Part II

Probability

In order to apply statistical procedures correctly you need to understand the probabilistic models used in them. Different stats procedures make different probability assumptions about how the data was generated and you need to learn some probability theory in order to understand these assumptions. Statistical procedures are like computer programs that work on different types of input files. It doesn't make sense to try to open a text file with a video editing program. Even if you somehow trick the program into opening the file, you can't expect a correct output will result since the video program only works correctly for video data. Similarly, it doesn't make sense to apply a stats procedure to experimental data that doesn't fit the statistical assumptions of that procedure.

It's not enough to follow a statistical procedure's steps like the steps of a recipe. Without understanding the assumptions you might pick the wrong stats procedure and come to invalid conclusions, like what happens when you open a text file with a video editing program. Unlike computer programs that will refuse to open files that they are not designed to process, statistical procedures won't "complain" if you use them the wrong way. This task is on you. You must learn probability theory so you can use stats correctly.

SEE `https://ermongroup.github.io/cs228-notes/preliminaries/probabilityreview/` and `https://github.com/ermongroup/cs228-notes/blob/master/preliminaries/probabilityreview/index.md`

Chapter 21

Introduction

Probability theory started when a bunch of mathematicians went to the casino and tried to use their math skills to compute the chance of winning at different games. Suppose someone offers you the chance to play a dice game. The game costs \$1 to play, and you win \$5 if you roll a ⚅ and you win nothing if you roll any other number. The die is six-sided are fair, meaning that you have an equal chance of rolling any of the six possible numbers {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}. Knowing the chance for each outcome is simply $\frac{1}{6}$, you calculate your expected gains like so:

$$\$5 \cdot \text{probability of rolling a ⚅} = \$5 \cdot \frac{1}{6} = 83 \text{ cents.}$$

Since the expected gains from this game are smaller than the cost of playing, this game is not worth playing. You'd lose 17 cents per game on average, so you might as well throw your money down the drain—it would be just as efficient. Surprise-surprise: the expected gains for playing casino games are always negative.

Though the early days of probability theory were concerned with games of chance, this framework can be applied to any situation where uncertainty plays a role including many real-world systems. Probability theory is the foundation for both statistics and machine learning, so our journey towards learning these subjects must start with an introduction of the laws of probability theory.

21.1 Probability building blocks

In probability theory lingo, the roll of a fair die is a *random phenomenon* because we can't know the precise result in advance. We do know all the potential results, however, and we call these six possibilities {⚀, ⚁, ⚂, ⚃, ⚄, ⚅} the *sample space*. Each element within a sample space (e.g. ⚄) is called an *outcome*. When we talk about the chance that any particular outcome will happen, we are talking about *probability*. We represent probability with a number that's between 0 (will almost never happen) and 1 (will almost surely happen). We denote the probability of an outcome with the notation Pr(outcome), for example, Pr(⚁)= $\frac{1}{6} = 0.167 = 16.7\%$ means you've got a 1 in 6 chance of rolling a two. The corresponding probability for each of the six possible outcomes $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$ is called the *probability distribution*. Notice that the total probability adds to one. This sums-to-one characteristic is true for all probability distributions, since we always distribute a "total amount of probability" across the all possible outcomes in the sample space.

21.2 Probability models

We use the sample space and the probability distribution to define a *probability model*. A probability model is a mathematical description of a random phenomenon. In the case of a die, the probability model looks like this:

outcome	⚀	⚁	⚂	⚃	⚄	⚅
probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

We can use this model to predict specific events. In probability theory, an *event* is one or more outcomes. For example, the probability of the event in which you roll a ⚀ on the first game, then a ⚃ the second game is Pr(⚀) $\cdot$ Pr(⚃) $= \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36} = 2.7\%$. The probability of losing the first game, then winning the second game is Pr(⚀, ⚁, ⚂, ⚃, ⚄) $\cdot$ Pr(⚅) $= (\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6}) \cdot \frac{1}{6} = \frac{5}{36} = 13.8\%$.

Using math tools like summations and functions transformations, we can build probability models that help us understand and make predictions about all kinds of real-world situations. To illustrate this, we'll focus on a particular application: predicting the number of hard disk failures in a data centre. We'll model the number of hard disk failures that occur during one year of operation as the variable Z. We denote as $\{Z = k\}$ the event "k hard disks failed in one year." We denote $\{a \leqslant Z \leqslant b\}$ the event where the number of hard disk failures is between a and b inclusively.

Because the number of hard disk failures is uncertain, Z is an *random variable*. As a matter of convention we use uppercase letters like Z to denote random variables, and lowercase letters like z to denote particular events. When you see the variable Z in an equation you have to keep in mind this is a random variable that can take on any value in the sample space $\{0, 1, 2, 3, \ldots\}$. On the other hand, if you see the expression $z = k$, then we're referring to a particular event which describes the occurrence of exactly k hard disks failing.

We will model the random variable Z with the *Poisson distribution* (to be defined formally in Section ???). The probability distribution of the Poisson model is controlled by a single parameter λ, which represents the rate at which errors occur on average. The Poisson distribution with $\lambda = c$ tells us c hard disks will fail on average, and it also allows us to estimate the probability of $\{Z = c + 1\}$ failures, and in fact any other outcome $\{Z = k\}$.

Figure 21.1 shows a schematic diagram of the Poisson probability model. The Poisson probability distribution is controlled by the parameter λ and produces the random variable Z.

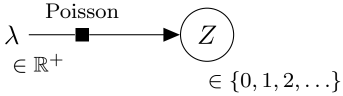

Figure 21.1: Graphical representation the `Poisson model` for the random variable Z distributed according to the Poisson probability distribution with parameter λ.

At the core of every probabilistic model is some probability distribution function that describes the probabilities for all possible outcomes. The probability distribution of the Poisson model is defined by the equation

$$f_Z(k) \equiv \frac{\lambda^k e^{-\lambda}}{k!}, \quad \text{for } k \in \{0, 1, 2, 3, \ldots\}.$$

The probability of the event $\{Z = k\}$ is given by the k^{th} value the probability mass function f_Z:

$$\Pr(\{Z = k\}) = f_Z(k), \quad \text{for } k \in \{0, 1, 2, 3, \ldots\}.$$

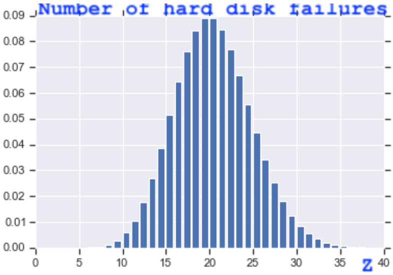

Figure 21.2: Graphical representation the `Poisson model` for the random variable Z with parameter $\lambda = 20$. We see the average number of hard disk failures is 20 but we outcomes from 15 to 25 have significant probability of occurring. Even extreme outcomes like 10 and 30 can be expected to occur.

Figure 21.2 illustrates the variability of possible outcomes we can expect to occur for an poisson model with parameter $\lambda = 20$ (twenty hard disk failures on average). A probabilistic model for the number of hard disk failures allow us to better describe the number of failures that might occur, as compared to just knowing the average $\lambda = 20$. It's true that the average number of failures is 20, but we also see that there is significant probability of observing more-that-average number failures like $\{Z = 25\}$ and $\{Z = 30\}$. The Poisson model allows us to calculate the probabilities of all possible outcomes, including really bad cases like $\{Z \geqslant 30\}$, which describes the case where 30 or more failures occur. Knowing these probabilities can help you plan a robust backup strategy for the data centre that includes worst-case scenarios, rather than just planning for the average-case scenario.

21.3 Applications

In this book we want to focus on applications of probability theory to statistics and machine learning. Probability theory has applications to countless areas like information theory, communications, randomized algorithms in computer science, physics, chemistry, biology, politics, cat memes, etc. To cover all these topics and keep the book a reasonable size would be impossible (Pr = 0). So we'll only mention these other applications in passing and instead focus on the topics that are direct prerequisites for understanding statistics and machine learning topics.

21.3.1 Statistical models

Statisticians use probability models to characterize the statistics of data samples taken from large populations. Given a sample of n observed values from a population and an assumed probability model for the population, statistical reasoning allows us to estimate the population parameters. The process of statistical inference consists of "reverse engineering" the probabilistic generative process. Instead of generating random numbers from a probability distribution with some *known parameters*, we start from some observed data and have to produce a "guess" for the *unknown model parameters*. Crucially to the statistics endeavour, every estimate of a population parameter that we compute comes with an estimate of its accuracy. You can think of statistics a bit like an "educated guessing" strategy: you state your guess *estimate* but also specify a *confidence interval* for your estimate.

To illustrate the notion of statistical inference, let's consider a real-world scenario that makes use of the Poisson model we introduced above. Suppose you're a hard disk manufacturer and you want to estimate the average failure rate λ for the hard drives you produce. You have collected sample data from n identical data centres $\{z_1, z_2, \ldots, z_n\}$. Each z_i is assumed to be an independent observation for the number of hard disk failures generated by the model shown in Figure 21.3. Your job is to compute an estimate $\hat{\lambda}$ (estimates are denoted with a hat on top) for the average error rate and also quantify the variance of your estimate $s^2_{\hat{\lambda}}$. Note what's going on here—we assumed the Poisson model is true and produce an estimate $\hat{\lambda}$ based on this assumption. Note the difference between λ and $\hat{\lambda}$. The parameter λ is an abstract quantity of the underlying math model, and we can never compute its value. We simply assume the parameter λ exists since we're using the Poisson model. In contrast, the estimate $\hat{\lambda}$ is a real quantity computed from the sample $\hat{\lambda} = g(\{z_1, z_2, \ldots, z_n\})$, for some function g. If g is a good estimator function then $\hat{\lambda} \approx \lambda$, but it's not like λ actually exists. It's just a variable in a math model that we assumed to be true. Statisticians often identify the parameters of the math models they use with the parameters of the entire population. In that context, you can think of λ as the parameter estimate you would obtain from an infinitely large sample that includes the whole population, $\lambda = g(\{z_1, z_2, z_3, \ldots\})$.

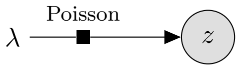

Figure 21.3: A statistical model for the number of hard disk failures. We assume the value z is observed (shown as filled in circle), while the model parameter λ is unknown. Using statistics computed from n independent draws from this model $\{z_1, z_2, \ldots, z_n\}$, we can compute an estimate $\hat{\lambda}$ and an estimate of the variance $s^2_{\hat{\lambda}}$. These estimates can help us make better business decisions.

As a business owner, knowing an estimate of the failure rate $\hat{\lambda}$ for the hard drives you produce can help you answer business questions. How many years of warranty should you offer? What is the estimated cost of replacements under this warranty program? How does the failure rate of hard disks produced by factory one compare with the failure rate of hard disks produced by factory two? Furthermore, knowing the variability of the estimates can help you plan for best case and worst case scenarios, not just the average case.

