have already used one trial, and we expect $E(X)$ more since nothing has changed from our original situation (the geometric distribution is memoryless). Hence

$$E(X) = p \cdot 1 + (1-p) \cdot (1 + E(X))$$

Solving this equation yields

$$E(X) = \frac{1}{p}$$

which is also intuitive: if we have, say, a 1/100 chance of success on each trial, we would naturally expect 100 trials until our first success. (Note: if the method above does not seem rigorous to you, then worry not. We will revisit the method under the framework of conditional expectation in the future.)

Here is a more computational way to obtain the formula. We are looking to evaluate the sum

$$E(X) = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = p\sum_{k=0}^{\infty}(k+1)(1-p)^k$$

Recall the following identity (from calculus, and geometric series):

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x} \tag{1.13}$$

Multiply both sides of the identity by x:

$$\sum_{k=0}^{\infty} x^{k+1} = \frac{x}{1-x}$$

Differentiate both sides with respect to x:

$$\sum_{k=0}^{\infty}(k+1)x^k = \frac{1}{(1-x)^2}$$

Set $x = 1-p$ to evaluate our original sum:

$$E(X) = p \cdot \frac{1}{(1-(1-p))^2} = p \cdot \frac{1}{p^2} = \frac{1}{p}$$

A third way of finding the expectation is to use the Tail Sum Formula (Theorem 1.2).

Memoryless Property

An important property of the geometric distribution is that it is *memoryless*, which is to say that a random variable following the geometric distribution only depends on its current state and not on its past. To make this notion formal, we shall show:

Theorem 1.3 (Memoryless Property)**.** *The geometric distribution satisfies*

$$\Pr(X > k + t \mid X > k) = Pr(X > t)$$

Proof.

$$\begin{aligned}
\Pr(X > k + t \mid X > k) &= \frac{\Pr(X > k + t, X > k)}{\Pr(X > k)} \\
&= \frac{\Pr(X > k + t)}{\Pr(X > k)} \\
&= \frac{(1-p)^{k+t}}{(1-p)^k} \\
&= (1-p)^t \\
&= \Pr(X > t)
\end{aligned}$$

□

Intuitively, the theorem says: suppose you have already tried flipping a coin k times, without success. The probability that it takes you at least t more coin flips until your first success is the *same* as the probability that your friend picks up a coin and it takes him/her at least t coin flips. Moral of the story: the geometric distribution does not care how many times you have already flipped the coin, because it is *memoryless*.

1.3.6 Poisson Distribution

The Poisson distribution with parameter λ, abbreviated Pois(λ), is an approximation to the binomial distribution under certain conditions: let the number of trials, n, approach infinity and the probability of success per trial, p, approach 0, such that the mean $E(X) = np$ remains a fixed value λ. An example is: suppose we cultivate 1000 agar dishes, and each dish has a probability 1/1000 that a bacterial colony will grow. The mean is $\lambda = np = 1$. Under these assumptions, k is typically very small compared to n, so that

$$\begin{aligned}
\Pr(X = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k} \\
&\approx \frac{n^k p^k}{k!} \left(1 - \frac{\lambda}{n}\right)^n \\
&\approx \frac{\lambda^k e^{-\lambda}}{k!}
\end{aligned}$$

where in the last line, we have used the identity

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x \tag{1.14}$$

The previous section was to motivate the form of the Poisson distribution. We now define the Poisson distribution:

$$\Pr(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k \geq 0$$

Check for yourself that the probabilities sum to 1, i.e.

$$\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = 1 \tag{1.15}$$

Remember to use the important identity

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} \tag{1.16}$$

(In many cases, the power series is seen as the definition of e^x. The power series converges for all real x.)

The expectation of the Poisson distribution is, as we would expect, λ. But let's prove it:

$$\begin{aligned}
E(X) &= \sum_{k=0}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} \\
&= \sum_{k=1}^{\infty} k \frac{e^{-\lambda}\lambda^k}{k!} \\
&= e^{-\lambda}\lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\
&= e^{-\lambda}\lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\
&= e^{-\lambda}\lambda e^{\lambda} \\
&= \lambda
\end{aligned}$$

Sums of Poisson Random Variables

Here, we will prove an important fact about the sums of Poisson random variables.

Theorem 1.4 (Sums of Independent Poisson Random Variables)**.** *Let $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$ be independent random variables. Then*

$$X + Y \sim \text{Pois}(\lambda + \mu)$$

Proof. We will compute the distribution of $X + Y$ and show that it is Poisson (using independence).

$$\begin{aligned}
\Pr(X + Y = k) &= \sum_{j=0}^{k} \Pr(X = j, Y = k - j) \\
&= \sum_{j=0}^{k} \frac{e^{-\lambda}\lambda^j}{j!} \frac{e^{-\mu}\mu^{k-j}}{(k-j)!} \\
&= \frac{e^{-(\lambda+\mu)}}{k!} \sum_{j=0}^{k} \frac{k!}{j!\,(k-j)!} \lambda^j \mu^{k-j} \\
&= \frac{e^{-(\lambda+\mu)}}{k!} \sum_{j=0}^{k} \binom{k}{j} \lambda^j \mu^{k-j} \\
&= \frac{e^{-(\lambda+\mu)}(\lambda+\mu)^k}{k!} \\
&= \Pr(\text{Pois}(\lambda + \mu) = k)
\end{aligned}$$

In the second to last line, we have used the binomial theorem. □

Poisson Thinning

Here, we introduce another important property of the Poisson distribution known as the **Poisson thinning** property, which will be proven first and motivated later.

Theorem 1.5 (Poisson Thinning)**.** *Suppose that* $X \sim \text{Pois}(\lambda)$ *and that given* $X = j$*,* Y *follows the* $\text{Bin}(j, p)$ *distribution. In the notation of probability theory,* $Y \mid X = j \sim \text{Bin}(j, p)$*. Then* $Y \sim \text{Pois}(\lambda p)$*.*

Proof. We use the definition of conditional probability and show that Y has the correct distribution. Notice that the sum starts from $j = k$ because $X \geq Y$.

$$\begin{aligned}
\Pr(Y = k) &= \sum_{j=k}^{\infty} \Pr(X = j, Y = k) \\
&= \sum_{j=k}^{\infty} \Pr(X = j) \Pr(Y = k \mid X = j) \\
&= \sum_{j=k}^{\infty} \frac{e^{-\lambda}\lambda^j}{j!} \binom{j}{k} p^k (1-p)^{j-k} \\
&= e^{-\lambda} \sum_{j=k} \frac{\lambda^j}{j!} \frac{j!}{k!\,(j-k)!} p^k (1-p)^{j-k} \\
&= \frac{e^{-\lambda}(\lambda p)^k}{k!} \sum_{j=k} \frac{(\lambda(1-p))^{j-k}}{(j-k)!} \\
&= \frac{e^{-\lambda}(\lambda p)^k}{k!} e^{\lambda(1-p)} \\
&= \frac{e^{-\lambda p}(\lambda p)^k}{k!} \\
&= \Pr(\text{Pois}(\lambda p) = k)
\end{aligned}$$

□

As an example: suppose that the number of calls that a calling center receives per hour is distributed according to a Poisson distribution with mean λ. Furthermore, suppose that each call that the calling center receives is independently a telemarketer with probability p (therefore, the distribution of telemarketing calls is binomial, conditioned on the number of calls received). Then, the number of telemarketing calls that the calling receives per hour (unconditional) follows a Poisson distribution with mean λp. This property of the Poisson distribution is also rather intuitive because it says that if a Poisson random variable is thinned out such that only a fraction p remains, then the resulting distribution remains Poisson. However, take time to appreciate that the Poisson thinning property is not immediately obvious without the mathematical proof.

1.4 Bonus: Computing a Difficult Sum

We now proceed to present a clever trick for solving

$$E(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$

without using linearity of expectation. Of course, this sum is still too difficult for us to solve initially, so we will consider a different sum instead. First, recall the Binomial Theorem:

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k} \tag{1.17}$$

Then, notice we can rewrite the following sum using the Binomial Theorem:

$$\sum_{k=0}^{n} e^{tk} \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (pe^t)^k (1-p)^{n-k} = (1-p+pe^t)^n$$

Differentiate both sides with respect to t:

$$\sum_{k=0}^{n} ke^{tk}\binom{n}{k}p^k(1-p)^{n-k} = n(1-p+pe^t)^{n-1}pe^t$$

Finally, set $t = 0$:

$$\sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} = np$$

Wait, wasn't that the original sum we were trying to solve? Wow! Hopefully this blew your mind, at least a little bit. ;) We next demystify the solution with three motivating observations:

Observation 1 The method described above is actually considerably more general than it might first appear. Observe that if you differentiate with respect to t twice, a little more work quickly produces $E(X^2)$. Differentiating a third time will produce $E(X^3)$, and so forth.

Observation 2 The sum we chose was not random. Observe that what we were essentially computing is

$$\boxed{M(t) := E(e^{tX})} \tag{1.18}$$

Furthermore, observe that $M'(0) = E(X)$, $M''(0) = E(X^2)$, and in general,

$$\boxed{M^{(k)}(0) = E(X^k)} \tag{1.19}$$

$E(X^k)$ is called the kth moment of X; hence, $M(t)$ is known as the **moment generating function** or the **m.g.f.** for short. (Note that $E(e^{tX})$ is not always defined.)

Observation 3 We used the Binomial Theorem to simplify the complicated sum. Observe that if X is the sum of i.i.d. random variables (see 1.3.3), i.e. $X = X_1 + \cdots X_n$, then

$$E(e^{tX}) = E(e^{t(X_1+\cdots+X_n)}) = E(e^{tX_1}\cdots e^{tX_n})$$

Since $X_1, \ldots X_n$ are independent, then using Theorem 1.1,

$$E(e^{tX_1}\cdots e^{tX_n}) = E(e^{tX_1})\cdots E(e^{tX_n})$$

Since each X_i has the same distribution (identically distributed), then we can simply write

$$E(e^{tX_1})\cdots E(e^{tX_n}) = \left[E(e^{tX_i})\right]^n$$

Observe that it was no accident that we obtained something of the form (expression)^n. Although applying the Binomial Theorem successfully seemed like a shining piece of luck, its success was inevitable.

We will say no more about moment generating functions and such. This was merely meant to wet your taste buds, so that you can explore further on your own if you desire.

Chapter 2

Variance and Probability Bounds

Previously, we have discussed the expectation of a random variable, which is a measure of the center of the probability distribution. Today, we discuss the variance, which is a measure of the *spread* of the distribution. Variance, in a sense, is a measure of how unpredictable your results are. We will also cover some important inequalities for bounding tail probabilities.

2.1 Variance

Suppose your friend offers you a choice: you can either accept \$1 immediately, or you can enter in a raffle in which you have a 1/100 chance of winning a \$100 payoff. The expected value of each of these deals is simply \$1, but clearly the offers are very different in nature! We need another measure of the probability distribution that will capture the idea of *variability* or *risk* in a distribution. We are now interested now in how often a random variable will take on values close to the mean.

Perhaps we could study the quantity $X - E(X)$ (the difference between what we expected and what we actually measured), but we quickly notice a problem:

$$E(X - E(X)) = E(X) - E(X) = 0 \tag{2.1}$$

The expectation of this quantity is always 0, no matter the distribution! Every random variable (except the constant random variable) can take on values above or below the mean by definition, so studying the average of the differences is not interesting. To address this problem, we could study $|X - E(X)|$ (thereby making all differences positive), but in practice, it becomes much harder to analytically solve problems using this quantity. Instead, we will study a quantity known as the variance of the probability distribution:

Definition 2.1 (Variance)**.** The **variance** of a probability distribution is

$$\boxed{\mathrm{Var}(X) = E((X - E(X))^2)} \tag{2.2}$$

Remark 2.2. We often denote the mean of the probability distribution as μ, and the variance as σ^2. We call $\sigma = \sqrt{\mathrm{Var}(X)}$ the **standard deviation** of X. The standard deviation is useful because it has the same units as X, allowing for easier comparison. As an example, if X represents the height of an individual, then σ_X would have units of meters, while $\mathrm{Var}(X)$ has units of meters^2.

2.1.1 The Computational Formula

The explicit formula for variance is

$$\boxed{\mathrm{Var}(X) = \sum_k (k - E(X))^2 \Pr(X = k)} \tag{2.3}$$

In practice, however, we tend to use the following formula to calculate variance:

Theorem 2.3 (Computational Formula for Variance)**.** *The variance of X is*

$$\boxed{\mathrm{Var}(X) = E(X^2) - E(X)^2} \tag{2.4}$$

Proof. We use linearity of expectation (note that $E(E(X)) = E(X)$ since $E(X)$ is just a constant):

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

□

This formula will be extremely useful to us throughout the course, so please memorize it!

2.1.2 Properties of the Variance

We next examine some useful properties of the variance.