Statistics is a field of study with more than three hundred years of history. Generations of statisticians have developed various techniques for estimating population parameters based on samples, and computing confidence intervals. It would be vastly optimistic to think that any single book could summarize all these techniques, so the goal of this book is to introduce only the fundamental concepts like *estimators*, *hypothesis testing*, and *sampling distributions*. This is what we'll do in PART III of this book.

21.3.2 Machine learning models

Let's continue with the same running example to describe what a machine learning model for hard disk failure rates could look like. Suppose you're the operator responsible for running your company's data centres. You're interested in the hard disk failure rates in order to plan a redundant data storage strategy that is robust to individual hard disk failures. You have at your disposal historical data of operations from various data centres. The dataset contains information about the workload, operating temperature, hard disk manufacturer, and the resulting number of failures that occurred.

Instead of using a basic model with a single constant parameter λ (the average failure rate), you want to leverage your data to build a rich model that captures how the failure rate depends on the variables workload w (measured as number of reads and writes), temperature t (measured in degrees), and manufacturer m (a categorical variable). Such a machine learning model can be useful for making business decisions like preferring to buy disks from one manufacturer over another because they produce more reliable disks, or setting the temperature in your data centre to reduce error rates.

Figure 21.4 shows a schematic for a machine learning model for this situation. The variables denoted with filled in circles represent observed quantities. Instead of treating the error rate λ as a constant parameter, the machine learning model treats λ as an unknown random variable that depends on the workload, temperature, and device manufacturer. Note this is a composite model that combines a `ReliabilityModel` and the `Poisson` model that we previously discussed.

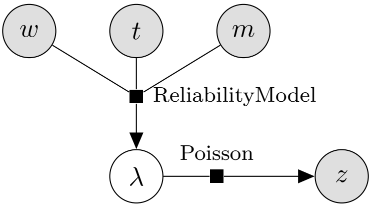

Figure 21.4: A machine learning model for predicting the number of hard disk failures z as a function of different operating conditions like workload w and temperature t, and hard disk manufacturer m.

As a data centre operator, the machine learning model allows you to simulate hypothetical situations. What will happen to the failure rates if we buy more discs from manufacturer A instead of manufacturer B? How much more failures can we expect if we operate the data centres at higher temperature?

In order to obtain this machine learning model, you will usually start from a data set of past observations from different data centres:

$$\begin{aligned}\texttt{data} = [&(w_1, t_1, m_1,\ z_1),\\ &(w_2, t_2, m_2,\ z_2),\\ &\quad\vdots\\ &(w_n, t_n, m_n,\ z_n)].\end{aligned}$$

Based on these `data`, you can train the machine learning model to predict the relationship between the operating conditions (w, t, m) and failure rates z. This is why it's called machine learning, since the goal is to learn the model parameters from past observations.

Machine learning as a field has grown rapidly over the last 50 years, and just like statistics, there are a lot of machine learning techniques you can learn about. The focus of the machine learning topics we'll discuss in PART IV of this book is on topics directly related to statistics. If you know probability and statistics there are some machine learning topics that require very little additional effort for you to learn. In other words, if you already "own" statistics, the machine learning expansion pack in PART IV of the book is "free" in terms of mental effort.

21.4 Overview of probability chapters

In Chapter 22 we'll delve into the core ideas of probability theory. We don't need to go too far into the math at first, and focus on definition the new concepts and general principles used for probabilistic thinking.

We'll follow this up with a condensed chapter of math prerequisites topics. To be frank with you, dear readers, there is no probability theory without math and equations. So no matter how much we as authors try to "simplify" things for you using words and diagrams, in the end of the day you'll need to learn how to handle equations and formulas if you want to truly understand what's going on. In Chapter **??** we'll review concepts from set, functions, calculus, and combinatorics, which are needed for understanding probability theory and statistics.

In Chapter 23 we'll discuss discrete random variables and Chapter 24 we'll discuss continuous random variables Together these chapters will give you an inventory of probability distributions that you can use to build mathematical models for describing random phenomena in later chapters.

In Chapter ???? we'll describe how to use computers to simulate the generative process of any random variable and product random numbers from the appropriate distribution.

Finally we conclude with Chapter **??** which contains extra topics in probability theory. Probability theory is such a vast topics that it would take forever to cover all aspects, so the best w can do is give you some pointers to areas of interest.

Chapter 22

Probability theory

In this chapter we'll introduce the fundamental ideas of probability theory that you need to know for statistics and machine learning. We'll define precisely the basic building blocks like random variables, probability distributions, and expectations.

There are two types of random variables that we'll study in this book: discrete random variables and continuous random variables. Because of the different nature of the underlying samples spaces, different "math machinery" is used for computing probabilities and expectations. In order to keep this chapter concise, we'll give all the probability theory definitions and formulas using the math machinery for discrete random variables. Rest assured, the same results also apply to continuous random variables by changing summations to integrations. We defer the detailed discussion on continuous random variables and continuous probability distributions until Chapter 24.

22.1 Definitions

Let's first establish the notation and terminology for the math objects used in probability theory. Pay attention because we'll be introducing a lot of new concepts and things will go quickly. Probability theory is like a new language you have to learn in order to understand stats and machine learning.

22.1.1 Random variables

Random variables are the main building blocks of probability theory. A random variable X is associated with with a probability distribution f_X that describes the probabilities of the different possible outcomes of a random phenomenon.

- $\mathcal{X}$: the *sample space* is the set of possible values for the random variable X.
- X: a random variable. We use capital letters to denote random variables.
- x: a particular value of the random variable X. We use lowercase letters to denote specific outcomes.

The *sample space* of the random variable X, denoted $\mathcal{X}$, is the set of all possible outcomes of the random variable. For example, we can describe the random phenomenon of rolling a six-sided die using the random variable $X \in \mathcal{X}$, where the sample space is $\mathcal{X} =$ {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}. We can describe the coin toss as a random variable $Y \in \mathcal{Y}$, where the sample space is $\mathcal{Y} = \{\texttt{heads}, \texttt{tails}\}$. In order to do probability calculations, we assign a value to each outcome. In the case of the die, $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$, and in the case of the coin, $\mathcal{Y} = \{0, 1\}$.

22.1.2 Probability mass function

In all the random phenomena we've described so far—rolling a dice, flipping a coin—the possible outcomes are discrete. We can describe the probability distribution of a discrete scenario using a *probability mass function*. The probability mass function f_X tells us the probability of each outcome of the random variable X. If you know f_X you can compute the probability of all possible random outcomes and events of X.

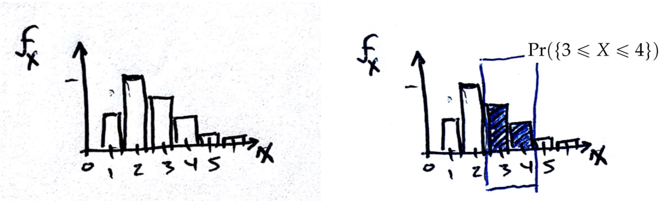

Figure 22.1: Illustration of the probability mass function f_X for some random variable X. The height of each column tells us the probability of this outcome occurring. The area highlighted in the left half of the figure shows the probability of the event $\{3 \leqslant X \leqslant 4\}$, which is the sum of all values $f_X(x)$ for which x satisfies the conditions of the event: $\Pr(\{3 \leqslant X \leqslant 4\}) = f_X(3) + f_X(4)$.

- $f_X \colon \mathcal{X} \to \mathbb{R}$: the *probability mass function*, or *pmf*, of a discrete random variable X tells us the probability of each of the possible outcomes:

$$f_X(x) \equiv \Pr(\{X = x\}), \text{ for all } x \in \mathcal{X}.$$

The function is non-negative: $f_X(x) \geqslant 0$ for all $x \in \mathcal{X}$. The total amount of probability is one $\sum_{x \in \mathcal{X}} f_X(x) = 1$.

- $\Pr(\{a \leqslant X \leqslant b\}) = \sum\limits_{x=a}^{b} f_X(x)$: the probability of the event in which the random variable X takes on any value between a and b, inclusively. This is calculated by taking the sum the probabilities for all outcomes within and including a and b.
- $F_X(x) \equiv \Pr(\{X \leqslant x\})$: the *cumulative probability distribution* of the random variable X, describes the probability of random variable being smaller than x. You can use F_X to compute the probability of X being between a and b using subtraction: $\Pr(\{a \leqslant X \leqslant b\}) = F_X(b) - F_X(a-1)$. The cumulative distribution function is oftern abbreviated as the *cdf* or *CDF*.

The probability mass function of a random throw of a six-sided die has the form $f_X : x \to \mathbb{R}$, $x \in \{1,2,3,4,5,6\}$ and its values are

$$\begin{aligned} f_X(1) &= \tfrac{1}{6}, \quad f_X(2) = \tfrac{1}{6}, \quad f_X(3) = \tfrac{1}{6}, \\ f_X(4) &= \tfrac{1}{6}, \quad f_X(5) = \tfrac{1}{6}, \quad f_X(6) = \tfrac{1}{6}. \end{aligned}$$

Written compactly, we could say $f_X(x) = \frac{1}{6}$, for all $x \in \{1,2,3,4,5,6\}$.

The probability distribution of the six-sided die can also be denoted $\texttt{DiscreteUniform}(1,6)$. It is an instance of the general model for discrete uniform distributions, denoted $\texttt{DiscreteUniform}(a,b)$, which assign equal probabilities to all integers between a and b: $\{a, a+1, a+2, \ldots, b\}$.

The probability distribution for the coin toss has the form $f_Y : \{0,1\} \to \mathbb{R}$ and its values are

$$f_Y(0) = \tfrac{1}{2} \qquad \text{and} \qquad f_Y(1) = \tfrac{1}{2},$$

where 0 is `heads` and 1 is `tails`. The coin toss random variable can also be denoted $f_Y = \texttt{Bernoulli}(\frac{1}{2})$ and is an instance of the general model of Bernoulli random variable $\texttt{Bernoulli}(p)$, which assigns probability p to success and probability $1-p$ for failure of some event.

Note the sum of the values for both probability distributions f_X and f_Y sum to one, which is the normalization convention used throughout probability theory—we assume the total probability is one, and represent probabilities of different events as fractions of this total. For example, the probability of rolling a number between 3 and 5 is given by

$$\begin{aligned}
\Pr(\{3 \leqslant X \leqslant 5\}) &= \sum_{x=3}^{5} f_X(x) \\
&= f_X(3) + f_X(4) + f_X(5) \\
&= \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = \tfrac{1}{2}.
\end{aligned}$$

Note the probability distribution of the composite event $\{3 \leqslant X \leqslant 5\}$ was computed by summing the probabilities of the individual outcomes that make up this event.

Here is a list of common discrete probability distributions:

- `DiscreteUniform`(a, b) variables assign equal probabilities to the integers between the a and b, inclusively.
- `Bernoulli`(p) variables that describe a coin toss with outcomes 0 or 1.
- `Binomial`(n, p) distribution that describes the number of successes in n repeated Bernoulli trials.
- `Geometric`(p) describes the waiting time until the first success in a series of Bernoulli trials.
- `NegativeBinomial`(r, p) is a generalization of a geometric distribution where we wait to obtain r successes.
- `Hypergeometric`(N, n, K) describes the number of successes that will be observed when sampling n balls **without replacement** from a bucket that contains a total of N balls of which K balls are labelled "success" and the remaining $N - K$ balls are labeled as "failure."
- `Poisson`(λ) models the number of times an event occurs in some interval, given that the average number of occurrences is λ.