Theorem 2.4 (Properties of the Variance)**.** *Let X be a random variable, $a \in \mathbb{R}$ be a constant, and c be the constant random variable. Then*

$$\boxed{\mathrm{Var}(aX + c) = a^2 \,\mathrm{Var}(X)} \tag{2.5}$$

Proof. We use the computational formula:

$$\begin{aligned} \mathrm{Var}(aX + c) &= E((aX + c)^2) - E(aX + c)^2 \\ &= E(a^2X^2 + 2acX + c^2) - (aE(X) + c)^2 \\ &= a^2E(X^2) + 2acE(X) + c^2 - a^2E(X)^2 - 2acE(X) - c^2 \\ &= a^2(E(X^2) - E(X)^2) \\ &= a^2 \,\mathrm{Var}(X) \end{aligned}$$

□

Corollary 2.5 (Scaling of the Standard Deviation)**.** *Let X be a random variable, $a \in \mathbb{R}$ be a constant, and c be the constant random variable. Then*

$$\boxed{\sigma_{aX+c} = |a|\sigma_X} \tag{2.6}$$

Proof. The corollary follows immediately from taking the square root of the result in Theorem 2.4. The theorem states that multiplying by a constant factor a scales the standard deviation appropriately, but adding a constant factor does not change the standard deviation. Intuitively, adding a constant factor shifts the distribution to the left or right, but does not affect its shape (and therefore its spread). □

We saw that linearity of expectation was an extremely powerful tool for computing expectation values. We would like to have a similar property hold for variance, but additivity of variance does not hold *in general*. However, we have the following useful theorem:

Theorem 2.6 (Variance of Sums of Random Variables)**.** *Let X and Y be random variables. Then*

$$\boxed{\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2(E(XY) - E(X)E(Y))} \tag{2.7}$$

Proof. As always, we start with the computational formula for variance.

$$\begin{aligned}\mathrm{Var}(X+Y) &= E((X+Y)^2) - E(X+Y)^2 \\ &= E(X^2+2XY+Y^2) - (E(X)+E(Y))^2 \\ &= E(X^2) + 2E(XY) + E(Y)^2 - E(X)^2 - 2E(X)E(Y) - E(Y)^2 \\ &= (E(X^2) - E(X)^2) + (E(Y)^2 - E(Y)^2) + 2(E(XY) - 2E(X)E(Y)) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2(E(XY) - E(X)E(Y))\end{aligned}$$

□

We will reveal the importance of the term $E(XY) - E(X)E(Y)$ later in the course. For now, we are more interested in the corollary:

Corollary 2.7 (Variance of Independent Random Variables)**.** *Let X and Y be independent. Then*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \tag{2.8}$$

Proof. When X and Y are independent, $E(XY) - E(X)E(Y) = 0$ according to our previous result. □

The general case is proven by induction. If $X_1, \ldots, X_n$ are pairwise independent random variables, then

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) \tag{2.9}$$

2.2 Probability Distributions Revisited

We will revisit the probability distributions introduced last time and proceed to calculate their variances.

2.2.1 Uniform Distribution

Recall that if $X \sim \mathrm{Unif}\{1, \ldots, n\}$,

$$E(X) = \frac{n+1}{2}$$

We compute:

$$\begin{aligned}E(X^2) &= \sum_{k=1}^{n} k^2 \cdot \frac{1}{n} \\ &= \frac{1}{n}\sum_{k=1}^{n} k^2 \\ &= \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} \\ &= \frac{(n+1)(2n+1)}{6}\end{aligned}$$

where we have used the identity (verified using induction)

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} \tag{2.10}$$

The variance is calculated to be (after a little algebra):

$$\begin{aligned}\mathrm{Var}(X) &= E(X^2) - E(X)^2 \\ &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\ &= \frac{n^2-1}{12}\end{aligned}$$

2.2.2 Bernoulli Distribution & Indicator Random Variables

Recall that if $X \sim \text{Ber}(p)$,

$$E(X) = p$$

Additionally, recall that Bernoulli random variables satisfy the important property $X^2 = X$. Hence,

$$E(X^2) = E(X) = p$$

The variance of a Bernoulli random variable is

$$\text{Var}(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p)$$

In the special case of indicator random variables, $p = \Pr(A)$ and

$$\boxed{\text{Var}(1_A) = \Pr(A)(1 - \Pr(A))} \tag{2.11}$$

2.2.3 Binomial Distribution

Recall that X is the sum of i.i.d. indicator random variables:

$$X = X_1 + \cdots + X_n$$

Hence,

$$\begin{aligned}\text{Var}(X) &= \text{Var}(X_1 + \cdots + X_n) \\ &= \text{Var}(X_1) + \cdots + \text{Var}(X_n)\end{aligned}$$

where we have used the independence of the indicator random variables to apply Corollary 2. Since each indicator random variable follows the $\text{Ber}(p)$ distribution, it follows that

$$\text{Var}(X) = n \cdot \text{Var}(X_i) = np(1-p)$$

2.2.4 Computing the Variance of Dependent Indicators

Unlike when we computed the mean of the binomial distribution (which did not require any assumptions except that the binomial distribution could be written as the sum of indicators), the calculation of the variance of the binomial distribution relied on a crucial fact: the indicator variables were *independent*. In this section, we outline a general method for computing the variance of random variables which can be written as the sum of indicators, even when the indicators are not mutually independent.

Let X be written as the sum of identically distributed indicators, which are *not* assumed to be independent:

$$X = 1_{A_1} + \cdots + 1_{A_n}$$

We first note that the expectation is easy, thanks to linearity of expectation (which holds regardless of whether the indicator random variables are independent or not):

$$E(X) = \sum_{i=1}^{n} E(1_{A_i}) = \sum_{i=1}^{n} \Pr(A_i) \tag{2.12}$$

Using the fact that the indicator variables are identically distributed:

$$\boxed{E(X) = n \Pr(A_i)} \tag{2.13}$$

Next, we compute $E(X^2)$:

$$E(X^2) = E((1_{A_1} + \cdots + 1_{A_n})^2)$$

Observe that the square $(1_{A_1} + \cdots + 1_{A_n})^2$ has two types of terms:

1. There are *like-terms*, such as $1_{A_1}^2$ and $1_{A_3}^2$. However, we know from the properties of indicators that $1_{A_i}^2 = 1_{A_i}$. There are n of these terms in total:

$$\sum_{i=1}^{n} 1_{A_i}^2 = \sum_{i=1}^{n} 1_{A_i} = 1_{A_1} + \cdots + 1_{A_n} = X \tag{2.14}$$

2. Then, there are *cross-terms*, such as $1_{A_2}1_{A_4}$ and $1_{A_1}1_{A_2}$. There are n^2 total terms in the square, and n of those terms are like-terms, which leaves $n^2 - n = n(n-1)$ cross-terms. We usually write the sum:

$$\sum_{i \neq j} 1_{A_i}1_{A_j} = 1_{A_1}1_{A_2} + \cdots + 1_{A_n}1_{A_{n-1}}$$

We can discover more about the cross-terms by examining their meaning. Consider the term $1_{A_i}1_{A_j}$: it is the product of two indicators. Each indicator 1_{A_i} is either 0 or 1; therefore, their product is also 0 or 1, which suggests that the product is also an indicator! The product is 1 if and only if each indicator is 1, which in the language of probability is expressed as

$$\Pr(1_{A_i}1_{A_j} = 1) = \Pr(1_{A_i} = 1, 1_{A_j} = 1) = \Pr(A_i \cap A_j) \tag{2.15}$$

We have arrived at a crucial fact: *the product of two indicators 1_{A_i} and 1_{A_j} is itself an indicator for the event $A_i \cap A_j$*. Therefore, we can rewrite the sum:

$$\sum_{i \neq j} 1_{A_i}1_{A_j} = \sum_{i \neq j} 1_{A_i \cap A_j} \tag{2.16}$$

Putting it together, we have that

$$X^2 = \left(\sum_{i=1}^{n} 1_{A_i}\right)^2 = X + \sum_{i \neq j} 1_{A_i \cap A_j} \tag{2.17}$$

The expectation of the square is

$$E(X^2) = E(X) + \sum_{i \neq j} E(1_{A_i \cap A_j}) = n \Pr(A_i) + n(n-1) \Pr(A_i \cap A_j) \tag{2.18}$$

(For simplicity, we made the assumption that all of the intersection probabilities $\Pr(A_i \cap A_j)$ are the same.) Finally, the variance is $E(X^2) - E(X)^2$, or

$$\boxed{\mathrm{Var}(X) = n \Pr(A_i) + n(n-1) \Pr(A_i \cap A_j) - n^2 (\Pr(A_i))^2} \tag{2.19}$$

Although the resulting formula looks rather complicated, it is a remarkably powerful demonstration of the techniques we have developed so far. The path we have taken is an amusing one: when the indicators are not independent, additivity of variance fails to hold, so the tool we ended up relying on was... linearity of expectation and indicators, of course!

2.2.5 Geometric Distribution

Next, we compute the variance of the geometric distribution. (It will mostly be an exercise in manipulating complicated series, but we include it for completeness.)

Recall that if $X \sim \mathrm{Geom}(p)$,

$$E(X) = \frac{1}{p}$$

We must compute

$$E(X^2) = p \sum_{k=1}^{\infty} k^2 (1-p)^{k-1}$$

Recall the formula

$$\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2} \tag{2.20}$$

Shifting the index k by 1 yields the equivalent formula

$$\sum_{k=0}^{\infty} (k+1)x^k = \frac{1}{(1-x)^2}$$

Differentiating this equation with respect to x yields

$$\sum_{k=1}^{\infty} k(k+1)x^{k-1} = \frac{2}{(1-x)^3} \tag{2.21}$$

Subtracting Equation 2.20 from Equation 2.21 yields

$$\sum_{k=1}^{\infty} k^2 x^{k-1} = \frac{2}{(1-x)^3} - \frac{1}{(1-x)^2} = \frac{1+x}{(1-x)^3}$$

Setting $x = 1-p$ yields

$$\sum_{k=1}^{\infty} k^2 (1-p)^{k-1} = \frac{2-p}{p^3}$$

Finally, we obtain

$$E(X^2) = p \cdot \frac{2-p}{p^3} = \frac{2-p}{p^2}$$

The variance is computed to be

$$\mathrm{Var}(X) = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$$

2.2.6 Poisson Distribution

Once again, computing the variance of the Poisson distribution will be an exercise in manipulating complicated sums (with one clever trick). The result, however, is extremely useful.

Recall that if $X \sim \mathrm{Pois}(\lambda)$,

$$E(X) = \lambda$$

We will proceed to calculate $E(X(X-1))$ instead. (The reasons for doing so will hopefully become clear.)

$$\begin{aligned} E(X(X-1)) &= \sum_{k=2}^{\infty} k(k-1)\frac{\lambda^k e^{-\lambda}}{k!} \\ &= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} \\ &= \lambda^2 e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\ &= \lambda^2 e^{-\lambda} e^{\lambda} \\ &= \lambda^2 \end{aligned}$$

Hence, by linearity of expectation,

$$\mathrm{Var}(X) = E(X^2) - E(X)^2 = E(X(X-1)) + E(X) - E(X)^2 = \lambda^2 + \lambda - \lambda^2$$

We have what may be considered to be a surprising result:

$$\mathrm{Var}(X) = \lambda$$

2.3 Bounds on Tail Probabilities

Often, probability distributions can be difficult to compute exactly, so we will cover a few important bounds.

2.3.1 Markov's Inequality

The following inequality is quite useful because it applies generally to any probability distribution with very few assumptions: the only piece of information it requires is the expectation of $f(X)$.

Theorem 2.8 (Markov's Inequality)**.** *Let X be a random variable, f be an increasing function, and $a \in \mathbb{R}$ such that $f(a) > 0$. Then the following inequality holds:*

$$\boxed{\Pr(X \geq a) \leq \frac{E(f(X))}{f(a)}} \tag{2.22}$$

Proof. Let $1_{X \geq a}$ be the indicator that $X \geq a$. There are two cases:

1. $X < a$: Then $1_{X \geq a} = 0 \leq f(X)/f(a)$, since $f(X)/f(a)$ is always positive by hypothesis.
2. $X \geq a$: Since f is increasing, we have $f(X) \geq f(a)$. Then $1_{X \geq a} = 1 \leq f(X)/f(a)$.

In either case, we have proven

$$1_{X \geq a} \leq \frac{f(X)}{f(a)} \tag{2.23}$$

Taking the expectation of both sides yields the desired inequality. □

Corollary 2.9 (Weak Markov's Inequality)**.** *Let X be a nonnegative random variable and $a > 0$. Then*

$$\Pr(X \geq a) \leq \frac{E(X)}{a} \tag{2.24}$$

Proof. The proof is immediate because $f(x) = x$ is an increasing function. □

2.3.2 Chebyshev's Inequality

We will use Markov's Inequality to derive another inequality which uses the variance to bound the probability distribution. Chebyshev's Inequality is useful for deriving confidence intervals and estimating sample sizes.