We'll describe each of these distributions and provide further details about their properties and applications in Section 23.1.

22.1.3 Expectations

Even if the outcomes of the random variable X are uncertain, we can compute various quantities that describe the outcomes of X, on average. The *expectation operator* computes expected values of quantities that depend on X and takes into account the probability of every possible outcome.

- $\mathbb{E}_X$: the expectation operator with respect random variable X.

- $\mathbb{E}_X[g(X)]$: the expected value of the quantity $g(X)$ that depends on the random variable X.
- $\mu_X \equiv \mathbb{E}_X[X]$: the *mean* or *expected value* of X
- $\mathbb{E}_X[X^2]$: the expectation of the quantity X^2. This quantity is also known as the *second moment* of X around the origin.
- $\sigma_X^2 \equiv \mathbb{V}[X] \equiv \mathbb{E}_X[(X-\mu)^2]$: the *variance* of X, also denoted $\mathrm{var}(X)$. The variance is the second moment of X computed around the mean.

Consider a function $g : \mathcal{X} \to \mathbb{R}$ that assigns values to each of the possible outcomes of a random variable X. You can think of g as the payout function in a game of chance based on the random variable X. You obtain $g(x)$ dollars when the outcome of X is x. The expected value of the function $g(X)$ is computed as follows:

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

Note the formula weights the value of each $g(x)$ by the probability of the outcome x, hence the name *expected value*—the value of $g(X)$ will take on different values depending on the random variable X, but $\mathbb{E}_X[g(X)]$ tells us the expected value of g on average, under the randomness encoded in X.

Example Consider the following game involving a six-sided die. You pay \$1 to roll the die and the payout for the game is as follows. If you roll a ⚀, a ⚁, a ⚂, or a ⚃, you win nothing. If you roll a ⚄, you win \$1. If you roll a ⚅, you win \$4. Should you play this game?

The payout function for this game is defined as follows:

$$g(\text{⚀}) = g(\text{⚁}) = g(\text{⚂}) = g(\text{⚃}) = \$0, \quad g(\text{⚄}) = \$1, \quad g(\text{⚅}) = \$4.$$

We'll model the die roll as a random variable X with distribution $f_X(x) = \frac{1}{6}$, for all $x \in \{\text{⚀},\text{⚁},\text{⚂},\text{⚃},\text{⚄},\text{⚅}\}$. The expected gains from this game are

$$\begin{aligned}\mathbb{E}_X[g(X)] &= \textstyle\sum_x g(x)\, f_X(x) \\ &= g(\text{⚀})\tfrac{1}{6} + g(\text{⚁})\tfrac{1}{6} + g(\text{⚂})\tfrac{1}{6} + g(\text{⚃})\tfrac{1}{6} + g(\text{⚄})\tfrac{1}{6} + g(\text{⚅})\tfrac{1}{6} \\ &= (\$0)\tfrac{1}{6} + (\$0)\tfrac{1}{6} + (\$0)\tfrac{1}{6} + (\$0)\tfrac{1}{6} + (\$1)\tfrac{1}{6} + (\$4)\tfrac{1}{6} \\ &= \tfrac{\$1+\$4}{6} \;=\; \tfrac{\$5}{6} \;\approx\; 83 \text{ cents}.\end{aligned}$$

The expected gains of this game is less than the cost of playing, so it doesn't make sense to play.

Computing expectations is crucial for many applications in statistics and machine learning. The function $g(X)$ could represent any quantity of interest in a particular situation. Two important quantities of interest for any random variable are its mean and its variance.

22.1.4 Mean and variance

The *mean* and the *variance* are two properties of any random variable X that capture important aspects of its behaviour.

We compute the *mean* of the random variable X is the expected value of the variable itself:

$$\mu_X \equiv \mathbb{E}_X[X] \equiv \sum_x x\, f_X(x).$$

The mean is a single number that tells us what value of X we can expect to obtain on average from the random variable X. The mean is also called the *average* or the *expected value* of the random variable X. The mean gives us an indication the *centre* of the probability distribution.

The *variance* of the random variable X is defined as follows:

$$\sigma_X^2 \equiv \mathbb{E}_X[(X-\mu_X)^2] = \sum_x (x-\mu_X)^2 f_X(x).$$

The variance formula computes the expectation of the squared distance of the random variable X from its expected value, and gives us an indication of how clustered or spread the values of X are. A small variance indicates the outcomes of X are tightly clustered near the expected value μ_X, while a large variance indicates the outcomes of X are widely spread. The variance σ_X^2 is also denoted $\text{var}(X)$ or $\mathbb{V}[X]$. The square root of the variance is called the *standard deviation*: $\sigma_X \equiv \sqrt{\sigma_X^2}$.

Please take a look a the above formulas and memorize them, because from now on we'll use concepts of μ_X and σ_X^2 a lot, and it's important for you to know the calculations that these quantities refer to.

The *mean* and the *variance* are two properties of any random variable X that capture two important aspects of its behaviour and will be used throughout the entire book. Readers familiar with concepts from physics can think of the mean as the *centre of mass* of the distribution, and the variance as the *moment of inertia* of the distribution.

Expectation formulas

Consider the new random variable $Z = mX + b$ that is defined as the transformation of another random variable X with sample space $\mathcal{X}$ and probability mass function f_X. The random variable Z describe the same underlying random events as the random variable X, but assign different values to the outcomes. The random variable Z describes an *affine transformation* (multiplication by a constant scaling factor m and addition of a constant offset b).

The mean and variance of the variables Z are related to the mean and variance of X.

$$\mu_Z = m\mu_X + b \qquad \text{and} \qquad \sigma_Z^2 = m^2\sigma_X^2.$$

The mean is transformed exactly like the values of the random variable, while the variance "ignores" the constant offset b.

To get a better feeling of the properties of expectation operator consider the following general rules that apply for all expectation calculations:

- The expected value of a constant is the constant itself:

$$\mathbb{E}[c] = c.$$

- Expectation of mX is m times the expected value of X:

$$\mathbb{E}[mX] = m\mathbb{E}[X]$$

- The expected value of a sum of two variables is the sum of their expectations:

$$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

- The variance of the variable $X + b$ the same as variance of the variable X:

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

- The variance of the variable mX is m^2 times the variance of the variable X:

$$\text{var}(mX) = m^2\,\text{var}(X).$$

- The variance can be obtained by computing the second moment of X around the origin $\mathbb{E}[X^2]$ and subtracting the mean squared:

$$\sigma_X^2 \equiv \mathbb{E}[(X-\mu_X)^2] = \mathbb{E}[X^2] - \mu_X^2.$$

  Readers familiar with mechanics will recognize this is an instance of the *parallel axis theorem* for computing the moment of inertia of objects.

Additionally, there are two important formulas that apply for two independent variables X and Y. The expected value of their product is the product of their expectations:

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y],$$

and the variance of the sum of the two variables is the sum of the individual variances

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y).$$

22.2 Explanations

If you're reading this, it means you "survived" your first exposure to math equations and computer code. That's good news, because there is a lot of math equations and algorithmic thinking coming up in the book.

So far in this chapters, we defined all the key concepts and gave some general examples about how probabilistic models are can be useful. The opted for a rushed presentations in order to introduce you to all the "moving pieces," without taking too long. Now that we're done with the whirlwind overview, let's take the time to provide some additional explanations and behind-the-scenes commentary.

22.2.1 Random events

For pedagogical purposes, we started the discussion of probability theory with random variables and skipped the notion of random events. It's now time to repay this "theory debt" and take a moment to dig into the math a little. Specifically, we'll take a look at the underlying representation for *random events* as subsets of a sample space Ω. The *sample space* Ω consists of all possible outcomes of a random phenomenon. An *event* E is a subset of the sample space. You can think of an event as one or more outcomes. Events are defined using a word description {descr}, a math equation $\{g(X) \geqslant c\}$, or in terms of set operation like union, intersection, subtraction, and negation.

- Ω: a sample space that describes all the possible outcomes of a random experiment.
- {descr}: a random event described by the conditions "descr"
- Pr({descr}): the probability of the random event "{descr}"
- $\varnothing$: the empty set denotes a set that contains no elements.
- X: a random variable is a function $X : \Omega \to \mathcal{X}$.

The set Ω is called the *sample space* and contains all possible outcomes of the random event. For example, the case of a random dice throw the sample space is $\Omega =$ {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}, and the sample space for a toss is $\Omega = \{\texttt{heads}, \texttt{tails}\}$.

22.2.2 Probabilities

The notation $\Pr(A)$ describes the probability of event A for some random phenomenon. For example, the probabilities of a fair coin toss are $\Pr(\{\texttt{heads}\}) = \frac{1}{2}$ and $\Pr(\{\texttt{tails}\}) = \frac{1}{2}$. For a six-sided die, the probabilities for different outcomes are $\Pr(\{\text{⚀}\}) = \frac{1}{6}$ and $\Pr(\{\text{⚀},\text{⚁}\}) = \frac{1}{3}$. For simple random events like coin tosses and a dice throws, it is not necessary to invoke the mathematical machinery of random variables and their associated probability distributions since we can compute probabilities directly. A coin toss has two possible outcomes. If the coin is fair, each outcome is equally likely so the probabilities are $\frac{1}{2}$ each. There are six possible outcomes for six sided die, and if the die is fair the probability of each outcome is $\frac{1}{6}$. The probability of the outcome being either ⚀ or ⚁ is defined as the event $A = \{\text{⚀},\text{⚁}\}$, and the probability of event A is computed by comparing the size of the set $|A|$ to the size of the sample space $|\Omega|$:

$$\Pr(\{\text{⚀},\text{⚁}\}) = \Pr(A) \equiv \frac{|A|}{|\Omega|} = \frac{|\{\text{⚀},\text{⚁}\}|}{|\{\text{⚀},\text{⚁},\text{⚂},\text{⚃},\text{⚄},\text{⚅}\}|} = \frac{2}{6} = \frac{1}{3}.$$

Instead of studying specific examples of change occurrences like a coin toss or the throw of a die, we can think about the general principles that apply to *all* random phenomena.

22.2.3 Rules of probability theory

Consider the random phenomenon with sample space Ω. The probability function Pr assigns probabilities to all possible outcomes of the random experiment. A random event A is defined as a subset of the sample space $A \subseteq \Omega$. The probability of the event is defined as

$$\Pr(A) \equiv \text{probability of random outcome falling in the set } A.$$

The number $\Pr(A)$ tells us the "probability weight" of the outcomes in the set A, relative to the weights of other outcomes.

All probabilities satisfy the following conditions:

Axiom 1: Probabilities are always nonnegative numbers:

$$0 \leqslant \Pr(A), \quad \text{for all} \;\; A \subseteq \Omega.$$

Axiom 2: The probability of the whole sample space is one:

$$\Pr(\Omega) = 1.$$

The sample space Ω is the set of all possible outcomes and it contains a total probability weight of one.

Axiom 3: The probability of the union of disjoint events $B_1, B_2, B_3, \ldots$, is given by the sum of the probabilities of the individual events:

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

Taken together, these three conditions are known as the *Kolmogorov axioms of probability*. Note there is nothing particularly profound about the first two axioms—we simply want to establish the general rules of the game for representing probabilities. We use nonnegative numbers to describe the "weights" to the different outcomes, and establish a convention that the "total amount of probability" is one. The third axiom describes the additive structure of probability theory, which tells us how to compute probabilities of composite events. Consider for example two disjoint events B_1 and B_2 ($B_1 \cap B_2 = \varnothing$). The set $B_1 \cup B_2$ describes the event when the result is either one of the outcomes of set B_1 or one of the outcomes of set B_2. The probability of this event is given by the sum of the probabilities of the individual events $\Pr(B_1 \cup B_2) = \Pr(B_1) + \Pr(B_2)$.

All concepts, rules, and equations of probability theory follow from these three basic assumptions. Assume that A and B are events of a random experiment with sample space Ω. Each event is a subset of the sample space $A \subseteq \Omega$ and $B \subseteq \Omega$. We'll now describe some general rules for computing probabilities that are a consequence of the set operations that implement them "under the hood."

- For any event A, the complement of A is denoted A^c, and describes the event that A *does not* happen:

  $$A^c = \Omega \setminus A.$$

  By definition, either A or A^c will happen so $A \cup A^c = \Omega$. Applying the third axiom we know

  $$\Pr(A^c) = 1 - \Pr(A).$$

  The probability of the complement A^c is equal to one minus the probability of event A.
- The probability of the empty set is zero:

  $$\Pr(\varnothing) = 0.$$

- Larger events have larger probabilities. Consider $B \subseteq A$, then

  $$\Pr(B) \leqslant \Pr(A).$$

  If B is a subset A, then the probability of B must be less than or equal to the probability of A.
- Probabilities are numbers between 0 and 1:

  $$0 \leqslant \Pr(A) \leqslant 1, \text{ for all } A \subseteq \Omega.$$

- The logical `OR` of the two events corresponding to the outcomes that are either in set A or in set B, which is the union of the two sets $A \cup B$. The probability of the event $A \cup B$ is given by

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

  The idea behind this formula is to avoid "double counting" the probability of the set $A \cap B$, which is included on both $\Pr(A)$ and $\Pr(B)$.

  Two events E and F are called *mutually exclusive* if $E \cap F = \varnothing$. The probability of their union of such events is equal to the sum of the probabilities: $\Pr(E \cup F) = \Pr(E) + \Pr(F)$.
- The logical `AND` of the two events corresponds to the intersection of the two sets $\Pr(A \cap B)$. Two events are *mutually exclusive* if $A \cap B = \varnothing$.
- For any event A and a sequence of mutually exclusive events $B_1, B_2, \ldots, B_n$ then

  $$\Pr(A) = \Pr(A \cap B_1) + \Pr(A \cap B_2) + \cdots + \Pr(A \cap B_n).$$
- Two events are called *independent* if $\Pr(A \cap B) = \Pr(A)\Pr(B)$.

22.2.4 Conditional probabilities

The *conditional probability* of event B *given* the output of the random event A has occurred is denoted by $\Pr(B|A)$ and computed as

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)}.$$

This formula computes the "weight" of the event B given that the event A has occurred.

For every event A with probability $\Pr(A) \geqslant 0$, we have

1. $\Pr(B|A) \geqslant 0$, for all events $B \subseteq \Omega$.
2. $\Pr(\Omega|A) = 1$
3. Given disjoint sets $B_1, B_2, \ldots$ the union of the conditional probability of these events is given by the sum of the individual conditional probabilities: $\Pr(\bigcup_{i=1}^{\infty} B_i|A) = \sum_{i=1}^{\infty} \Pr(B_i|A)$.

Note these three properties are identical to the Kolmogorov's axioms, but this time we've restricted (conditioned) the events on subsets of events that overlap with the event A.

For any events A, B, and C with $\Pr(A) > 0$, the following probability rules apply:

- $\Pr(B^c|A) = 1 - \Pr(B|A)$
- If $B \subseteq C$ then $\Pr(B|A) \leqslant \Pr(C|A)$.
- $\Pr(B \cup C|A) = \Pr(B|A) + \Pr(C|A) - \Pr(B \cap C|A)$.
- Observe that $A \cap B$ and $A^c \cap B$ form mutually exclusive events and that $B = (A \cap B) \cup (A^c \cap B)$. The probability of event B can therefore be written as two parts:

  $$\begin{aligned} \Pr(B) &= \Pr(A \cap B) &+& \Pr(A^c \cap B) \\ &= \Pr(B|A)\Pr(A) &+& \Pr(B|A^c)\Pr(A^c) \end{aligned}$$

  The first term corresponds to the probability of event B occurs with event A, while the second term computes the probability of the event B and event `NOT`-A, denoted A^c.
- The probability of the event $A \cap B$ can be computed from the conditional probability distribution:

  $$\Pr(A \cap B) = \Pr(B|A) \cdot \Pr(A).$$

  More generally

  $$\Pr(A \cap B \cap C) = \Pr(C|A \cap B) \cdot \Pr(B|A) \cdot \Pr(A).$$

The ability to decompose a complex event $A \cap B \cap C$ as a sequence of products is important for computing probabilities of random experiments that have a sequential structure.

22.2.5 Independent events

We say that the two random events are *independent* if knowledge of one of them does not give information about the other. Independence is very important concept to understand because it serves as the basis of many complex calculations in probability theory and statistics.

If the events A and B are independent then the following statements are true

- $\Pr(A \cap B) = \Pr(A)\Pr(B)$
- $\Pr(A|B) = \Pr(A)$
- $\Pr(B|A) = \Pr(B)$

Furthermore, if A and B are independent, then the complement of the two events are also independent:

- A and B^c are independent
- A^c and B are independent

- A^c and B^c are independent

For three events A, B, and C to be independent, the following four conditions must be satisfied:

- $\Pr(A \cap B) = \Pr(A)\Pr(B)$
- $\Pr(A \cap C) = \Pr(A)\Pr(C)$
- $\Pr(B \cap C) = \Pr(B)\Pr(C)$
- $\Pr(A \cap B \cap C) = \Pr(A)\Pr(B)\Pr(C)$

22.2.6 Bayes' Theorem

A useful result that allows us to compute the conditional probability $\Pr(A|B)$ in terms of $\Pr(B|A)$ or vice-versa is obtained by manipulating our previously developed formulae and expressed as

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}.$$

Conversely we can also compute $\Pr(B|A)$ in terms of $\Pr(A|B)$, $\Pr(B)$ and $\Pr(A)$:

$$\Pr(B|A) = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}.$$

TODO: Example medical test

A collection $B_1, B_2, \ldots, B_n$ is said to be exhaustive if the union of its events is the sample space, in which case $\cup_{i=1}^{n} B_i = \Omega$. The events are said to be mutually exclusive if $B_i \cap B_j = \emptyset$, for all $i \neq j$. A collection of mutually exclusive and exhaustive events forms a partition of Ω. If $B_1, B_2, \ldots, B_n$ are mutually exclusive and exhaustive events and A is any event with $\Pr(A) > 0$, then

$$\Pr(B_k|A) = \frac{\Pr(A|B_k)\Pr(B_k)}{\sum_{i=1}^{n} \Pr(A|B_k)\Pr(B_k)}.$$

TODO: EXAMPLE

Observe that rules of probability do not dictate how the probabilities Pr are computed. The most common approach for describing random events is through the use of random variables and their associated probability distributions.

22.2.7 Random variables and probability distributions

A random variable X is described by a probability distribution f_X. We denote by $\mathcal{X}$ (calligraphic X) the *sample space* of the random variable X, which is the set of all possible outcomes of the random variable. For example, we can describe the outcome of rolling a six-sided die using the random variable $X \in \mathcal{X}$, where the sample space is $\mathcal{X} = \{1,2,3,4,5,6\}$. The number of possible outcomes is six: $|\mathcal{X}| = 6$.

We use the capital letter X when referring to the random variable, and the lowercase letter x to refer to particular outcomes of X. In the standard terminology for probability theory, we refer to the particular outcome x as a *realization* of the random variable X.

22.2.8 Probability mass function

A discrete random variable $X \in \mathcal{X}$ is defined by a *probability mass function* $f_X\colon \mathcal{X} \to \mathbb{R}$ which tells us the probability of each of the possible outcomes:

$$f_X(x) \equiv \Pr(\{X = x\}), \text{ for all } x \in \mathcal{X}.$$

The probability functions is non-negative: $f_X(x) \geqslant 0$ for all $x \in \mathcal{X}$. The total amount of probability is one $\sum_{x\in\mathcal{X}} f_X(x) = 1$. The abbreviation *pmf* is often used to refer to the probability mass function.

Using mathematical notation, we can describe the requirements for a probability mass function f_X as follows:

$$f_X(x) \geqslant 0, \forall x \in \mathcal{X} \;\text{ and }\; \sum_{x\in\mathcal{X}} f_X(x) = 1.$$

The above conditions are a restatement of the Kolmogorov's axioms of probability applied to the probability mass function: the entries of the probability mass function must be nonnegative numbers and the sum of the entries must be one.

The probability mass function is used to compute the probability of the random variable taking on a value in between a and b. The probability of the event $\{a \leqslant X \leqslant b\}$ is given by the expression:

$$\Pr(\{a \leqslant X \leqslant b\}) = \sum_{x=a}^{x=b} f_X(x) = f_X(a) + f_X(a+1) + \cdots + f_X(b).$$

This sum describes the probability of the random variable X taking on one of the values between a and b, inclusively.

Cumulative probability distribution

The *cumulative probability distribution* of the random variable X describes the probability of random variable being smaller than x:

$$F_X(x) \equiv \Pr(\{X \leqslant x\}).$$

You can F_X to compute the probability of X being between a and b using subtraction:

$$\Pr(\{a \leqslant X \leqslant b\}) = F_X(b) - F_X(a-1)$$

The cumulative distribution function is often abbreviated as the *cdf* or *CDF*.

TODO: EXAMPLE

Inverse of the cumulative probability distribution

intervals

TODO: EXAMPLE

22.3 Multiple random variables

So far we discussed random variables in isolation, described by probability distribution, and showed calculate some of their summary statistics. Many real-world scenarios require us to model multiple random variables. In order to work with two random variables X and Y, we can define a *joint probability distribution* f_{XY} that describes the probability of different outcomes of the two variables.

Consider the pair of random variables, defined in the sample space $\mathcal{X} \times \mathcal{Y}$. We denote as $\{X = x, Y = y\}$ the random event where the variable X takes on value x and random variable Y is y. The *joint probability distribution* is a function of the form $f_X \colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, that tells us the probability of each of the possible outcomes:

$$f_{XY}(x,y) \equiv \Pr(\{X = x, Y = y\}), \text{ for all } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}.$$

Like all probability distributions, the joint probability distribution has non-negative values, $f_{XY}(x,y) \geqslant 0$ for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$, and the total amount of probability is one $\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f_{XY}(x,y) = 1$.