Theorem 2.10 (Chebyshev's Inequality)**.** *Let X be a random variable and $a > 0$. Then*

$$\boxed{\Pr(|X - E(X)| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}} \tag{2.25}$$

Proof. Let $Y = |X - E(X)|$ and $f(y) = y^2$. Since f is an increasing function, apply Markov's Inequality:

$$\Pr(Y \geq a) \leq \frac{E(Y^2)}{a^2}$$

Note, however, that $E(Y^2) = E(|X - E(X)|^2) = E((X - E(X))^2) = \mathrm{Var}(X)$. This completes the proof. □

Corollary 2.11 (Another Look at Chebyshev's Inequality)**.** *Let X be a random variable with standard deviation σ and $k > 0$. Then*

$$\boxed{\Pr(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}} \tag{2.26}$$

Proof. Set $a = k\sigma$ in Chebyshev's Inequality. □

Notably, Chebyshev's Inequality justifies why we call the variance a measure of the spread of the distribution. The probability that X lies more than k standard deviations away from the mean is bounded by $1/k^2$, which is to say that a larger standard deviation means X is more likely to be found away from its mean, while a low standard deviation means X will remain fairly close to its mean.

2.4 Weak Law of Large Numbers

Now, we can justify why the expectation is called the *long-run average* of a sequence of values. Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables, which we can think of as successive measurements of a true variable X. The idea is that X is some quantity which we wish to measure, and X follows some probability distribution with unknown parameters: mean μ and variance σ^2. Each random variable X_i is a measurement of X, which is to say that each X_i *follows the same probability distribution as* X. In particular, this means that each X_i also has mean μ and variance σ^2. We are interested in the *average* of the samples we collect:

$$\bar{X} := \frac{X_1 + \cdots + X_n}{n} \tag{2.27}$$

What is the expectation of $\bar{X}$? Since we would like to measure the parameter μ, we are hoping that $E(\bar{X}) = \mu$. (In other words, we are hoping that the average of our successive measurements X_i will be close to the true parameter μ.) We can use linearity of expectation to quickly check that this holds:

$$E(\bar{X}) = \frac{1}{n}(E(X_1) + \cdots + E(X_n)) = \frac{1}{n} \cdot n\mu = \mu$$

We therefore call $\bar{X}$ an **unbiased estimator** of μ.

The next question to ask is: on average, we expect $\bar{X}$ to estimate μ. But for a given experiment, *how close to μ do we expect $\bar{X}$ to be? How long will it take for $\bar{X}$ to converge to its mean, μ?* These are questions that involve the variance of the distribution. First, let us compute

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n} \tag{2.28}$$

(Notice that the answer is *not* σ^2!) In calculating our answer, we have used the scaling property of the variance and the assumption of independence. The dependence of $\sigma_{\bar{X}}$, the standard deviation of $\bar{X}$, is

$$\sigma_{\bar{X}} \sim \frac{1}{\sqrt{n}} \tag{2.29}$$

a dependence that is well-worth remembering. In particular, this result states that *the more samples we collect, the smaller our standard deviation becomes*! This result is what allows the scientific method to work: without it, gathering more samples would not make us any more certain of our results. We next state and prove a famous result, which shows that $\bar{X}$ converges to its expected value μ after enough samples are drawn.

Theorem 2.12 (Weak Law of Large Numbers)**.** *For all $\epsilon > 0$, in the limit as $n \to \infty$,*

$$\boxed{\Pr(\left|\bar{X} - \mu\right| \geq \epsilon) \to 0} \tag{2.30}$$

Proof. We will use Chebyshev's Inequality:

$$\Pr(\left|\bar{X} - \mu\right| \geq \epsilon) \leq \frac{\mathrm{Var}(\bar{X})}{\epsilon^2}$$

Filling in what we know about $\mathrm{Var}(\bar{X})$, we have

$$\Pr(\left|\bar{X} - \mu\right| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

which tends to 0 as $n \to \infty$. □

Intuitively, the theorem is asking: what is the probability that $\bar{X}$ is ϵ-far away from μ? The answer is: as $n \to \infty$, the probability becomes 0. *Increasing the number of samples decreases the probability that the sample average will be far from the true average.*

Chapter 3

LLSE, MMSE, and Conditional Expectation

A fundamental question in statistics is: given a set of data points $\{(X_i, Y_i)\}$, how can we *estimate* the value of Y as a function of X? First, we will discuss the covariance and correlation of two random variables, and use these quantities to derive the best *linear* estimator of Y, known as the LLSE. Then, we will define the conditional expectation, and proceed to derive the best *general* estimator of Y, known as the MMSE. The notion of conditional expectation will turn out to be an immensely useful concept with further applications.

3.1 Covariance

We have already discussed the case of independent random variables, but many of the variables in real life are dependent upon each other, such as the height and weight of an individual. Now we will consider how to quantify the dependence of random variables, starting with the definition of covariance.

Definition 3.1 (Covariance)**.** The **covariance** of two random variables X and Y is defined as

$$\boxed{\text{Cov}(X,Y) := E((X - E(X))(Y - E(Y)))} \tag{3.1}$$

The covariance is the product of the deviations of the two variables from their respective means. Suppose that whenever X is larger than its mean, Y is also larger than its mean; then, the covariance will be positive, and we say the variables are *positively correlated*. On the other hand, if whenever X is larger than its mean, Y is smaller than its mean, then the covariance is negative and we say that the variables are *negatively correlated*. In other words: *positive correlation means that X and Y tend to fluctuate in the same direction, while negative correlation means that X and Y tend to fluctuate in opposite directions.*

Just as we had a computational formula for variance, we have a computational formula for covariance.

Theorem 3.2 (Computational Formula for Covariance)**.** *Let X and Y be random variables. Then*

$$\boxed{\text{Cov}(X,Y) = E(XY) - E(X)E(Y)} \tag{3.2}$$

Proof. We take the definition of covariance and expand it out:

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

□

Corollary 3.3 (Covariance of Independent Random Variables)**.** *Let X and Y be independent random variables. Then we have* $\mathrm{Cov}(X,Y) = 0$.

Proof. By the assumption of independence, we have $E(XY) = E(X)E(Y)$; the corollary follows immediately. □

Important: Note that the converse is not true, i.e. $\mathrm{Cov}(X,Y) = 0$ does *not* imply that X and Y are independent! (Try to find a counterexample.)

Corollary 3.4 (Covariance & Variance)**.** *Let X be a random variable. Then*

$$\boxed{\mathrm{Var}(X) = \mathrm{Cov}(X,X)} \tag{3.3}$$

Proof. The proof is straightforward.

$$\mathrm{Cov}(X,X) = E(X\cdot X) - E(X)E(X) = E(X^2) - E(X)^2 = \mathrm{Var}(X)$$

□

The corollary is not very useful for calculating $\mathrm{Var}(X)$, but it does help to elucidate the relationship between covariance and variance. In fact, variance can be seen as a special case of covariance.

Corollary 3.5 (Variance of Sums of Random Variables)**.** *Let X and Y be random variables. Then*

$$\boxed{\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)} \tag{3.4}$$

Proof. The majority of the work for this proof was completed in the previous set of notes, in which we found

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2(E(XY) - E(X)E(Y))$$

The proof is immediate once we recognize $E(XY) - E(X)E(Y)$ as $\mathrm{Cov}(X,Y)$. □

The utility of the last result is that it holds true for *any* random variables, even ones that are not independent.

3.1.1 Bilinearity of Covariance

Next, we show that the covariance is bilinear, i.e. linear in each of its arguments.

Theorem 3.6 (Bilinearity of Covariance)**.** *Suppose X_i, Y_i are random variables, and $a \in \mathbb{R}$ is a constant. Then the following properties hold:*

$$\mathrm{Cov}(X_1 + X_2, Y) = \mathrm{Cov}(X_1, Y) + \mathrm{Cov}(X_2, Y) \tag{3.5}$$

$$\mathrm{Cov}(X, Y_1 + Y_2) = \mathrm{Cov}(X, Y_1) + \mathrm{Cov}(X, Y_2) \tag{3.6}$$

$$\operatorname{Cov}(aX, Y) = a\operatorname{Cov}(X, Y) = \operatorname{Cov}(X, aY) \tag{3.7}$$

Proof. The proofs are mostly straightforward using linearity of expectation. To prove Equation 3.5,

$$\begin{aligned}\operatorname{Cov}(X_1 + X_2, Y) &= E((X_1 + X_2)Y) - E(X_1 + X_2)E(Y)\\ &= E(X_1Y) - E(X_1)E(Y) + E(X_2Y) - E(X_2)E(Y)\\ &= \operatorname{Cov}(X_1, Y) + \operatorname{Cov}(X_2, Y)\end{aligned}$$

Similarly, to prove Equation 3.6,

$$\begin{aligned}\operatorname{Cov}(X, Y_1 + Y_2) &= E(X(Y_1 + Y_2)) - E(X)E(Y_1 + Y_2)\\ &= E(XY_1) - E(X)E(Y_1) + E(XY_2) - E(X)E(Y_2)\\ &= \operatorname{Cov}(X, Y_1) + \operatorname{Cov}(X, Y_2)\end{aligned}$$

Finally, to prove Equation 3.7,

$$\operatorname{Cov}(aX, Y) = E(aXY) - E(aX)E(Y) = a(E(XY) - E(X)E(Y)) = a\operatorname{Cov}(X, Y)$$

$$\operatorname{Cov}(X, aY) = E(XaY) - E(X)E(aY) = a(E(XY) - E(X)E(Y)) = a\operatorname{Cov}(X, Y)$$ □

3.1.2 Standardized Variables

Sometimes, we would like to write random variables in a standard form for easier computations.

Definition 3.7 (Standard Form)**.** Let X be a non-constant random variable. Then

$$\boxed{X^* := \frac{X - E(X)}{\sigma_X}} \tag{3.8}$$

is called the **standard form** of X. In statistics books, X^* is also denoted the **z-score** of X.

Next, we explain why X^* is called the standard form.

Theorem 3.8 (Mean & Variance of Standardized Random Variables)**.** *Let X^* be a standardized random variable. Then the mean and variance of X^* are*

$$\boxed{E(X^*) = 0} \tag{3.9}$$

$$\boxed{\operatorname{Var}(X^*) = E((X^*)^2) = 1} \tag{3.10}$$

Proof. First, we prove that the mean is 0 using linearity of expectation.

$$E(X^*) = \frac{E(X) - E(X)}{\sigma_X} = 0$$

Next, since $E(X^*) = 0$, then $\operatorname{Var}(X^*) = E((X^*)^2) - E(X^*)^2 = E((X^*)^2)$. Using the properties of variance,

$$\operatorname{Var}(X^*) = \operatorname{Var}\left(\frac{X - E(X)}{\sigma_X}\right) = \frac{\operatorname{Var}(X)}{\sigma_X^2} = \frac{\operatorname{Var}(X)}{\operatorname{Var}(X)} = 1$$ □

Standardizing the random variable X is equivalent to shifting the distribution so that its mean is 0, and scaling the distribution so that its standard deviation is 1. The random variable X^* is rather convenient because it is dimensionless: for example, if X and Y represent measurements of the temperature of a system in degrees Fahrenheit and degrees Celsius respectively, then $X^* = Y^*$.

3.1.3 Correlation

We next take a slight detour in order to define the correlation of two random variables, which appears frequently in statistics. Although we will not use correlation extensively, the exposition to correlation presented here should allow you to interpret the meaning of correlation in journals.

Definition 3.9 (Correlation)**.** The **correlation** of two random variables X and Y is

$$\boxed{\text{Corr}(X,Y) := \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}} \tag{3.11}$$

The correlation is often denoted by ρ or r, and is sometimes referred to as Pearson's correlation coefficient.

The next result provides an interpretation of the correlation.

Theorem 3.10 (Covariance & Correlation)**.** *The correlation of two random variables X and Y is*

$$\boxed{\text{Corr}(X,Y) = \text{Cov}(X^*, Y^*) = E(X^*Y^*)} \tag{3.12}$$

Proof. We calculate the covariance of X^* and Y^* using the properties of covariance.

$$\text{Cov}(X^*, Y^*) = \text{Cov}\left(\frac{X - E(X)}{\sigma_X}, \frac{Y - E(Y)}{\sigma_Y}\right) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

(Remember that the covariance of a constant with a random variable is 0.) The second equality follows easily because $\text{Cov}(X^*, Y^*) = E(X^*Y^*) - E(X^*)E(Y^*)$ and $E(X^*) = E(Y^*) = 0$. □

The result states that we can view the correlation as a *standardized* version of the covariance. As a result, we can also prove a result about the possible values of the correlation:

Theorem 3.11 (Magnitude of the Correlation)**.** *If X, Y are non-constant random variables,*

$$\boxed{-1 \leq \text{Corr}(X,Y) \leq 1} \tag{3.13}$$

Proof. The expectation of the random variable $(X^* \mp Y^*)^2$ is non-negative; hence

$$0 \leq E((X^* \mp Y^*)^2) = E((X^*)^2) \mp 2E(X^*Y^*) + E((Y^*)^2) = 2 \mp 2E(X^*Y^*)$$

We have the inequality

$$\pm \text{Corr}(X,Y) = \pm E(X^*Y^*) \leq 1$$

and the result follows by considering the two cases. □

Corollary 3.12 (Correlations of ± 1)**.** *Let X and Y be non-constant random variables and suppose that $\text{Corr}(X,Y) = 1$ or $\text{Corr}(X,Y) = -1$. Then Y is a linear function of X.*

Proof. From the above proof, $\text{Corr}(X,Y) = \pm 1$ if and only if $0 = E((X^* \mp Y^*)^2)$, which can only happen if $\forall \omega \in \Omega\ (X^* \mp Y^*)(\omega) = 0$, i.e. $X^* \mp Y^* = 0$. This implies that $Y^* = \pm X^*$, which is true if and only if $Y = aX + b$ for constants $a, b \in \mathbb{R}$. (What are the constants a and b?) □

Now, we can see that correlation is a useful measure of the degree of *linear* dependence between two variables X and Y. If $\text{Corr}(X,Y)$ is -1 or 1, then X and Y are perfectly linearly correlated, i.e. a plot of Y versus X would be a straight line. The closer the correlation is to ± 1, the closer the data resembles a straight line

relationship. If X and Y are independent, then the correlation is 0 (but the converse is not true). As a final remark, the square of the correlation coefficient is called the **coefficient of determination** (usually denoted R^2). The coefficient of determination appears frequently next to best-fit lines on scatter plots as a measure of how well the best-fit line fits the data.