- $F_{XY}(x,y) = \Pr(\{X \leqslant x, Y \leqslant y\})$: joint cumulative distribution
- $f_X(x) = \sum_{y\in\mathcal{Y}} f_{XY}(x,y)$: the *marginal distribution* for the random variable X.
- $f_Y(y) = \sum_{x\in\mathcal{X}} f_{XY}(x,y)$: the *marginal distribution* for the random variable Y.
- $f_{X|Y}(x|y) = \Pr(\{X = x\}|\{Y = y\})$: the conditional distribution of X given Y.
- $f_{Y|X}(y|x) = \Pr(\{Y = y\}|\{X = x\})$: the conditional distribution of Y given X.

- $f_{XY}(x,y) = f_X(x)f_Y(y)$: the probability distribution of two *independent* random variables. When X and Y are independent random variables, their joint distribution consists of a product of two independent distributions. The randomness of X does not depend on the randomness of Y and vice versa.
- $\textrm{cov}(X,Y) \equiv \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$: the *covariance* of the random variables X and Y
- $\rho_{XY} = \frac{\textrm{cov}(X,Y)}{\sigma_X\sigma_Y}$: the *correlation coefficient* of the random variables X and Y is computed as the correlation of the two variables divided by the product of their standard deviations.

The joint probability distribution f_{XY} is our main tool for modelling relationships between two random variables X and Y. By choosing the appropriate function f_{XY}, we can describe and model various relationships between random variables. We can model cases when one random variable depends on the other, cases when the random variables are correlated, or the case when the variables are *independent*. The joint distribution when the random variables X and Y are *independent* can be written as the product of two single-variable distributions $f_{XY}(x,y) = f_X(x)f_Y(y)$.

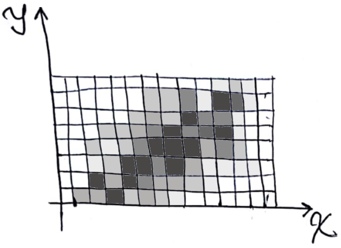

Figure 22.2: Graphical representation of a joint probability distribution $f_{XY} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where $|\mathcal{X}| = 14$ and $|\mathcal{Y}| = 8$. The darkness of each square (x,y) represents is proportional to its mass.

The marginal distributions f_X and f_Y are obtained from the joint distribution f_{XY} by summing over all possible values for the other variable:

$$f_X(x) = \sum_{y\in\mathcal{Y}} f_{XY}(x,y) \quad \textrm{and} \quad f_Y(y) = \sum_{x\in\mathcal{X}} f_{XY}(x,y).$$

The idea for a marginal distribution f_X is to get rid of the Y randomness by *marginalizing* it, which means summing over all its possible values. The marginal distribution f_X describes the randomness of X

when we don't know the value of Y. Similarly the marginal distribution f_Y describes the randomness of Y, ignoring the random variable X.

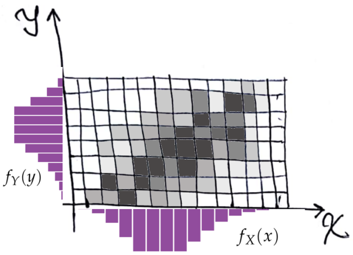

Figure 22.3: Marginal distribution f_X is obtained by summing all the values of f_{XY} in each column. Marginal distribution f_Y is obtained by summing all the values of f_{XY} in each row.

The marginal distributions are also used to define the *conditional distributions* $f_{X|Y}$ and $f_{Y|X}$:

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} \quad \text{and} \quad f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}.$$

The vertical bar is pronounced "given" and describes situations where the realization of some random variables is known. For example, the conditional distribution $f_{Y|X}(y|x_a)$ describes the probabilities of the random variable Y, given we know the value of the random variable X is x_a. The distribution $f_{Y|X}(y|x_b)$ describes the separate case when $X = x_b$, and in general there is a different distribution for each of the possible $x \in \mathcal{X}$.

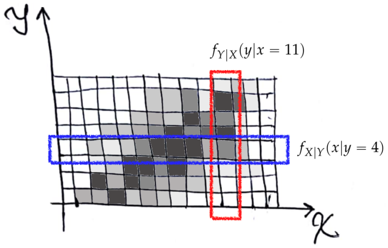

Figure 22.4: Conditional distributions $f_{Y|X}(y|x)$ represent different vertical slices through the joint distribution. Similarly, conditional distributions $f_{X|Y}(x|y)$ are horizontal slices of the joint distribution.

Independent, identically distributed variables

The analysis of samples consisting of multiple, independent draws from a random variable is very important in probability theory and in statistics. Consider some random variable X and a sample that consists of n draws from the variable.

- X: a random variable with probability distribution f_X
- $(X_1, X_2, \ldots, X_n)$: n draws from the random variable X. Each draw represents an independent copy of the random variable X with the same distribution, $X_i \sim f_X$.
- $\{x_1, x_2, \ldots, x_n\}$: a particular sample of size n.

The joint probability distribution for the n random variables is

$$f_{X_1 X_2 \cdots X_n}(x_1, x_2, \cdots, x_n) = f_X(x_1) f_X(x_2) \cdots f_X(x_n).$$

Note the joint probability distribution is the product of n copies of the probability distribution of the random variable X. We call this an *independent, identically distributed* sample, or *i.i.d.* for short. This *product structure* implies that the random variables are independent, and each copy of X has the same distribution f_X, hence the name.

Collecting and analyzing samples form distributions is an important strategy for studying random variables. Observing a single instance x_i of a random variable X doesn't tell us much since the variable is random and can take on any value in the set $\mathcal{X}$. However, if we collect n independent observations all drawn from the same distribution, $\{x_1, x_2, \ldots, x_n\}$, we can start to see patterns in the randomness, compute statistics, and make inferences about the probability distribution.

Correlated random variables

If X and Y are not independent, we can quantify the amount of correlation between them using the *covariance* calculations

$$\text{cov}(X, Y) \equiv \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y],$$

which computes the expectation of the product XY and subtracts the product of the individual random variables. Thus correlation coefficient $\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$ computes the ratio of the correlation to the product of the standard deviations of the two variables. The covariance and correlation coefficient both provide indicators about an underlying linear relationship between the random variables X and Y. Recall that for independent random variables $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ so the covariance and correlation coefficient will both be zero.

22.4 Probability models

Throughout this book we'll use random variables to describe various situation in statistics and machine learning. The probability distributions mentioned above (Uniform, Binomial, Poisson, etc.) are examples of the basic building blocks of randomness that are available to use when trying to model random scenarios. Different situations call for different probability distributions, but there are some general ideas apply to all probability models.

- `model`: the probability model that describes the scenario.
- θ: the model parameters
- X: the random variable produced by a probability model
- $\{x_1, x_2, \ldots, x_n\}$: a sample of n observations from the model
- $\hat{\theta}$: an estimate of the model parameters, usually obtained from a sample of observations.

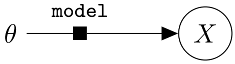

Figure 22.5: Graphical representation for the probabilistic model `model` with parameters θ, which describes how the random variable X is generated.

The model parameters θ represents the control knobs for the probability model. Each probability model has different parameters that control the distribution. For example, a Bernoulli trial (coin flip) has a single parameter p, which represents the probability of success (heads). Another example of a model parameter is the average error rate λ, which we used to build the Poisson model for the number of hard disk failures. Every probability model has a different set of parameters, which are often denoted with lowercase letters of the Greek alphabet. In this book, we use the symbol θ to refer to the model parameters "in general" without specifying a particular choice of model.

22.4.1 Computer models for random variables

Once we know the probability distribution f_X for a random variable, we can compute the probability of any outcome and calculate any statistical property of interest for that random variable. Computers models can be very helpful for such statistical calculations. Doing hands on calculations, plotting, and generating samples from probability distributions are very useful to help you understand and "own" the equations.

Here are some things you can do using a computer model of a random variable X:

- Obtain values of the probability mass function f_X for every possible outcome $\{X = k\}$ and composite events $\{a \leqslant X \leqslant b\}$ using the summation $f_X(a) + f_X(a+1) + \cdots + f_X(b)$.
- Obtain values of the cumulative distribution F_X and compute probability of the event $\{a \leqslant X \leqslant b\}$ using the difference $F_X(b) - F_X(a-1)$.
- Compute values of the percent point function, which corresponds to the inverse cumulative distribution $F_X^{-1}(q)$. Percentiles and quartiles represent values of x for which some percentage of the population is smaller. The first quartile $x_{25} = F_X^{-1}(0.25)$ corresponds to the value of x where the cumulative distribution F_X equals 0.25. The second quartile is $x_{50} = F_X(0.5)$ and is equal to the median of the distribution: the value of x for which half of probability mass f_X is on the left and the other half of f_X is on the right. Percentiles are useful when we're interested in giving guarantees. For example we could give a "worst case" scenario for the number of hard disk failure rates we guarantee that that the number of hard disk failure rates won't be exceeded this value 90% of the time.
- Compute α-confidence intervals I_α for the outcomes of X. The α-confidence interval is defined as $I_\alpha \equiv [F^{-1}(\frac{\alpha}{2}), F^{-1}(1-\frac{\alpha}{2})]$, and contains $(1-\alpha)$ of the probability mass of the function f_X. Confidence intervals are used to give two-sided guarantees for the range outcome values that are likely to occur. The probability of observing a value outside of the confidence interval is bounded by α: $\Pr(\{X \notin I_\alpha\}) \leqslant \alpha$.
- Compute the expectations of any function $g(X)$ which depends on the randomness of X.
- Compute properties of the random variable like the mean, median, mode, variance, standard deviation, and others.
- Generate random samples from the random variable. You can ask the computer generate a random draw where the probability of each outcome is proportional to f_X. Such computer-generated samples can be useful for visualization and numerical simulation experiments.

The `SciPy` library provides the module `scipy.stats.distributions`, which contains ready-made models for various probability distributions. Let's illustrate some of the "features" of these computer models by working through a concrete example.

Example

Suppose you're the operator of a data centre and you want to estimate the number X of hard disk failures you can expect this month. Using some numbers from the manufacturer's data sheets and the specifics of your data centre you know that $\mu = 20$ hard disk failures will occur on average.

Sometimes knowing the average is not good enough for the type of questions you need to answer. What is the probability of observing 21 hard disk failures? What is the probability of having 25 hard disk failures or less? Can you give a range of outcomes that will occur 95% time? This is the type of questions your colleagues are interested in. The software engineering department needs to know the probabilities of different events to design the redundant storage system, and legal wants to know "worst case" scenarios in order to know how to draft the service level agreement (SLA) documents. The finance people are interested in estimating costs of replacement disks, so they're also asking you to give certain estimates.

All these requests for estimates are piling up in your inbox, but despite your interest in data science topics you never seem to find the time to do the probabilistic modelling exercise needed to answer the questions because you're busy fixing the servers, managing network capacity, and paying electricity bills. One day during a high-level meeting, your colleagues decide to gang up on you and to complain loudly about the lack of estimates, and put you on the spot in front of everyone. You decide to get this done right then and there, and tell everyone

> "Relax, everyone, we can do this right now. I know Python."

Everyone is immediately reassured.

You know the Poisson family of probability models is well suited to describe the random number of hard disk failures in general. The parameter $\mu = \lambda = 20$ for the Poisson model describes your data centre in particular, since the expected number of hard disk failures for your data centre is $\mu \equiv \mathbb{E}_X[X] = 20$. You proceed to share your screen so everyone can see, open a Python shell, and start typing

```
>>> from scipy.stats.distributions import poisson
>>> rv = poisson(mu=20)
```

The code above imports the `poisson` model from the `SciPy` package `scipy.stats.distributions` and creates a instance of it called `rv` with parameter $\mu = \lambda = 20$. Before you proceed with the calculations, you want to review the names of the methods available on the `rv` that you'll be using. You type in "`rv.`" then press TAB to see all the methods available on that object. The most useful methods are listed below.

- `rv.pdf`$(k) \equiv f_X(k)$. Use this method to obtain the value of the probability mass function $f_X(k) = \Pr(\{X = k\})$.
- `rv.cdf`$(k) \equiv F_X(k)$. Use this method to evaluate the cumulative distribution function $F_X(k) = \Pr(\{X \leqslant k\})$, which corresponds to the sum of the probabilities of events smaller than k. The CDF is useful for answering questions about the probabilities falling in any interval.
- `rv.ppf`$(q) \equiv F_X^{-1}(q)$. The *percent point function* corresponds to the inverse of the CDF function. This function is needed for inverse-probability questions like percentiles which correspond to guarantee values that the random variable won't exceed 90%, 95%, or 99% of the time.
- `rv.interval`$(\alpha) \equiv [F^{-1}(\frac{\alpha}{2}), F^{-1}(1-\frac{\alpha}{2})]$. Use this method to compute two-sided confidence intervals.
- `rv.rvs(size=`n`)`. This method is useful for generating instances of the random variable. Calling `rv.rvs(size=1)` will return an instance of the random variable. Calling `rv.rvs(size=100)` will return a sample of size $n = 100$ generated from the same random variable.
- The object `rv` also has numerous methods for computing statistics of the random variable.

  ▷ `rv.mean()` $\equiv \mu_X$: computes the mean of the distribution.

  ▷ `rv.var()` $\equiv \sigma_X^2$: computes the variance of the distribution. Recall that the *standard deviation* is defined as the square root of the variance: $\sigma_X \equiv \sqrt{\sigma_X^2}$. The method `rv.std()` $\equiv \sigma_X$ can also be used to compute the standard deviation.

  ▷ `rv.stats()` $\equiv [\mu_X, \sigma_X^2]$: computes both the mean and the variance of the distribution.

  ▷ `rv.median()` $\equiv F_X^{-1}(\frac{1}{2})$: computes the median of the distribution.

- `rv.expect`$(g) \equiv \mathbb{E}_X(g(X))$. The `expect` method computes the expected value of any function $g(X)$ with respect to the distribution. Recall that the mean is defined as the expectation $\mu_X \equiv \mathbb{E}_X[X]$, and similarly the variance is defined as $\sigma_X^2 \equiv \mathbb{E}_X[(X-\mu_X)^2]$. Additionally, there are methods for computing the n^{th} moment `rv.moment`$(n) \equiv \mathbb{E}_X[X^n]$ and the entropy of the distribution `rv.entropy()` $\equiv \mathbb{E}_X[\log(X)]$
- `poisson.fit`$(\{x_1, x_2, x_3, \ldots, x_n\}) = \hat{\lambda}$. The `fit` method can be used to compute an estimate $\hat{\lambda}$ of the model parameters from a sample data $\{x_1, x_2, x_3, \ldots, x_n\}$. We assume that sample comes from some Poisson distribution with an unknown parameter λ, and compute a "guess" $\hat{\lambda}$ for this parameter.

Feeling reassured by the plethora of methods available to you, you explain what you want from your colleagues during the meeting: "I want you to give me any probability question related to hard disk failure rates, and I'll use the probability model to answer your question to the best of my ability. Right now. Live."

There is a moment of silence in the room as people are processing your directives. You decide to use the time to compute the probability of some outcomes:

```
>>> rv.pmf(20)
0.0888353173920848

>>> rv.pmf(21)
0.08460506418293791

>>> rv.pmf(22)
0.07691369471176195
```

You explain to your colleagues this means the probability of observing 20 failures next month is 8.88%, the probability of observing 21 failures is 8.46%, and the probability of 22 failures is 7.69%.

Alice from accounting interrupts with a question. "Wait, I thought you said the expected value is 20. Now you're telling us there is just 8% chance of that happening?"

"Yes, the average is $\mu = 20$, but we could have 21, 22, 23, or any other number of failures next month."

"So we can't know for sure how many failures will occur?"

"No, we can't know for sure since failures are random, but we can think about the different possible outcomes and plan accordingly. For example, we could run simulations to—." You stop yourself mid-sentence because you sense this meeting can go on forever if you start explaining each concept in detail. Better show than tell.

In order to better describe the range of values for the random variable X, you compute the two important statistics of the probability distribution:

```
>>> rv.mean()
20.0

>>> rv.std()
4.47213595499958
```

You interpret these numbers for your colleagues by saying: "This means that we can expect 20 plus or minus 5 failures on average."

"What do you mean 'plus or minus 5'?" asks Bob from sales.

"I mean that the number of failures will likely be roughly between 20-5=15 and 20+5 = 25." You then proceed to compute the exact probability by summing the probabilities of the individual outcomes in that range.

```
>>> sum( [rv.pmf(k) for k in range(15,25+1)] )
0.782950746174042      # = Pr({ 15 <= X <= 25 })
```

In other words, 78.2% of data centres like ours will experience between 15 and 25 failures. Mathematically speaking, the number above corresponds to the probability of the event $\{15 \leqslant X \leqslant 25\}$, which can also be calculated using `rv.cdf(25) - rv.cdf(15-1)`.

You then say "Here is a plot that shows the probabilities of all the outcomes," while typing in the commands:

```
>>> n = 40
>>> k = numpy.arange(0,n)
>>> fX = rv.pmf(k)
>>> matplotlib.pyplot.bar(k, fX, color='b')
```

The graph generated as the output of this command is shown in Figure 22.6.

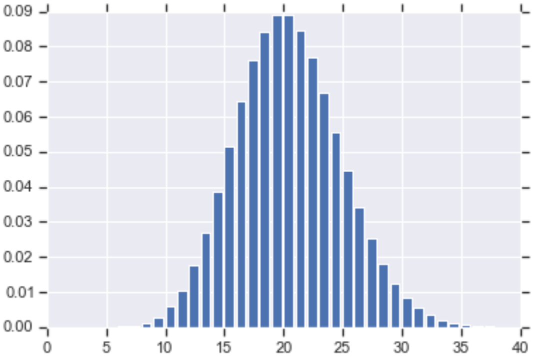

Figure 22.6: Plot of the probability mass distribution of a poisson model with parameter $\mu = \lambda = 20$. The possible outcomes are clustered around $k = 20$ with most of the probability mass falling in the range $[15, 25]$.

Desiring to keep the conversation going, you ask "Other questions?"

"I have one." says Charlotte from software engineering. "I want to know, what is the maximum number of failure rates that I should plan for."

"I can't answer that question because, theoretically speaking, any number of failures can occur. What I can do is give you a 95% confidence interval," you explain as you're typing this in:

```
>>> rv.ppf(0.95)
28
```

To explain what this number means, you say "95% of the data centres like ours will not observe more than 28 failures." If you plan for

28 failures as the worst-case scenario, you know there is only a 5% chance that what happens next month will not be covered.

David from the marketing department has a question. “Is this thing that you did just now AI?,” he asks, thinking about how he can use this in the webcopy for the new company website.

“I suppose you *could* say that, since we’re using probability distributions and probability distributions are used in AI,” you explain, stretching the definitions.

“What about blockchain? Are we using a blockchain for this?”

“No blockchain,” you interject. “Listen David, let’s proceed one buzzword at a time. You can have “AI” for now. Show me you can sell $1M worth of product with it, then come back to me and I’ll find another buzzword for you.”

“Okay, deal! I can work with that. AI is hot these days.”

Looking around the room you sense the meeting is coming to a close. Everyone is feeling good about their first data science experience. You decide to wrap things up with some random number generation. “To close the meeting, let me show you some examples of the possible number of hard disk failures we can expect to see during the next year,“ you say while running the command needed to generate 12 random samples from the random variable rv:

```
>>> rv.rvs(12)
[20, 26, 18, 23, 13, 23, 22, 15, 26, 21, 19, 11]
```

You hear a few people in the room say “wow.” Visuals always work. Finally people get it—the average is 20, but the number of failures can vary a lot around that average.

Later that day you receive a followup email from Emily from the purchasing department. She wants an estimate of the total cost she should budget for replacement hard disks, given a base price of $200/disk and a bulk discount of $150/disk if buying 20 or more disks. In other words Emily is asking you to compute $\mathbb{E}_X[g(X)]$ where $g(k)$ is the cost function for purchasing k replacement disks. You quickly compute the answer by following the formula:

```
>>> def g(k):
...     if k > 20:
...         return 150*k
...     else:
...         return 200*k

>>> EXg = sum([g(k)*rv.pmf(k) for k in range(0,100)])
>>> EXg
3470.2572668392268
```

The expected value figure is computed by taking all possible outcomes and weighing the cost in each case by the probability of this outcome to occur.

I hope reading about this real-world scenario convinced you of the general usefulness of computers for doing probability calculations. The above code examples were using Python, but you obtain the same answers just as easily using the probability functions available in R or in spreadsheet software (see problems P23.7 and P23.8 for that).

Note that the `poisson` is just one of the random variable models available in `scipy.stats.distributions`. Some other models defined in that package are: `randint`, `bernoulli`, `binom`, `geom`, `nbinom`, `hypergeom`, `poisson`. Each of these models provides you with the same set of methods for computing probabilities, confidence intervals, and generating random values. You'll have to wait until Chapter 23 to learn the detailed story about these different families of probabilistic models. For now you can think of probability models as various building blocks available to describe different random processes, with applications in different situations.

22.4.2 Statistical inference

At the core of both statistics and machine learning lies the general notion of *statistical inference*, which is the act of computing estimates of model parameters based on observed data. Listen up, because this section contains the "main idea" that we'll use throughout the rest of the book.

Consider the random variable X, which is an instance of the model `ProbabilityModel` with parameters θ:

$$\texttt{X} = \texttt{ProbabiliyModel}(\theta).$$

In all the examples above we assumed that the model parameters θ were known and used the model to probabilities of different outcomes.