3.2 LLSE

We will immediately apply our development of the covariance to the problem of finding the best linear predictor of Y given X. We begin by presenting the main result, and then proceed to prove that the result satisfies the properties we desire. [1]

Definition 3.13 (Least Linear Squares Estimate)**.** Let X and Y be random variables. The **least linear squares estimate** (LLSE) of Y given X is defined as

$$\boxed{L(Y \mid X) := E(Y) + \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(X - E(X))} \tag{3.14}$$

Observe that the LLSE is a random variable: in fact, it is a function of X.

3.2.1 Projection Property

Theorem 3.14 (Projection Property of LLSE)**.** *The LLSE satisfies*

$$\boxed{E(Y - L(Y \mid X)) = 0} \tag{3.15}$$

$$\boxed{E((Y - L(Y \mid X))X) = 0} \tag{3.16}$$

Proof. The proofs are actually relatively straightforward using linearity. Proof of Equation 3.15:

$$\begin{aligned} E(Y - L(Y \mid X)) &= E\left(Y - E(Y) - \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(X - E(X))\right) \\ &= E(Y) - E(Y) - \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(E(X) - E(X)) \\ &= 0 \end{aligned}$$

Proof of (Equation 3.16):

$$\begin{aligned} E((Y - L(Y \mid X))X) &= E\left(X\left(Y - E(Y) - \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(X - E(X))\right)\right) \\ &= E(XY) - E(X)E(Y) - \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(E(X^2) - E(X)^2) \\ &= \text{Cov}(X,Y) - \frac{\text{Cov}(X,Y)}{\text{Var}(X)} \cdot \text{Var}(X) \\ &= \text{Cov}(X,Y) - \text{Cov}(X,Y) \\ &= 0 \end{aligned}$$

□

If you have not studied linear algebra, then the rest of the section can be safely skipped. Linear algebra is not necessary to understand the properties of the LLSE, although linear algebra certainly enriches

[1] The material for this section and the next section on the MMSE relies heavily on Professor Walrand's notes, although I have inserted my own interpretation of the material wherever appropriate.

the theory of linear regression. We discuss linear algebra concepts solely to motivate the Projection Property.

Given a probability space Ω, then the space of random variables over Ω is a vector space (that is, random variables satisfy the axioms of vector spaces). Indeed, we have already introduced how to add and scalar multiply random variables. Specifically, since random variables are functions, the vector space of random variables is the vector space of functions $X : \Omega \to \mathbb{R}$, which is also called the *free space of* $\mathbb{R}$ *over the set* Ω (denoted $\mathbb{R}\langle\Omega\rangle$). Let V be the vector space of random variables over Ω. Let δ_γ be the random variable

$$\delta_\gamma(\omega) = \begin{cases} 1, & \omega = \gamma \\ 0, & \omega \neq \gamma \end{cases}$$

Then $\{\delta_\omega : \omega \in \Omega\}$ is a basis for V. Hence, if Ω is a finite probability space, then we see that $\dim(V) = |\Omega|$.

The map $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ given by $\langle X, Y \rangle := E(XY)$ is an inner product (sometimes called a non-degenerate bilinear form), which makes V into an inner product space (of course, the axioms of an inner product space must also be verified). Therefore, we can say that two random variables X and Y are *orthogonal* if $E(XY) = 0$. According to the Projection Property, we see that the random variable $Y - L(Y \mid X)$ is orthogonal to both the constant random variable 1 and the random variable X. Hence, $Y - L(Y \mid X)$ is orthogonal to the plane spanned by the two, i.e. $\mathcal{L}(X) := \text{span}\{1, X\}$. Note that $\mathcal{L}$ is the subspace of V containing all linear functions of X, e.g. $aX + b$, and $L(Y \mid X) \in \mathcal{L}$. Geometrically, we have that *the projection of* Y *onto* $\mathcal{L}$ *is* $L(Y \mid X)$, which is to say that $L(Y \mid X)$ is, in a sense, a linear function of X which is *closest* to X. We will make this notion more precise in the next section.

3.2.2 Linear Regression

Next, we show that the Projection Property implies that $L(Y \mid X)$ is the best linear predictor of Y.

Theorem 3.15 (Least Squares Property)**.** *$L(Y \mid X)$ is the best linear predictor of Y given X, i.e. $L(Y \mid X)$ minimizes the mean squared error. In other words, let $\mathcal{L}(X) := \{aX + b : a, b \in \mathbb{R}\}$ be the set of linear functions of X. Then $L(Y \mid X)$ has the property that for any element $aX + b \in \mathcal{L}(X)$,*

$$E((Y - L(Y \mid X))^2) \leq E((Y - aX - b)^2) \tag{3.17}$$

Proof. According to the Projection Property, we can combine Equation 3.15 and Equation 3.16 to obtain

$$E((Y - L(Y \mid X))(cX + d)) = 0$$

for any $cX + d \in \mathcal{L}(V)$. Let $\hat{Y} := L(Y \mid X)$ to simplify the notation. We calculate:

$$\begin{aligned} E((Y - aX - b)^2) &= E(([Y - \hat{Y}] + [\hat{Y} - aX - b])^2) \\ &= E((Y - \hat{Y})^2) + 2E([Y - \hat{Y}][\hat{Y} - aX - b]) + E((\hat{Y} - aX - b)^2) \\ &= E((Y - \hat{Y})^2) + E((\hat{Y} - aX - b)^2) \\ &\geq E((Y - \hat{Y})^2) = E((Y - L(Y \mid X))^2) \end{aligned}$$

In the second line of the proof, the term $E([Y - L(Y \mid X)][L(Y \mid X) - aX - b])$ vanishes by the Projection Property because $L(Y \mid X) - aX - b \in \mathcal{L}(X)$. Essentially, the proof states that $Y - L(Y \mid X)$ has a lower mean squared error than any other linear function of X. □

The least-squares line is used everywhere as a visual summary of a trend. Now you have seen the theory behind this powerful tool! As one last remark, you may be wondering how the common phrase *linear regression* relates to the LLSE. The answer has more to do with linear Gaussian models, which are a more advanced topic in continuous probability, but in short: the slope of the LLSE depends on the correlation of X and Y, and the correlation is rarely ± 1. Hence, we observe a phenomenon called *regression to the mean*, which is illustrated by the following example: suppose that the performance of two twins is highly correlated

with correlation $\rho = 0.9$. The first twin scores 1.2 standard deviations above the mean on an exam. Due to the high correlation, we would expect the second twin to also score higher than average on the exam; however, since the correlation is not perfectly 1, *we do not expect the second twin to score equally as high.* Instead, we might expect the second twin to score, say, 1 standard deviation above the mean. To learn more about this subject, look up the properties of the bivariate Gaussian distribution.

3.3 Conditional Expectation

Next, we will search for an even more powerful predictor of Y given X. [2] We define the conditional expectation $E(X \mid Y = y)$ using the following formula:

$$\boxed{E(X \mid Y = y) = \sum_x x \Pr(X = x \mid Y = y)} \tag{3.18}$$

In other words, $E(X \mid Y = y)$ is the expectation *with respect to the probability distribution of X conditioned on $Y = y$.* It is important to stress that $E(X \mid Y = y)$ is just a real number, just like any other expectation.

Notice that for every possible value of Y, we can assign a real number $E(X \mid Y = y)$. Hence, let us define $E(X \mid Y)$ to be a *function* of Y in the following manner:

Definition 3.16 (Conditional Expectation)**.** Let X and Y be random variables. Then $E(X|Y)$ is also a random variable, called the **conditional expectation** of X given Y, which has the value $E(X \mid Y = y)$ with probability $\Pr(Y = y)$. Observe that $E(X \mid Y)$ is a function of Y, i.e. $E(X \mid Y) = f(Y)$.

This point cannot be stressed enough: $E(X \mid Y)$ is a *random variable*! Although the conditional expectation may seem mysterious at first, there is an easy rule for writing down $E(X \mid Y)$. Let us consider an example for concreteness: let $Y \sim \text{Unif}\{1, \ldots, 5\}$ be the number of dice rolls, and X be the sum of the dice rolls. Conditioned on $Y = 1$ (that is, we roll one die), then $X \mid Y = 1 \sim \text{Unif}\{1, \cdots, 6\}$ so that $E(X \mid Y = 1) = 7/2$. Similarly, conditioned on $Y = 2$, we roll two dice, and so $E(X \mid Y = 2) = 7$ (the expected sum of two dice is 7). In general, conditioned on $Y = y$ dice rolls, we have $E(X \mid Y = y) = 7y/2$. Now, the random variable $E(X \mid Y)$ assigns to every possible value of Y the real number $E(X \mid Y = y)$; hence, we can write $E(X \mid Y) = 7Y/2$ (which is a function of Y, following the discussion above). At first, it may appear that in going from the expression for $E(X \mid Y = y)$ to $E(X \mid Y)$, we merely replaced the y with Y. Of course, there is more going on than just a simple substitution ($E(X \mid Y)$ and $E(X \mid Y = y)$ are completely different objects), but in one sense, *substituting $y \mapsto Y$ is exactly the procedure for writing down $E(X \mid Y)$.* Don't worry if conditional expectation is difficult to grasp at first. Mastery of this concept requires practice, and we will soon see how to apply conditional expectation to the problem of prediction.

3.3.1 The Law of Total Expectation

First, we prove an amazing fact about conditional expectation. We noted that $E(X|Y)$ is a random variable, and of course, we are always interested in the expectation values of random variables. Hence, we can ask the question: what is the expectation of $E(X \mid Y)$? To answer this question, recall that $E(X \mid Y)$ is a function of Y, that is, $E(X \mid Y) = f(Y)$. Then, to calculate the expectation of $E(X \mid Y)$, we see that *we should compute the expectation with respect to the probability distribution of Y.* Once you understand this point, the following proof is straightforward in its details.

Theorem 3.17 (Law of Total Expectation)**.** *Let X and Y be random variables. Then*

$$\boxed{E(E(X \mid Y)) = E(X)} \tag{3.19}$$

[2] The material presented in this section appears slightly differently from the presentation in Professor Walrand's lecture notes, but the lecture notes are still the source for these discussion notes.

Proof. As noted above, we compute $E(E(X \mid Y))$ with respect to the probability distribution of Y.

$$\begin{aligned}
E(E(X \mid Y)) &= \sum_y E(X \mid Y = y) \Pr(Y = y) \\
&= \sum_y \left(\sum_x x \Pr(X = x \mid Y = y) \right) \Pr(Y = y) \\
&= \sum_y \sum_x x \Pr(X = x, Y = y) \\
&= \sum_x x \left(\sum_y \Pr(X = x, Y = y) \right) \\
&= \sum_x x \Pr(X = x) \\
&= E(X)
\end{aligned}$$

□

What a marvelous proof! Once you understand this proof, you will have understood most of the concepts we have covered so far. In the first line, we use the definition of the expectation of a function of Y, i.e. $E(f(Y)) = \sum_y f(y) \Pr(Y = y)$; in the second line, we use the definition of $E(X \mid Y = y)$, which we defined in the previous section; in the third line, we use our knowledge of conditional probability; in the fifth line, we recall that summing over all possible values of Y in the joint distribution $\Pr(X = x, Y = y)$ yields the marginal distribution $\Pr(X = x)$; and finally, in the last line, we come back to the definition of $E(X)$.

Going back to the example in the previous section, we saw that $E(X \mid Y) = 7Y/2$. Then, we have that $E(E(X \mid Y)) = E(7Y/2) = 7E(Y)/2 = 21/2$ since $E(Y) = 3$ (recall that $Y \sim \text{Unif}\{1, \ldots, 5\}$). By the previous theorem, we have also found that $E(X) = 21/2$, with no additional work! The theorem provides a powerful way of computing the unconditional expectation of random variables which would normally be quite difficult to compute. For instance, consider trying to compute $E(X)$ without conditional expectation. X is the sum of the dice rolls, and the number of dice rolls is also a random variable (which means the method of indicators cannot be used, since we do not know how many indicators we should have). Calculating $E(X)$ is actually surprisingly difficult!

3.4 MMSE

3.4.1 Orthogonality Property

Before we apply the conditional expectation to the problem of prediction, we first note a few important properties of conditional expectation. The first is that the conditional expectation is linear, which is a crucial property that carries over from ordinary expectations. For example, $E(X + Y \mid Z) = E(X \mid Z) + E(Y \mid Z)$. The justification for this, briefly, is that $E(X + Y \mid Z = z) = E(X \mid Z = z) + E(Y \mid Z = z)$.

The second important property is: let $f(X)$ be any function of X and $g(Y)$ any function of Y. Then the conditional expectation $E(f(X)g(Y) \mid Y) = g(Y)E(f(X) \mid Y)$. The intuitive idea behind this property is that when we condition on Y, then any function of Y is treated as a constant and can be moved outside of the expectation by linearity. In other words, $E(f(X)g(Y) \mid Y = y) = E(f(X)g(y) \mid Y = y) = g(y)E(f(X) \mid Y = y)$. Then, to obtain $E(f(X)g(Y) \mid Y)$, we perform our usual procedure of substituting $y \mapsto Y$.

Let us now prove the analogue of the Projection Property for $E(Y \mid X)$.

Theorem 3.18 (Orthogonality Property)**.** *Let X and Y be random variables, and let $\phi(X)$ be any*

function of X. Then we have that

$$\boxed{E((Y - E(Y \mid X))\phi(X)) = 0} \tag{3.20}$$

Proof. We first calculate $E((Y - E(Y \mid X))\phi(X) \mid X)$ using the properties of conditional expectation.

$$\begin{aligned} E((Y - E(Y \mid X))\phi(X) \mid X) &= \phi(X)E(Y - E(Y \mid X) \mid X) \\ &= \phi(X)[E(Y \mid X) - E(E(Y \mid X) \mid X)] \\ &= \phi(X)[E(Y \mid X) - E(Y \mid X)] = 0 \end{aligned}$$

Note: Why is $E(E(Y \mid X) \mid X) = E(Y \mid X)$? We have that $E(Y \mid X)$ is a function of X, and conditioned on the value of X, $E(Y \mid X)$ is essentially a constant. Now observe that the law of total expectation gives us

$$E((Y - E(Y \mid X))\phi(X)) = E(E((Y - E(Y \mid X))\phi(X) \mid X)) = E(0) = 0$$ □

In our proof, we used the useful trick of conditioning on a variable first in order to use the law of total expectation. Compare with the Projection Property: the Orthogonality Property is stronger in that $\phi(X)$ is allowed to be any function of X, whereas the Projection Property was proven for linear functions of X.

3.4.2 Minimizing Mean Squared Error

Definition 3.19 (Minimum Mean Square Error)**.** The **minimum mean square error** (MMSE) estimator of Y given X is the function $f(X)$ which minimizes the mean squared error, i.e. for any function $g(X)$,

$$E((Y - f(X))^2) \leq E((Y - g(X))^2)$$

Compared to the task of finding the best *linear* estimator of Y given X, finding the best *general* estimator of Y given X seems to be an even more difficult task. However, the MMSE will simply turn out to be our new friend, the conditional expectation. In fact, the proof is virtually the same as the proof for the LLSE.

Theorem 3.20 (Conditional Expectation Is the MMSE)**.** *Let X and Y be random variables. Then the MMSE of Y given X is the function $E(Y \mid X)$, which is to say that for any function $g(X)$,*

$$\boxed{E((Y - E(Y \mid X))^2) \leq E((Y - g(X))^2)} \tag{3.21}$$

Proof. We have, using the Orthogonality Property,

$$\begin{aligned} E((Y - g(X))^2) &= E(([Y - E(X \mid Y)] + [E(X \mid Y) - g(X)])^2) \\ &= E((Y - E(X \mid Y))^2) + 2E([Y - E(X \mid Y)][E(X \mid Y) - g(X)]) + E((E(X \mid Y) - g(X))^2) \\ &= E((Y - E(X \mid Y))^2) + E((E(X \mid Y) - g(X))^2) \\ &\geq E((Y - E(X \mid Y))^2) \end{aligned}$$

The term $E([Y - E(X|Y)][E(X|Y) - g(X)])$ vanishes by the Orthogonality Property since $E(X|Y) - g(X)$ is just another function of X. □

We have come a long way, and the answer is surprisingly intuitive. We have just found the *best estimator* of Y given X (in the mean squared error sense) is the expected value of Y given X!

3.5 Bonus: Conditional Variance

In the bonus section, we will introduce the concept of conditional variance and prove a similar result to the law of total expectation. This material is not required for the course, but it is fun to see how far

probability theory can go. Just as the law of total expectation allowed us to compute the expectation by first conditioning on a random variable, the result of this section will allow you to calculate the variance by a similar conditioning procedure.

Definition 3.21 (Conditional Variance)**.** Let X and Y be random variables. We define $\text{Var}(X \mid Y = y)$ to mean the variance of the conditional probability distribution $\Pr(X = x \mid Y = y)$. Furthermore, the **conditional variance** $\text{Var}(X \mid Y)$ is defined to be the random variable that takes on the value $\text{Var}(X \mid Y = y)$ with probability $\Pr(Y = y)$. Note that $\text{Var}(X \mid Y)$ is a function of Y.

Now, we prove our main result:

Theorem 3.22 (Law of Total Variance)**.** *Let X and Y be random variables. Then*

$$\boxed{\text{Var}(X) = E(\text{Var}(X \mid Y)) + \text{Var}(E(X \mid Y))} \tag{3.22}$$

Proof. First, we use the computational formula for the variance.

$$\text{Var}(X) = E(X^2) - E(X)^2$$

We calculate each term by the law of total expectation.

$$\begin{aligned}
\text{Var}(X) &= E(E(X^2 \mid Y)) - E(E(X \mid Y))^2 \\
&= E(E(X^2 \mid Y)) - E(E(X \mid Y)^2) + E(E(X \mid Y)^2) - E(E(X \mid Y))^2 \\
&= E(E(X^2 \mid Y) - E(X \mid Y)^2) + \text{Var}(E(X \mid Y)) \\
&= E(\text{Var}(X \mid Y)) + \text{Var}(E(X \mid Y))
\end{aligned}$$

□

Although the formula is somewhat more complicated than the law of total expectation, it still holds a certain charm. For fun, try going back to the example in the conditional expectation section and calculating $\text{Var}(X)$.

Chapter 4

Continuous Probability

Our study of discrete random variables has allowed us to model coin flips and dice rolls, but often we would like to study random variables that take on a *continuous* range of values (i.e. an uncountable number of values). At first, understanding continuous random variables will require a conceptual leap, but most of the results from discrete probability carry over into their continuous analogues, with sums replaced by integrals.

4.1 Continuous Probability: A New Intuition

Let us pick a random number in the interval $[0, 1] \subseteq \mathbb{R}$. What is the probability that we picked *exactly* the number $2/3$? There are uncountably many real numbers in the interval $[0, 1]$, so it seems overwhelmingly unlikely that we pick any particular number. We must have that the probability of choosing exactly $2/3$ is 0 (i.e. it is impossible). In fact, for any real number $a \in [0, 1]$, the probability of choosing a must also be 0. But we are guaranteed to choose *some* number in $[0, 1]$, so how is that possible that whatever number we chose was chosen with probability 0? Furthermore, if I consider the probability of choosing a number less than $1/2$, then intuitively, we would like to say the probability is $1/2$. Does this not imply that when we add up a bunch of zero probabilities, we manage to get a non-zero probability? Clearly, our theory of discrete probability breaks down when we consider continuous random variables.

Therefore, let us begin with a few definitions. It is natural if you find that you cannot interpret these definitions immediately, since our intuition from discrete probability will require some updates. Over the course of working with continuous probability, you will start to build a new intuition.

The **density function** of a continuous random variable X (also known as the probability density function, or p.d.f.), is a real-valued function $f_X(x)$ such that

1. f_X is non-negative, i.e. $\forall x \in \mathbb{R}\ f_X(x) \geq 0$.
2. f_X is **normalized**, which is to say that f_X satisfies

$$\boxed{\int_{\mathbb{R}} f_X(x)\,\mathrm{d}x = 1} \tag{4.1}$$

Compare the normalization condition to the normalization condition in discrete probability, which says

$$\sum_x \Pr(X = x) = 1$$

We can interpret the continuous normalization condition to mean that "the probability that $X \in \mathbb{R}$ is 1". Similarly, we can define the probability that X lies in some interval $[a, b]$ as

$$\boxed{\Pr(X \in [a, b]) := \int_a^b f_X(x)\,\mathrm{d}x} \tag{4.2}$$

Remark: It does not matter whether I write the interval as $[a, b]$ (including the endpoints) or (a, b) (excluding the endpoints). The endpoints themselves do not contribute to the probability, since the probability of a single point is 0, as discussed above.

Therefore, the probability of an interval in $\mathbb{R}$ is interpreted as the *area under the density function above the interval*. (Fun fact: Thinking about continuous probability distributions is how I first discovered the Fundamental Theorem of Calculus!) Similarly, we can calculate the probability of the union of disjoint intervals by adding together the probabilities of each interval.

When we discuss continuous probability, it is also extremely useful to use the **cumulative distribution function** of X, or the c.d.f., defined as

$$\boxed{F_X(x) := \Pr(X \leq x) = \int_{-\infty}^{x} f(x')\,\mathrm{d}x'} \tag{4.3}$$

Remark: Once again, it makes no difference whether I write $\Pr(X \leq x)$ or $\Pr(X < x)$, since we have that $\Pr(X = x) = 0$. From now on, I will be sloppy and use the two interchangeably.

To obtain the p.d.f. from the c.d.f., we use the Fundamental Theorem of Calculus once again:

$$\boxed{f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x)} \tag{4.4}$$

We have given an intepretation of the area under the density function $f_X(x)$ as a probability. The natural question is: what is the interpretation of $f_X(x)$ itself? In the next two sections, we present an interpretation of $f_X(x)$ and introduce two ways of approaching continuous probability.

4.1.1 Differentiate the C.D.F.

To motivate the discussion, we will consider the following example: suppose you pick a random point within a circle of radius 1. Let R be the random variable which denotes the radius of the chosen point (i.e. the distance away from the center of the circle). What is $f_R(r)$?

The first method is to simply work with the c.d.f. and to obtain $f_R(r)$ by simply differentiating $F_R(r)$. Since $F_R(r) := \Pr(R < r)$, and we have chosen a point uniformly randomly inside of the circle, then the probability we are looking for is the ratio of area of the inner circle (which has radius r) to the area of the total circle (which has radius 1).

$$\Pr(R < r) = \frac{\text{Area Inside a Circle of Radius } r}{\text{Total Area Inside Circle}} = \frac{\pi r^2}{\pi} = r^2$$

Differentiating quickly yields $f_R(r) = 2r$ for $0 < r < 1$. Often, differentiating the c.d.f. is a simple way of finding the density function, and you may feel more comfortable with the process.

4.1.2 The Differential Method

The second method works with the density function directly, and therefore involves manipulation of differential elements (such as $\mathrm{d}r$). If you have taken many physics courses before, then you may already be familiar with the method. A word of advice: if you do not feel comfortable with the next section, you can always use the previous method of differentiating the c.d.f. instead.

To briefly motivate the procedure, let us consider $F_R(r + \mathrm{d}r)$. Using a Taylor expansion,

$$F_R(r + \mathrm{d}r) = F_R(r) + \left(\frac{\mathrm{d}}{\mathrm{d}r} F_R(r)\right) \mathrm{d}r + O((\mathrm{d}r)^2)$$

where the notation $O((\mathrm{d}r)^2)$ includes terms of order $(\mathrm{d}r)^2$ or higher. Recalling that $F_R(r) = \Pr(R < r)$ and the derivative of the c.d.f. is the density function,

$$\Pr(R < r + \mathrm{d}r) - \Pr(R < r) = f_R(r)\,\mathrm{d}r + O((\mathrm{d}r)^2)$$

Let us immediately apply the formula we have derived to the motivating example. From the c.d.f., we have that the probability of picking a point with $R < r + \mathrm{d}r$ is $(r + \mathrm{d}r)^2$, and the probability of picking a point with $R < r$ is r^2. Therefore,

$$\begin{aligned}\Pr(R < r + \mathrm{d}r) - \Pr(R < r) &= (r + \mathrm{d}r)^2 - r^2 \\ &= r^2 + 2r\,\mathrm{d}r + (\mathrm{d}r)^2 - r^2 \\ &= 2r\,\mathrm{d}r + (\mathrm{d}r)^2\end{aligned}$$

The expression above must equal $f_R(r)\,\mathrm{d}r + O((\mathrm{d}r)^2)$. Therefore, by looking at the term which is proportional to $\mathrm{d}r$, we can identify $f_R(r) = 2r$ and we obtain the same answer!