Suppose instead that θ as an unknown parameter, and instead we have a sample of observations $\{x_1, x_2, \ldots, x_n\}$ from the random variable X. Figure 22.7 is a graphical representation of the probability scenario. The node x is shown as "filled in" which means it is an observed quantity. The variable x represents a particular outcome of the random variable we're studying, and the fact that we have n independent observations is denoted with the box with label n.

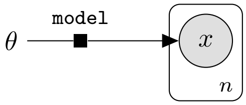

Figure 22.7: Graphical representation for the probabilistic model for analyzing samples of n observations $\{x_1, x_2, \ldots, x_n\}$. The probabilistic model `model` with parameters θ defines the probability function f_X for how the random variable X is generated.

Statistical inference is the process of computing an estimate of the model parameters, denoted $\hat{\theta}$, based on observed data $\{x_1, x_2, \ldots, x_n\}$, which we assume comes from the model.

The rest of this book is dedicated to various statistical procedures for computing estimates of model parameters. In PART III of this book, we'll learn various ways to compute estimates of model parameters $\hat{\theta}$ and also quantify the variance of these estimate $s^2_{\hat{\theta}}$. In PART IV of the book we'll discuss statistical inference in the context of machine learning applications. I want you to keep this in mind as you plot through the next few chapters of probability theory perequisites. I won't lie to you and tell you it will be easy. There will be lots of equations, code samples, and expectations imposed upon you to try things out for yourself. There is no other way out of this. In order to understand statistics and machine learning topics, you'll need a solid grasp of all the concepts of probability theory. I'm talking not just "I've heard of these things"-kind of knowledge, but real intuitive understanding and hands-on experience. This is the goal rest of the chapters in PART II of the book.

22.5 Discussion

Before we move on, here are two important comments and clarifications about the above material.

22.5.1 Continuous random variables

So far in this chapter we focussed on *discrete random variables* whose sample space is a finite set, like the throw of a dice $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$, or a countably infinite set, like the number of hard disk failures $\mathcal{Z} = \{0, 1, 2, 3, \ldots\} = \mathbb{N}$.

In contrast, *continuous random variables* can take on continuous values and represent quantities like time, length, or other smoothly varying quantities. The sample space $\mathcal{X}$ for a continuous random variable consist of some subset of the real numbers, and the probabilities are computed using a probability density function.

All the rules of probability and formulas we introduced in this chapter apply to continuous random variables as well, with the change of summation to integration. Recall the expectation operator $\mathbb{E}_X$, which computes the expected value of quantities that depend on the randomness in X. Suppose X is a continuous random variable and random variable $G = g(X)$ is defined for some function $g : \mathcal{X} \to \mathbb{R}$, then the expected value of G is

$$\mathbb{E}_X[G] = \mathbb{E}_X[g(X)] = \int_{x\in\mathcal{X}} g(x) f_X(x)\, dx.$$

Note the idea behind the expectation operator is that same—weight each value $g(x)$ by the probability of that outcome—but instead of a summation over a finite set of values, we perform integration over a continuous range of x values.

The expectation operator is important because it's used computing the mean μ

$$\mu \equiv \mathbb{E}_X[X] = \int_{x\in\mathcal{X}} x f_X(x)\, dx,$$

and variance σ^2 statistics:

$$\sigma^2 \equiv \mathbb{E}_X[(X-\mu)^2] = \int_{x\in\mathcal{X}} (X-\mu)^2 f_X(x)\, dx.$$

The mean tells is where the centre of distribution lies, while the variance tells you how spread out the distribution is around that centre. We defer the detailed discussion on continuous variables and their properties until Chapter 24.

22.5.2 Probability verbs

I know this has been a long chapter with a lot of information, thanks for sticking with it until the end. Before we move on, let's review the new terminology that we introduced in this chapter. We can think of probability theory as a language that includes specialized nouns, verbs, adjectives, and adverbs for describing random events. Let's see some of that.

First let's review the probability nouns like *sample space*, *event*, *random variable* X, and *probability distribution* (a function $f_X : \mathcal{X} \to \mathbb{R}$). We also referred to the specific outcomes of the random variable as *realization* and denoted with lowercase x.

We also introduced a lot of new probability verbs to describe random variables. When using the passive voice to referring to the random variables, we say they are *distributed* (according to a distribution) or *drawn* (from a distribution). If instead we take probability distributions as the subject of a sentence, then we say the probability distribution *assigns* certain values to given outcomes of the random variable. We can also put ourselves in the driver seat and say that we *generate* or *draw* random samples from the distribution. Verbs like *compute* and *calculate* are also used often in probability, but note the same verb could refer to very different types of computations. We can compute the value of a probability distribution, or compute an expectation (a weighted sum of some quantity according by the probability), or compute an estimate of a model parameter (by following a statistical inference procedure).

We also learned about some adjectives and adverbs that we can apply to random events. We used the adjective *random* throughout the chapter to describe probabilistic events. The adverb *randomly* can similarly be applied to verbs: *randomly distributed*, *randomly chosen/selected*, *randomly drawn*, etc. The adjective *independent* and the adverb *independently* describe random events that do not influence each other. Recall also the mouthful-of-an-expression *independent, identically distributed* (or *i.i.d* for short), which refers to multiple draws: $X_1, X_2, \ldots, X_n$, where the distribution of each X_i is identical (come from the same probability distribution $X_i \sim f_X$), and the draws are independent. The probability distribution for i.i.d. variables $X_1 X_2 \cdots X_n$ is $f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i)$.

Chapter 23

Discrete probability distributions

Feedback Some suggestions to consider

- I think we could move discussion of distributions and pmfs from 02 - probability theory to the introduction of this chapter.
- Consistency note: We use `this formatting` for outcomes and distribution families sometimes, and sometimes we don't. Sometimes the families are capitalized and sometimes they're not.
- In my opinion, it would be better to introduce all the families of distributions first, then go over how to simulate them on the computer and how they are related to one another.
- Either here or in the prob theory chapter, we need to tell readers what it means when we say a "draw from a distribution"
- We're using capital N here for number of trials, but only in some cases. Changed to lower case n. Also changed some capital Ks to lower-case ks.
- Just a reminder that we have code here for Excel, R, and Python. Not sure if you've decided yet whether we have space to include them all in every chapter.
- Not sure which distributions you'll use in machine learning, but I think you could cover geometric, hypergeometric, and negative binomial in less detail.

23.0.1 Probability distributions and probability mass functions

The *probability distribution* of a random variable X is a description of the probabilities associated with the possible values of X. In the case of discrete random variables, the distribution is easily specified by a table of possible values and their probability of occurring . In some cases, we can express the probability as a formula.

Definition 23.0.1. For a discrete random variable X with possible values $x_1, x_2, \dots, x_n$, a *probability mass function* is a function, f_X, is defined pointwise on $\Omega = \{x_1, x_2, \dots, x_n\}$ such that $f(x_i) = P(\{X = x_i\})$ and furthermore satisfies;
(i) $f_X(x_i) \geqslant 0, \forall i$
(ii) $\sum_{i=1}^{n} f_X(x_i) = 1$

23.0.2 Cumulative distribution functions

Definition 23.0.2. The *cumulative distribution function* (CDF), $F(x)$, of a discrete random variable X is defined by

$$F(x) = P(X \leqslant x) = \sum_{x_i \leqslant x} f(x_i)$$

and has the following properties
(i) $0 \leqslant F(x) \leqslant 1$
(ii) If $x \leqslant y$ then $F(x) \leqslant F(y)$.

23.0.3 Mean and variance of a discrete random variable

- The *mean* or *expected value*, μ or $E(X)$, of a discrete random variable X is
$$\mu = E(X) = \sum_{x \in X} x f(x)$$
- The *variance*, σ^2 or $V(X)$, of X is
$$\sigma^2 = V(X) = E(X - \mu)^2 = \sum_{x \in X} (x - \mu)^2 f(x) = \sum_{x \in X} x^2 f(x) - \mu^2$$
- The *standard deviation* of X is $\sigma = \sqrt{\sigma^2}$.

Furthermore, the expected value (mean) of a function of a discrete random variable is

$$E\left[h(x)\right] = \sum_{x} h(x) f(x)$$

. The expected value is linear so that $E(aX + b) = aE(X) + b$. In contrast, the variance of a linear term scales quadratically with the coefficient $V(aX + b) = a^2 V(X)$.

23.1 Discrete distributions reference

We'll now give an overview of some of the most important discrete probability distributions. The idea is to give you a quick overview of the these building blocks, and provide you with a fact sheet to refer to when you need to lookup facts and formulas related to these distributions. In this chapter, we use the notation $X \sim f_X$, which is read as "X is distributed according to f_X". This means that the random variable X has properties of the distribution specified by f_X.

23.1.1 Bernoulli

A *Bernoulli trial* is an experiment which results in either true of false, positive or negative, heads or tails, zero or one, or some other binary choice. The distribution is named after the mathematician Bernoulli who did some important early work in probability theory.

A random variable $X \sim \texttt{Bernoulli}(p)$ has the probability mass function

$$f_X(0) = 1 - p, \qquad f_X(1) = p.$$

In the case of a coin toss, the `heads` outcome $\{X = 1\}$ has probability p, while the `tails` outcome $\{X = 0\}$ has probability $1 - p$. If the coin is fair, then $p = 0.5$ and $1 - p = 1 - 0.5 = 0.5$. Both `heads` and `tails` have a $p = 0.5$ chance of occuring. Many random phenomena besides coin tosses have two possible outcomes: `success` or `failure`, such as... . Bernoulli trials also serve as building blocks for other probability distributions that we're about to learn about.

- The binomial distribution is defined as the count of success outcomes in n Bernoulli trials.
- The geometric distribution describes the waiting time until first success in sequence of repeated Bernoulli trials.

Can we use computers to generate random numbers? I'm asking you now in case you already know because we've just seen the mathematical notion of a coin flip. How can you make computers flip the coins for you?

It turns out computers have random number generators that can be used for this exact purpose. The random numbers generated by computers are called *pseudorandom*. Essentially, we start with some seed number and then perform lots of operations on it, twisting and shifting it until we generate a "random" output number. The number is not truly random, since the a generator that starts with the same seed always produces the same output number. This is why we call these computer-generated random numbers "*pseudo*random".

The the promises provided by a random number generation algorithm Is not that it will produce random numbers for you out of nowhere, What if you feed it with a random number you to start generating additional random numbers based on the random seed.

For the purposes of this book, we'll be doing computer simulations of random phenomena. For these, we'll use pseudorandom numbers to represent random outcomes. As long as the result looks random, it will work for these demonstrations. Know that the notion of random number generation for the purposes of cryptography has much more stringent requirements about how random numbers are generated. The discussion of cryptography and actual random number generation is beyond the scope of this book.

Most computer languages have a built-in functionality for generating pseudorandom numbers. In Excel the function `RAND()` to generate a random number between 0 and 1. The equivalent function in R is `runif(1)`, which is short for random uniform and the number indicates we just want one draw. In Python, you can import the module called `random` (use the statement `import random`), then generate random numbers between 0 and 1 using `random.random()`.