Initially, the differential method seems to require more calculations and is not as familiar as working with the c.d.f. instead. However, through this discussion we have obtained an *interpretation* for the density function: observe that $\Pr(X < x + \underset{\cdot}{\mathrm{x}}) - \Pr(X < x) = \Pr(X \in (x, x + \mathrm{d}x))$. Hence

$$\boxed{\Pr(X \in (x, x + \mathrm{d}x)) = f_X(x)\,\mathrm{d}x} \tag{4.5}$$

In words: the probability that the random variable X is found in the interval $(x, x + \mathrm{d}x)$ is proportional to both the length of the interval $\mathrm{d}x$ and the density function evaluated in the interval. The interpretation can be rephrased in the following way: the density function $f_X(x)$ is the *probability per unit length* near x.

The basic procedure for obtaining the density function directly is:

1. Use the information given in the problem to find $\Pr(X \in (x, x + \mathrm{d}x))$.
2. Drop terms of order $(\mathrm{d}x)^2$ or higher.
3. Identify the term multiplied by $\mathrm{d}x$ as the density function $f_X(x)$.

4.2 Continuous Analogues of Discrete Results

Now, we will return to familiar ground by re-introducing the results from discrete probability in the continuous case. In most cases, replacing summations with integration over $\mathbb{R}$ will work as intended.

The **expectation** of a continuous random variable X is

$$\boxed{E(X) := \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x} \tag{4.6}$$

Similarly, the expectation of a function of X is

$$\boxed{E(g(X)) := \int_{\mathbb{R}} g(x) f_X(x)\,\mathrm{d}x} \tag{4.7}$$

We can continue to use the formula $\mathrm{Var}(X) = E(X^2) - E(X)^2$ to obtain the variance of a continuous random variable. Observe that since integrals are linear, linearity of expectation still holds.

The **joint distribution** of two continuous random variables X and Y is $f_{X,Y}(x, y)$. The joint distribution represents everything there is to know about the two random variables. The joint distribution must satisfy the normalization condition

$$\boxed{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y = 1} \tag{4.8}$$

We say that X and Y are **independent** if and only if the joint density factorizes:

$$\boxed{f_{X,Y}(x,y) = f_X(x) f_Y(y)} \tag{4.9}$$

To obtain the **marginal distribution** of X from the joint distribution, integrate out the unnecessary variables (in this case, we integrate out Y):

$$\boxed{f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y} \tag{4.10}$$

The joint distribution can be extended easily to multiple random variables $X_1, \ldots, X_n$. The joint density satisfies the normalization condition

$$\boxed{\int_{\mathbb{R}^n} f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n = 1} \tag{4.11}$$

How do we use the joint distribution to compute quantities? (For simplicity, we will return to the joint density of X and Y.) Consider a region $J \subseteq \mathbb{R}^2$. The probability that $(X,Y) \in J$ is

$$\boxed{\Pr((X,Y) \in J) := \int_J f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y} \tag{4.12}$$

In other words, *to find the probability that (X,Y) is in a region J, we integrate the joint density over the region J.* As in multivariable calculus, it is often immensely helpful to draw the region of integration (i.e. draw J) before actually computing the integral.

To calculate the expectation of a function of many random variables,

$$\boxed{E(g(X_1,\ldots,X_n)) = \int_{\mathbb{R}^n} g(x_1,\ldots,x_n) f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n} \tag{4.13}$$

The **conditional density** of Y given $X = x$ is

$$\boxed{f_{Y \mid X=x}(y) := \frac{f_{X,Y}(x,y)}{f_X(x)}} \tag{4.14}$$

Note that the equation is superficially very similar to the definition of conditional probability for the discrete case. There are some subtleties, however. The function $f_{Y \mid X=x}(y)$ is a function of y only. However, to calculate the conditional density, we divide a function of x and y by a function of x, so in most cases, our function $f_{Y \mid X=x}(y)$ will actually involve the variable x as well! What is going on here? The answer is that x is a *parameter* of the function, not an *argument*. The distinction may seem blurry, but as an illustration, the normalization condition for the conditional density is

$$\boxed{\int_{\mathbb{R}} f_{Y \mid X=x}(y)\,\mathrm{d}y = 1} \tag{4.15}$$

In fact, we can prove this fact quite easily.

Proof.

$$\begin{aligned}\int_{\mathbb{R}} f_{Y \mid X=x}(y)\,\mathrm{d}y &= \int_{\mathbb{R}} \frac{f_{X,Y}(x,y)}{f_X(x)}\,\mathrm{d}y \\ &= \frac{1}{f_X(x)}\int_{\mathbb{R}} f_{X,Y}(x,y)\,\mathrm{d}y \\ &= \frac{1}{f_X(x)} f_X(x) = 1\end{aligned}$$

□

To make the concept even clearer, note that x is just some number, e.g. $x = 0$. Then we can write

$$f_{Y \mid X=0}(y) = \frac{f_{X,Y}(0,y)}{f_X(0)}$$

Now it should be very clear that $f_{Y \mid X=0}(y)$ is indeed a function of y only.

4.2.1 Tail Sum Formula

Yes, there is a continuous analogue of the tail sum formula.

Theorem 4.1 (Continuous Tail Sum Formula)**.** *Let X be a non-negative random variable. Then*

$$\boxed{E(X) = \int_0^\infty (1 - F_X(x))\,\mathrm{d}x} \tag{4.16}$$

Proof.

$$\begin{aligned}E(X) &= \int_0^\infty x f_X(x)\,\mathrm{d}x \\ &= \int_0^\infty \int_0^x f_X(x)\,\mathrm{d}t\,\mathrm{d}x \\ &= \int_0^\infty \int_t^\infty f_X(x)\,\mathrm{d}x\,\mathrm{d}t \\ &= \int_0^\infty \Pr(X > t)\,\mathrm{d}t \\ &= \int_0^\infty (1 - F_X(t))\,\mathrm{d}t\end{aligned}$$

The proof is quite similar to the discrete case. Interchanging the bounds of integration in line 3 is justified by Fubini's Theorem from multivariable calculus. □

4.3 Important Continuous Distributions

4.3.1 Uniform Distribution

The first distribution we will look at in detail is the $\mathrm{Unif}(0,1)$ distribution, which means that X is chosen uniformly randomly in the interval $(0,1)$. The property of being *uniform* means that the probability of choosing a number within an interval should only depend on the length of the interval. We can produce this by requiring the density function to equal a constant c in the interval $(0,1)$. Of course, since the density must be normalized to 1, then $c = 1$ and the density function is

$$f_X(x) = 1, \quad 0 < x < 1 \tag{4.17}$$

The c.d.f. is found by integrating:

$$F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 0 < x < 1 \\ 1, & x > 1 \end{cases} \tag{4.18}$$

Similarly, suppose that X and Y are i.i.d. Unif$(0,1)$ random variables. Since they are independent, the joint distribution is simply the product of their respective density functions:

$$f_{X,Y}(x,y) = f_X(x) f_Y(y) = 1, \quad 0 < x < 1, 0 < y < 1$$

The uniform distribution is especially simple because the density is constant. Suppose we want to find the probability that (X,Y) will lie in a region J. We integrate the joint density to find

$$\begin{aligned} \Pr((X,Y) \in J) &= \int_J f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_J \mathrm{d}x\,\mathrm{d}y \\ &= \text{Area}(J) \end{aligned}$$

The procedure for computing probabilities is therefore very simple. *To find the probability of an event involving two i.i.d.* Unif$(0,1)$ *random variables X and Y, draw the unit square, and shade in the region in the square which corresponds to the given event. The area of the shaded region is the desired probability.* As a result, many questions involving uniform distributions have very geometrical solutions.

To compute the expectation, we can simply observe that the distribution is symmetric about $x = 1/2$, so we can immediately write down $1/2$ as the expectation. To make the argument slightly more formal, consider the random variable $Y = 1 - X$. Observe that Y and X have identical distributions (but they are *not* independent). Therefore, $E(X) = E(Y) = E(1 - X)$, or $E(X) = 1 - E(X)$ using linearity. Hence,

$$E(X) = \frac{1}{2} \tag{4.19}$$

One question you may have is: why are X and Y identically distributed? To answer that question in a rigorous way, we will need the change of variables formula in order to show that X and Y have the same density function (and hence the same distribution). For now, accept the intuition! (Alternatively, carry out the integral $E(X) = \int_0^1 x\,\mathrm{d}x$, but that's boring.)

We compute variance in the standard way:

$$\begin{aligned} E(X^2) &= \int_0^1 x^2\,\mathrm{d}x \\ &= \frac{1}{3}x^3\Big|_0^1 \\ &= \frac{1}{3} \end{aligned}$$

Then, use $\text{Var}(X) = E(X^2) - E(X)^2 = 1/3 - 1/4$ to show that

$$\text{Var}(X) = \frac{1}{12} \tag{4.20}$$

4.3.2 Exponential Distribution

We will search for a continuous analogue of the discrete geometric distribution. Recall that the geometric distribution satisfied the memoryless property, i.e. $\Pr(X > s + t \mid X > s) = \Pr(X > t)$. In some sense, the memoryless property *defines* the geometric distribution, so we can look for a continuous distribution that satisfies the same property. Here is an amazing theorem:

Theorem 4.2 (The Memoryless Property Uniquely Defines the Exponential Distribution)**.** *Let T be a random variable satisfying the memoryless property. Then T has the density function*

$$f_T(t) = \lambda e^{-\lambda t}, \quad t > 0 \tag{4.21}$$

where $\lambda > 0$ is a parameter. We say that T follows the ***exponential distribution****, or $T \sim \text{Exp}(\lambda)$.*

Proof. Let $G_T(t) := 1 - F_T(t)$ be the **survival function** of T. Then

$$\begin{aligned} f_T(t)\,\mathrm{d}t &= \Pr(T \in (t, t+\mathrm{d}t)) \\ &= \Pr(T > t, T < t + \mathrm{d}t) \\ &= \Pr(T < t + \mathrm{d}t \mid T > t)\Pr(T > t) \\ &= (1 - \Pr(T > t + \mathrm{d}t \mid T > t))\Pr(T > t) \\ &= (1 - \Pr(T > \mathrm{d}t))\Pr(T > t) \\ &= (1 - G_T(\mathrm{d}t))G_T(t) \end{aligned}$$

Notice the use of the memoryless property in the fifth line. Consider the boundary condition of $G_T(t)$: we know that $G_T(0) = \Pr(T > 0) = 1$. Additionally, since $\mathrm{d}t$ is a small quantity, we can take the Taylor expansion of $G_T(\mathrm{d}t)$ about $t = 0$ and drop terms of order $(\mathrm{d}t)^2$ and higher. Therefore

$$\begin{aligned} f_T(t)\,\mathrm{d}t &\approx (1 - (G_T(0) + G_T'(0)\,\mathrm{d}t))\,G_T(t) \\ &= (1 - (1 + G_T'(0)\,\mathrm{d}t))G_T(t) \\ &= -G_T'(0)G_T(t)\,\mathrm{d}t \end{aligned}$$

Remembering that

$$-\frac{\mathrm{d}}{\mathrm{d}t}G_T(t) = -\frac{\mathrm{d}}{\mathrm{d}t}(1 - F_T(t)) = f_T(t)$$

We obtain a differential equation for $G_T(t)$:

$$\frac{\mathrm{d}}{\mathrm{d}t}G_T(t) = G_T'(0)G_T(t)$$

The solutions to the differential equation are all of the form $G_T(t) = Ce^{G_T'(0)t}$, where $C \in \mathbb{R}$ is a constant. Using the boundary condition $G_T(0) = 1$, we see that $C = 1$, so that

$$G_T(t) = e^{G_T'(0)t}$$

Set $\lambda = -G_T'(0)$. The c.d.f. is

$$F_T(t) = 1 - e^{-\lambda t}, \quad t > 0 \tag{4.22}$$

The density is easily obtained by differentiating. Note that the condition $\lambda > 0$ arises because if $\lambda < 0$, then the density function does not satisfy the condition of being non-negative. When $\lambda = 0$, the density is $f_T(t) = 0$ and cannot normalize to 1, so $\lambda = 0$ is also not a valid choice. □

In the lecture notes and slides, it was shown that the exponential distribution has the memoryless property. Here, we have shown that a distribution with the memoryless property is exponential. Hence, we have proven both directions: *a probability distribution is exponential if and only if it satisfies the memoryless property*! This is a remarkable result, and it is also true for the geometric distribution in the discrete case (although we have not proved it). Essentially, the memoryless property should be thought of as the defining characteristic of the exponential distribution.

We will proceed to compute the basic properties of the exponential distribution. First, check for yourself

that the exponential distribution is propertly normalized:

$$\int_0^\infty \lambda e^{-\lambda t}\,\mathrm{d}t = 1 \tag{4.23}$$

We claim that the normalization condition itself is already most of the work necessary to find the expectation of T. Normally, we would need to calculate

$$E(T) = \int_0^\infty t\lambda e^{-\lambda t}\,\mathrm{d}t$$

which can be solved using integration by parts. I hate integrating by parts, so here is a trick:

$$\begin{aligned} E(T) &= \lambda \int_0^\infty te^{-\lambda t}\,\mathrm{d}t \\ &= \lambda \int_0^\infty -\frac{\partial}{\partial\lambda} e^{-\lambda t}\,\mathrm{d}t \\ &= -\lambda \frac{\mathrm{d}}{\mathrm{d}\lambda} \int_0^\infty e^{-\lambda t}\,\mathrm{d}t \\ &= -\lambda \frac{\mathrm{d}}{\mathrm{d}\lambda} \frac{1}{\lambda} \\ &= -\lambda \cdot -\frac{1}{\lambda^2} \end{aligned}$$

Therefore,

$$E(X) = \frac{1}{\lambda} \tag{4.24}$$

Similarly, the variance is computed by finding

$$\begin{aligned} E(T^2) &= \lambda \int_0^\infty t^2 e^{-\lambda t}\,\mathrm{d}t \\ &= \lambda \int_0^\infty \frac{\partial^2}{\partial\lambda^2} e^{-\lambda t}\,\mathrm{d}t \\ &= \lambda \frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} \int_0^\infty e^{-\lambda t}\,\mathrm{d}t \\ &= \lambda \frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} \frac{1}{\lambda} \\ &= \lambda \cdot \frac{2}{\lambda^3} = \frac{2}{\lambda^2} \end{aligned}$$

Hence, the variance is $\mathrm{Var}(T) = 2/\lambda^2 - 1/\lambda^2$ or

$$\mathrm{Var}(T) = \frac{1}{\lambda^2} \tag{4.25}$$

The Minimum of Exponential Random Variables

Theorem 4.3. *Let $T_1, \ldots, T_n$ be independent exponential random variables with parameters $\lambda_1, \ldots, \lambda_n$ respectively. Then the minimum of the random variables is also exponentially distributed:*

$$\min\{T_1, \ldots, T_n\} \sim \mathrm{Exp}(\lambda_1 + \cdots + \lambda_n) \tag{4.26}$$

Proof. The easiest way to prove this is to once again consider the survival function $G_T(t) = Pr(T > t)$.

$$\begin{aligned}\Pr(\min\{T_1, \ldots, T_n\} > t) &= \Pr(T_1 > t, \ldots, T_n > t) \\ &= \Pr(T_1 > t) \cdots \Pr(T_n > t) \\ &= e^{-\lambda_1 t} \cdots e^{-\lambda_n t} \\ &= e^{-(\lambda_1 + \cdots + \lambda_n)t}\end{aligned}$$

We have the survival function of an exponential distribution with parameter $\lambda_1 + \cdots + \lambda_n$. □

4.4 Change of Variables

Often, we wish to find the density of a function of a random variable, such as the density of X^2 (assuming we already know $f_X(x)$). The problem can be solved in a satisfying and general way.

Theorem 4.4 (Change of Variables)**.** *Let X be a random variable and let $g : \mathbb{R} \to \mathbb{R}$ be a function which is one-to-one on the domain of X (i.e. g is one-to-one on the set of values for which $f_X(x) > 0$). Let H be the inverse of g (h exists since g is one-to-one). Then*

$$\boxed{f_Y(y) = f_X(h(y))|h'(y)|} \tag{4.27}$$

Proof. The method I prefer to use is to first manipulate the c.d.f. and then to differentiate. First, we consider the case that g is strictly increasing.

$$\begin{aligned}F_Y(y) &= \Pr(Y < y) \\ &= \Pr(g(X) < y) \\ &= \Pr(X < h(y)) \\ &= F_X(h(y))\end{aligned}$$

Differentiate both sides:

$$f_Y(y) = f_X(h(y))h'(y)$$

Since h is strictly increasing, then $h'(y) > 0$ and the change of variables equation holds. Now consider the case in which g is strictly decreasing. The main change appears on line 3 of the above steps: now that h is strictly increasing, applying h to both sides of the inequality also flips the direction of the inequality. Hence

$$\begin{aligned}F_Y(y) &= \Pr(X > h(y)) \\ &= 1 - \Pr(X < h(y)) \\ &= 1 - F_X(h(y))\end{aligned}$$

Differentiate both sides:

$$f_Y(y) = -f_X(h(y))h'(y)$$

Here, since h is strictly decreasing, then $h'(y) < 0$ and therefore $|h(y)| = -h(y)$. Hence, the change of variables formula still holds. □

My advice is to not bother remembering the change of variables formula. Instead, remember the basic outline of the proof: write down the c.d.f. of Y, then write in terms of the c.d.f. of X, and then differentiate.

Why did we assume that the function g was one-to-one? Actually, we assumed that g was one-to-one out of convenience: the condition that g is one-to-one is not necessary for change of variables to work, although

the change of variables formula is somewhat more complicated:

$$\boxed{f_Y(y) = \sum_{x:g(x)=y} f_X(x)|h'(y)|} \tag{4.28}$$

Essentially, the idea is that since g is no longer one-to-one, there may be many values of x such that $g(x) = y$, so we must sum up over all x such that $g(x) = y$. How far can we go? Can we define change of variables in the discrete case? Actually, the discrete case is rather easy.

$$\boxed{\Pr(Y = y) = \sum_{x:g(x)=y} \Pr(X = x)} \tag{4.29}$$

If you think about the above equation, you will realize there was no need to even write down the equation. We have been using the formula all along without knowing it.

We will use change of variables in the next section about the normal distribution.

4.5 Normal Distribution

We turn our attention to one of the most important probability distributions: the standard normal distribution (also called a Gaussian), which we denote $\mathscr{N}(0, 1)$. (The first parameter is the mean, and the second parameter is the variance). The density will be proportional to $e^{-x^2/2}$ (defined over all of $\mathbb{R}$), which has no known elementary antiderivative. Therefore, we devote the next section to integrating this function.

4.5.1 Integrating the Normal Distribution

Let us find the normalization constant, i.e. we must find $c \in \mathbb{R}$ such that

$$c \int_{\mathbb{R}} e^{-x^2/2} \, \mathrm{d}x = 1$$

In fact, we will solve a slightly more general integral by considering $e^{-\alpha x^2/2}$ instead. (We can always set $\alpha = 1$ at the end of our computations, after all.) The reason for doing so may not be clear, but it will actually save us some time later on.

As I mentioned above, $e^{-x^2/2}$ cannot be integrated normally, so we will need to use another trick (hopefully you find all of these tricks somewhat interesting). The trick here is to consider the square of the integral:

$$I^2 = \int_{\mathbb{R}} e^{-\alpha x^2/2} \, \mathrm{d}x \int_{\mathbb{R}} e^{-\alpha y^2/2} \, \mathrm{d}y = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\alpha(x^2+y^2)/2} \, \mathrm{d}x \, \mathrm{d}y$$

Notice that the integral depends only on the quantity $x^2 + y^2$, which you may recognize as the square of the distance from the origin. (The variables x and y were chosen suggestively to bring to mind the picture of integration on the plane $\mathbb{R}^2$.) Therefore, it is natural to change to polar coordinates with the substitutions

$$x^2 + y^2 = r^2 \tag{4.30}$$

$$\mathrm{d}x \, \mathrm{d}y = r \, \mathrm{d}r \, \mathrm{d}\theta \tag{4.31}$$

(Don't forget the extra factor of r that arises due to the Jacobian. For more information, consult a multivariable calculus textbook which develops the theory of integration under change of coordinates.)

In polar coordinates, the integral can now be evaluated.

$$\begin{aligned} I^2 &= \int_0^{2\pi} \int_0^{\infty} e^{-\alpha r^2/2} r \, \mathrm{d}r \, \mathrm{d}\theta \\ &= \int_0^{2\pi} \mathrm{d}\theta \int_0^{\infty} r e^{-\alpha r^2/2} \, \mathrm{d}r \\ &= 2\pi \cdot -\frac{1}{\alpha} e^{-\alpha r^2/2} \Big|_0^{\infty} \\ &= \frac{2\pi}{\alpha} \end{aligned}$$

Set $\alpha = 1$ and $I^2 = 2\pi$. We obtain the surprising result that $I = \sqrt{2\pi}$. The normalized density is

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{4.32}$$

As noted before, there is no elementary antiderivative of the density function, so we cannot write down the c.d.f. in terms of familiar functions. The c.d.f. of the standard normal distribution is often denoted

$$\boxed{\Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-(x')^2/2} \, \mathrm{d}x'} \tag{4.33}$$

4.5.2 Mean and Variance of the Normal Distribution

We sure hope that the mean and variance of the $\mathcal{N}(0,1)$ distribution are 0 and 1, as we claimed. Here, we verify that this is the case. First, notice that the density $f_X(x)$ depends only on x^2, so interchanging $x \mapsto -x$ leaves the density unchanged. The density is therefore symmetric about $x = 0$, and we write down

$$E(X) = 0$$

The variance is slightly trickier.

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2} \, \mathrm{d}x \end{aligned}$$

Although the integral smells like integration by parts, recall how we managed to avoid integration by parts in a similar integral when we computed the mean and variance of the exponential distribution. We would like to apply a similar trick here, so let us instead consider the integral of $e^{-\alpha x^2/2}$ (with the intention of setting $\alpha = 1$ at the end of our computations).

$$\begin{aligned} E(X^2) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\alpha x^2/2} \, \mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}} \int_{\infty}^{\infty} (-2) \frac{\partial}{\partial \alpha} e^{-\alpha x^2/2} \, \mathrm{d}x \\ &= \frac{-2}{\sqrt{2\pi}} \frac{\mathrm{d}}{\mathrm{d}\alpha} \int_{-\infty}^{\infty} e^{-\alpha x^2/2} \, \mathrm{d}x \\ &= \frac{-2}{\sqrt{2\pi}} \frac{\mathrm{d}}{\mathrm{d}\alpha} \sqrt{\frac{2\pi}{\alpha}} \\ &= -2 \cdot -\frac{1}{2} \alpha^{-3/2} \\ &= \alpha^{-3/2} \end{aligned}$$

Again, set $\alpha = 1$, and since $E(X)^2 = 0$,

$$\mathrm{Var}(X) = 1$$

Hopefully, it should be clear by now why we need the α: it is simply a tool that we use to integrate more easily, and then we discard it after we finish the actual integration. In any case, we have verified that the mean and variance are indeed 0 and 1 respectively.

We now apply the change of variables technique to the standard normal distribution, both as an illustration of the technique, and also to obtain the general form of the normal distribution. Consider the function $g(x) = \mu + \sigma x$ (with $\mu, \sigma \in \mathbb{R}$). Let $Y = g(X)$. We proceed to find the density of Y. First, note that the inverse function $h = g^{-1}$ is

$$h(y) = \frac{y - \mu}{\sigma}$$

and

$$|h'(y)| = \frac{1}{|\sigma|}$$

Then we have that

$$f_Y(y) = f_X(h(y))|h'(y)| = \frac{1}{|\sigma|} f_X((y - \mu)/\sigma)$$

Plugging in $(y - \mu)/\sigma$ into the standard normal density yields

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/(2\sigma^2)} \tag{4.34}$$

What are the mean and variance of Y? The answer is simple once we recall that $Y = \mu + \sigma X$. Using the basic properties of linearity and scaling,

$$E(Y) = \mu$$
$$\mathrm{Var}(Y) = \sigma^2$$

We call Y a normal random variable with mean μ and variance σ^2, which is signified $Y \sim \mathcal{N}(\mu, \sigma^2)$. We have seen that any normal distribution is found from the standard normal distribution by the following two-step procedure: 1) scale by σ, 2) shift by μ.

4.5.3 Sums of Independent Normal Random Variables

In this section, we prove the crucial fact that the normal distribution is *additive*, that is, the sums of independent normal random variables is also normally distributed. The theorem is breathtaking, but let us precede the theorem with some discussion about the joint density of two normal random variables.

Let X, Y be i.i.d. standard normal random variables. Since they are independent, their joint density is the product of the individual densities.

$$\begin{aligned} f_{X,Y}(x, y) &= f_X(x) f_Y(y) \\ &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \\ &= \frac{1}{2\pi} e^{-(x^2+y^2)/2} \end{aligned}$$

Observe that the joint density only depends on $x^2 + y^2 = r^2$, which is the square of the distance from the origin. (In fact, we already noticed this property when we were integrating the Gaussian.) In polar coordinates, the density only has a dependence on r, not on θ, which exhibits an important geometric property: the joint density is *rotationally symmetric*, that is, the Gaussian looks exactly the same if you rotate your coordinate axes. How can we utilize this geometric property to prove our result?

Let $Z = X + Y$. We can work with the c.d.f. of Z, that is, consider $F_Z(z) = \Pr(Z < z)$. We can write

$$F_Z(z) = \Pr(Z < z) = \Pr(X + Y < z)$$

To compute this probability, we integrate the joint density

$$F_Z(z) = \int_J f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y$$

where J is the region

$$J := \{(x, y) : x + y < z\} \subseteq \mathbb{R}^2$$

We know that $x + y = z$ is a line in $\mathbb{R}^2$. Perhaps we can align our coordinate axes with the line $x + y = z$, and the integral will be simplified. Let us consider a new coordinate system given by

$$x = \frac{1}{\sqrt{2}}x' - \frac{1}{\sqrt{2}}y'$$
$$y = \frac{1}{\sqrt{2}}x' + \frac{1}{\sqrt{2}}y'$$

Let us plug this into our region J to obtain

$$J = \{(x', y') : \sqrt{2}x' < z\} \subseteq \mathbb{R}^2$$

Since the joint density is invariant under rotations, changing variables from $x \mapsto x'$ and $y \mapsto y'$ should *not* affect the value of the integral. Hence,

$$\begin{aligned}
F_Z(z) &= \Pr(Z < z) \\
&= \int_J f_{X,Y}(x', y')\,\mathrm{d}x'\,\mathrm{d}y' \\
&= \int_{-\infty}^{z/\sqrt{2}} f_X(x')\,\mathrm{d}x' \int_{-\infty}^{\infty} f_Y(y')\,\mathrm{d}y' \\
&= \int_{-\infty}^{z/\sqrt{2}} f_X(x')\,\mathrm{d}x' \\
&= \Pr(X < z/\sqrt{2}) \\
&= \Pr(\sqrt{2}X < z)
\end{aligned}$$

Compare the top and bottom lines of the above derivation and we see that Z has the same distribution as $\sqrt{2}X$, where $X \sim \mathcal{N}(0, 1)$. From our previous section, we saw that scaling a standard normal random variable by $\sqrt{2}$ yields the $\mathcal{N}(0, 2)$ distribution (the normal distribution with mean 0 and variance 2). Hence,

$$Z \sim \mathcal{N}(0, 2)$$

However, $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(0, 1)$, and we have just found that $Z = X + Y \sim \mathcal{N}(0, 2)$. Could summing up independent normal random variables really be as easy as adding their parameters?

Lemma 4.5 (Rotational Invariance of the Gaussian)**.** *Let X, Y, be i.i.d. standard normal random variables and let $\alpha, \beta \in [0, 1]$ so that $\alpha^2 + \beta^2 = 1$. Then $\alpha X + \beta Y \sim \mathcal{N}(0, 1)$.*

Proof. We simply extend the previous argument to any arbitrary rotation. By the conditions given in the lemma, we can write down

$$\sin(\theta) = \alpha$$
$$\cos(\theta) = \beta$$

for some $\theta \in [0, \pi/2]$. We can write the c.d.f. of $Z = \alpha X + \beta Y$ as

$$\begin{aligned} F_Z(z) &= \Pr(Z < z) \\ &= \Pr(\alpha X + \beta Y < z) \\ &= \int_J f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y \end{aligned}$$

where

$$J = \{(x, y) : x\cos(\theta) + y\sin(\theta) < z\} \subseteq \mathbb{R}^2$$

Under the change of coordinates

$$\begin{aligned} x &= x'\cos(\theta) - y'\sin(\theta) \\ y &= x'\sin(\theta) + y'\cos(\theta) \end{aligned}$$

the area of integration becomes

$$J = \{(x', y') : x' < z\} \subseteq \mathbb{R}^2$$

Hence we conclude that

$$\begin{aligned} F_Z(z) &= \Pr(Z < z) \\ &= \int_J f_{X,Y}(x', y')\, \mathrm{d}x'\, \mathrm{d}y' \\ &= \int_{-\infty}^{z} f_X(x')\, \mathrm{d}x' \int_{-\infty}^{\infty} f_Y(y')\, \mathrm{d}y' \\ &= \int_{-\infty}^{z} f_X(x')\, \mathrm{d}x' \\ &= \Pr(X < z) \end{aligned}$$

Z has the same distribution as X, and $X \sim \mathcal{N}(0, 1)$, so we're done. □

Theorem 4.6 (Sums of Independent Normal Random Variables)**.** *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$. Then*

$$\boxed{X := X_1 + \cdots + X_n \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)} \tag{4.35}$$

Proof. We will prove the result for the sum of two independent normal random variables. The general result follows as a quick exercise in induction. Let $Z_i = (X_i - \mu_i)/\sigma_i$ be the standardized form of X_i. Then $Z_i \sim \mathcal{N}(0, 1)$. Additionally, observe that

$$Z = \sqrt{\frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}}\frac{X_1 - \mu_1}{\sigma_1} + \sqrt{\frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}}\frac{X_2 - \mu_2}{\sigma_2} = \frac{X_1 + X_2 - (\mu_1 + \mu_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

Apply Lemma (4.5) to Z to obtain that $Z \sim \mathcal{N}(0, 1)$. Finally, since $X = X_1 + X_2 = \mu_1 + \mu_2 + \sqrt{\sigma_1^2 + \sigma_2^2}Z$, it follows from the change of variables that

$$Z \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

□

The theorem is beautiful, so do not misuse it! The most common mistake that students make is that they forget the important rule: *variances add, standard deviations do not.*

4.5.4 Central Limit Theorem

We end with one of the most marvelous results in all of probability theory: the Central Limit Theorem. To start, why do we care so much about the normal distribution? It certainly has nice properties, but the exponential distribution is also easy to work with and models many real-life situations very well (e.g. particle decay). The normal distribution, however, goes even farther: I claim that it allows us to model *any* situation whatsoever. How can such a bold statement possibly be true?

First, let us make the statement precise. Suppose (X_i) is a sequence of i.i.d. random variables with mean μ and finite moments. Define

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$$

Then as $n \to \infty$, $\sqrt{n}(\bar{X}_n - \mu)$ *converges in distribution* to the normal distribution.

Remember that the Weak Law of Large Numbers tells us that as we increase the number of samples, the sample mean *converges in probability* to the expected value. The Central Limit Theorem is stronger; it states that the distribution of the sample mean also converges to a particular distribution, and it is our good friend the normal distribution! The power is that we can start from any distribution at all. Each X_i can be uniform or exponential, or a crazy distribution without a name, yet the sample mean will still converge to a single distribution. In fact, the version of the Central Limit Theorem is much weaker than it needs to be. There are versions of the Central Limit Theorem which drop the assumption of identical distributions; another version drops the assumption of independence. Of all of the theorems in mathematics, the Central Limit Theorem is definitely one of my favorites.

The Central Limit Theorem is the basis for much of modern statistics. When statisticians carry out hypothesis testing or construct confidence intervals, they do not care that they do not know the exact distribution of their data. They simply collect enough samples (30 is usually sufficient) until their sampling distributions are roughly normally distributed.

Of course, we have not answered many questions, such as, "what does it mean for a distribution to converge?" and "why is the Central Limit Theorem true?" The former question requires a rigorous study of real analysis, and requires far more probability theory than we have covered in this course. Hence, the latter question will only be answered partially in the next section.

4.6 Bonus: CLT Proof Sketch

Our approach for providing justification for the Central Limit Theorem will be to ignore the finer points of convergence and analysis in favor of relying on important, yet plausible, results. If the following treatment of the Central Limit Theorem is unsatisfactory to you, then I urge you to take additional probability and statistics courses (as well as real analysis, as the mathematical background for measure-theoretic probability).

4.6.1 Characteristic Functions

In 1.4, I introduced the moment-generating functions, which were a convenient way of quickly calculating the various moments of a probabiity distribution. One problem with the moment-generating function is that it does not always converge, which is to say that not every distribution has a moment-generating function. The characteristic function does not have such a problem.

Definition 4.7 (Characteristic Function)**.** The **characteristic function** of a probability distribution for a random variable X is defined to be

$$\boxed{\varphi_X(t) := E(e^{itX})} \tag{4.36}$$

Perhaps, you may recognize the characteristic function of a probability density as the *Fourier transform* of the probability distribution. The characteristic function has many nice properties, and the one that we will use in particular is that *if the characteristic functions of a sequence of random variables converges to a single characteristic function φ_X, then the sequence of random variables converges in distribution to X*. The result we are quoting is commonly known as **Lévy's Continuity Theorem**.

We can make the theorem plausible by making a few non-rigorous appeals to intuition: the characteristic function is another representation of the probabiity distribution, just like the c.d.f. or the survival function. From the characteristic function, it is possible to use the inverse Fourier transform to recover the density function. The characteristic function, in a sense, captures all of the information in a probability distribution, so convergence of the characteristic functions should be equivalent to convergence of the distributions themselves. Try not to worry about the exceptions to the statements I have made until learning about the Central Limit Theorem in a more formal context.

Let us attempt to compute the characteristic function of the standard normal distribution. Since

$$-\frac{x^2}{2} + itx = -\frac{1}{2}(x - it)^2 - \frac{t^2}{2}$$

We have that

$$\begin{aligned}
\phi_X(t) &= \int_{\mathbb{R}} e^{itx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-x^2/2 + itx} \, \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \int_{\mathbb{R}} e^{-(x-it)^2/2} \, \mathrm{d}x
\end{aligned}$$

Uh oh, we ran into a problem. You might expect the integral to come out to be $\sqrt{2\pi}$ since the integral looks like the integral over density of a normal distribution with mean *it* and variance 1. However, we cannot assume that integration over complex numbers behaves the same way that it does over real numbers. We will spare you further confusion of working out the complex integral and reveal that the integral does indeed come out to $\sqrt{2\pi}$. Therefore, the characteristic function is

$$\varphi_X(t) = e^{-t^2/2} \tag{4.37}$$

Neither the above paragraph, nor the plausibility arguments for the continuity theorem, are rigorous proofs. Then again, our goal at this stage should not be to look for a rigorous proof, but rather to see if we can discover pieces of the truth which hint at the reason why the Central Limit Theorem is true.

A quick result about characteristic functions:

Theorem 4.8 (Characteristic Function of Independent, Identically Distributed Random Variables)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables with X being their sum. Then*

$$\boxed{\varphi_X(t) = [\varphi_{X_i}(t)]^n} \tag{4.38}$$

Proof.

$$\begin{aligned}
\varphi_X(t) &= E(e^{itX}) \\
&= E(e^{it(X_1+\cdots+X_n)}) \\
&= E(e^{itX_1}\cdots e^{itX_n}) \\
&= E(e^{itX_1})\cdots E(e^{itX_n}) \\
&= \varphi_{X_1}(t)\cdots\varphi_{X_n}(t) \\
&= [\varphi_{X_i}(t)]^n
\end{aligned}$$

The assumptions of independence and identical distributions were used in lines 4 and 6 respectively. □

4.6.2 Proof Sketch Attempt

We will give a proof sketch of the Central Limit Theorem in which the X_i variables are i.i.d. with mean 0 and variance 1. Our plan of attack is to show that the characteristic functions converge to $e^{-t^2/2}$, which is the characteristic function of the $\mathcal{N}(0,1)$ distribution.

Theorem 4.9 (Central Limit Theorem)**.** *Let (X_n) be a sequence i.i.d. random variables with mean 0 and variance 1. Define*

$$Z_n := \frac{1}{1/\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

Then (Z_n) is a sequence of random variables which converges in distribution to the $\mathcal{N}(0,1)$ distribution.

Proof. Recall that if

$$\bar{X} = \frac{X_1+\cdots+X_n}{n}$$

where $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, then $E(\bar{X}) = \mu$ and $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X_n)/n$. Here, we have $\mu = 0$ and $\sigma = 1$, so $\bar{X}_n$ has mean 0 and variance $1/\sqrt{n}$. Hence, the variable Z_n can be thought of as the standardized version of $\bar{X}_n$ (standardization is a property we desire if we want to converge to the *standard* normal distribution). Observe that we can write

$$Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$$

Using the theorem we proved for characteristic functions,

$$\varphi_{Z_n}(t) = [\varphi_{X_i/\sqrt{n}}(t)]^n = [E(e^{itX_i/\sqrt{n}})]^n$$

Let us calculate the last quantity by expanding as a Taylor series:

$$\begin{aligned}
\varphi_{Z_n}(t) &= \left[E\left(1 + \frac{it}{\sqrt{n}}X_i - \frac{t^2}{2n}X_i^2 + \cdots\right)\right]^n \\
&= \left(1 + \frac{it}{\sqrt{n}}E(X_i) - \frac{t^2}{2n}E(X_i^2) + \cdots\right)^n \\
&\approx \left(1 - \frac{t^2}{2n}\right)^n
\end{aligned}$$

In the last line, we have carried out several steps at once. We used the fact that $E(X_i) = 0$ and that $E(X_i^2) = \mathrm{Var}(X_i) = 1$. Also, since we will be taking the limit as $n \to \infty$, we dropped terms which were

of order $1/n^2$ or higher. What were are left with is

$$\lim_{n\to\infty} \varphi_{Z_n}(t) \approx \lim_{n\to\infty} \left(1 - \frac{t^2}{2n}\right)^n$$

Recalling that

$$e^x = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n \tag{4.39}$$

Then we can see that

$$\lim_{n\to\infty} \varphi_{Z_n}(t) = e^{-t^2/2}$$

We have our desired result! □