If you want to simulate the outcomes of the random variable $X \sim \texttt{Bernoulli}(p)$, you can draw a random number using one of the above functions, then declare success if the random outcome you see is smaller or equal to p and failure if the value you see is greater than p. Since the random numbers generated by the computer are uniformly distributed between 0 and 1, your outputs will be "success" with probability p of the time an "failure" with the remaining probability $(1-p)$. .

The `scipy` module for Python, provides the function `bernoulli.rvs(p)` which you can use to draw random variables from $\texttt{Bernoulli}(p)$.

The random number generator functions `RAND()/runif(1)/random.random` are essential building blocks we need to generate random samples from *any* distribution. We'll talk more about using computers to generated samples from various distributions in Section [COMPUTER-GEN PROB CHAPTER IN PROB PART].

23.1.2 Uniform

The *discrete uniform distribution* $\texttt{DiscreteUniform}(a, b)$ describes the random phenomenon of picking a number between a and b at random, where each number has an equal chance of being selected. This distribution is called "uniform" since it assigns the same probability to each of the possible outcomes.

A random variable $X \sim \texttt{DiscreteUniform}(a, b)$ has the probability mass function

$$f_X(k) = \frac{1}{b-a+1},$$

where k can be any number between a and b. Note the sample space $\mathcal{X} = \{a, a+1, \ldots, b\}$ has a total of $n = b - a + 1$ elements. Note also that the probability of each outcome does not depend on the value of k—the same probability value is assigned to all outcomes in the sample space.

The cumulative distribution function that corresponds to the discrete uniform distribution looks like a step function, with $(b - a + 1)$ steps of height $\frac{1}{b-a+1}$ each. It starts at zero, jumps to $F_X(a) = \frac{1}{b-a+1}$ at $x = a$, then jumps by the same amount at every integer until $x = b$ where it reaches one the value one, $F_X(b) = 1$.

The mean and variance of a random variable $X \sim \texttt{DiscreteUniform}(a, b)$ are

$$\mu_X = \frac{a+b}{2} \qquad \text{and} \qquad \sigma_X^2 = \frac{(b-a+1)^2 - 1}{12}.$$

Intuitively speaking, The mean $\mu_X = \frac{a+b}{2}$ tells us the average value of X and is the centre of a and b. It's possible that the halfway point between a and b ends up being a fraction, and thus not part of the sample space. For example, the process of rolling six sided die can be represented as the random variable X with sample space $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$ and probability mass function $\texttt{DiscreteUniform}(1, 6)$.

The action of "pick a number at random between a and b" is one of the fundamental types of random phenomena and is implemented in most programming languages. Below are some relevant computer functions you should know about:

- In Excel, you can use the function `RANDBETWEEN(a, b)` to obtain a random number between a and b, inclusiverly.
- In R use `sample(a:b,1)` to choose a "sample" of size one from the list `a:b`.
- In Python, you can import the module called `random` using the command `import random`, then generate random numbers between a and b using `random.randint(a,b)`.

Relations to other distributions:

- The distribution $\texttt{DiscreteUniform}(0, 1)$ is identical to the distribution $\texttt{Bernoulli}(\frac{1}{2})$.

23.1.3 Binomial distribution

The binomial distribution models the number of successes in n consecutive draws from a Bernoulli distribution. Recall that the Bernoulli distribution with parameter p describes a random binary phenomenon where the probability of "success" is p and the probability of "failure" is $1 - p$. If we repeat this Bernoulli trial n times, then the *number of successes* is described by the binomial random variable $X \sim \texttt{Binomial}(n, p)$.

A random variable $X \sim \texttt{Binomial}(n, p)$ has the probability mass function

$$f_X(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

for k in the set $\{0, 1, \ldots, n\}$. The formula consists of the product of probabilities for k successes, $n - k$ failures, and the binomial coefficient $\binom{n}{k}$. This takes into account the number of ways the k successes can occur within the sequence of n trials. The name for this distribution comes from the binomial coefficient $\binom{n}{k}$ that appears in the formula.

The mean and the variance of the distribution $\texttt{Binomial}(n, p)$ are

$$\mu = \mathbb{E}_X[X] = np \quad \text{and} \quad \sigma^2 = \mathbb{V}[X] = \mathbb{E}_X[(X - \mu_X)^2] = np(1-p).$$

We can verify the above formulas using the fact that the Binomial distribution is the sum of the outcomes of n Bernoulli trials:

$$X = B_1 + B_2 + \cdots + B_n,$$

where $B_i \sim \texttt{Bernoulli}(p)$ are all draws from the Bernoulli distribution with parameter p, each of which have mean p and variance $p(1-p)$. Recall that the expectation of a sum of random variables equals the sum of the expectations: $\mathbb{E}[Y + Z] = \mathbb{E}[Y] + \mathbb{E}[Z]$. Applying this property to the equation that describes the Binomial random variable we get

$$\begin{aligned}
\mu &= \mathbb{E}_X[X] \\
&= \mathbb{E}_{B_1 B_2 \cdots B_n}[B_1 + B_2 + \cdots + B_n] \\
&= \mathbb{E}_{B_1}[B_1] + \mathbb{E}_{B_2}[B_2] + \cdots \mathbb{E}_{B_n}[B_n] \\
&= \mathbb{E}_B[B] + \mathbb{E}_B[B] + \cdots \mathbb{E}_B[B] \\
&= n\mathbb{E}_B[B] = np.
\end{aligned}$$

The variance of the distribution $\texttt{Binomial}(n, p)$ can similarly be computed:

$$\begin{aligned}
\sigma^2 &= \mathbb{V}[X] \\
&= \mathbb{V}_{B_1 B_2 \cdots B_n}[B_1 + B_2 + \cdots + B_n] \\
&= \mathbb{V}_{B_1}[B_1] + \mathbb{V}_{B_2}[B_2] + \cdots \mathbb{V}_{B_n}[B_n] \\
&= n\mathbb{V}_B[B] = np(1-p),
\end{aligned}$$

where in second equation we use the additive property of variance for the sum of two random variables $\mathbb{V}[Y+Z] = \mathbb{V}[Y] + \mathbb{V}[Z]$. See E23.1 and P23.1 for an alternative derivation of the equations for the mean that make use of calculus techniques.

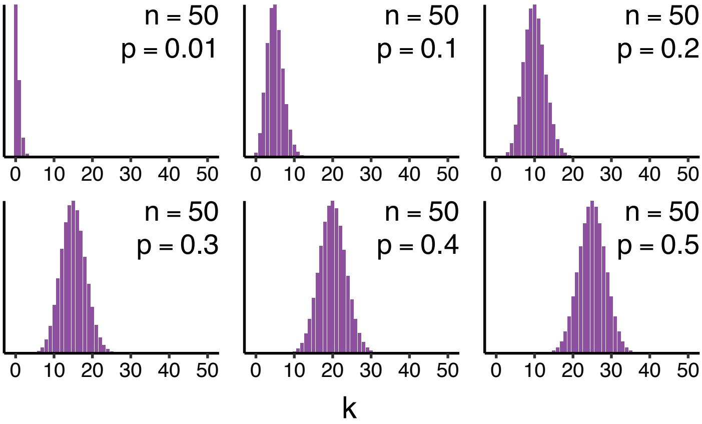

Figure 23.1: Plot of the probability mass function of the binomial distribution with $n = 100$ for different values of p.

Figure 23.1 shows the probability mass function for different values of p. Note the mean is np and the values get more spread out as p increases since $\sigma^2 = np(1-p)$.

Computer functions:

- In Excel use the `BINOM.DIST(k,n,p,FALSE)` to obtain the values from the pmf, and `BINOM.DIST(k,n,p,TRUE)` to obtain value of cdf.
- In R use `dbinom(k,n,p)` for the pmf and `cbinom(k,n,p)` for the cdf.
- In Python use `binom.pmf(k,n,p)` for the pmf and `binom.cdf(k,n,p)` for the cdf.

An example of a phenomenon that follows the binomial distribution is this: Imagine a bucket that contains n balls of which k are "success balls" and $n-k$ are "failure balls". We choose one of the n balls at random, record it's value, then put it back in the bucket. (This is called sampling with replacement.) The number of success outcomes will follow the binomial distribution with parameter $p = \frac{k}{n}$.

Relations to other distributions:

- The Bernoulli distribution is a special case of the binomial distribution with $n = 1$.

- The distribution $\texttt{Binomial}(\frac{k}{n}, n)$ describes the count of successes for n draws with replacement from an that contains a total of n balls of which k are "success balls." If the sampling is performed without replacement (choose one of the n balls at random, record it's value, put it away, then repeat the procedure by choose one of the remaining balls in the), then the number of successes is described by the hypergeometric distribution $\texttt{Hypergeometric}(N, n, K)$.
- If the size of the sample n is large ($n \geqslant 20$), the normal distribution $X' \sim \texttt{Normal}(\mu = np, \sigma^2 = np(1-p))$ can be used to approximate the binomial distribution $\texttt{Binomial}(n, p)$. This is known as the Moivre–Laplace approximation. We need to apply a *continuity correction* of 0.5 when using the normal approximation to the binomial. For example, if we're interested in computing $\Pr(X = 7)$ we compute $\Pr(6.5 \leqslant X' \leqslant 7.5)$.

23.1.4 Geometric

The geometric distribution describes the distribution of the waiting time until the first success in a series of independent Bernoulli trials, where each Bernoulli trial has probability of a success p. The probability mass function of a random variable $X \sim \texttt{Geometric}(p)$ is

$$f_X(k) = (1-p)^{k-1}p,$$

for k in the set $\{1,2,3,\ldots\}$. The formula consists of the product of $k-1$ failure probabilities and once success. Note the sample space is $\mathcal{X} = \mathbb{N}^+$, which is a countably infinite set. Indeed, there is no theoretical limit to the "bad luck" scenario in which the sequence of Bernoulli trials continues to result in failure. By definition, the trials must continue until the first success so the distribution is defined for all positive integers.

The mean and the variance of the distribution $\texttt{Geometric}(p)$ are

$$\mu = \frac{1}{p} \qquad \text{and} \qquad \sigma^2 = \frac{1-p}{p^2}.$$

See E23.2 and P23.2 for the derivations.

Note certain textbooks and computer programs use an alternative definition of the the geometric random variable Y in terms of the number of failures that occur before the first success. The probability mass function for this alternative formulation is $f_Y(k') = (1-p)^{k'}p$ and it has mean $\mu_Y = \frac{1-p}{p}$ and variance $\sigma_Y^2 = \frac{1-p}{p^2}$.

The values of the probability mass function decrease geometrically by a factor of $r = (1-p)$. Each subsequent trial, $f_X(k+1) =$

$(1-p)f_X(k)$. This is where the name "geometric" comes from. Recall the geometric sequence has the form $a_n = ar^n$, and its infinite series given by the formula $\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}$.

Computer functions:

- In Excel use the formula `POWER((1-p),k-1)*p` to obtain the values from the pmf.
- In R use `dgeom(k-1,p)` for the pmf and `pgeom(k-1,p)` for the cdf.
- In Python use `geom.pmf(k,p)` for the pmf and `geom.cdf(k,p)` for the cdf.

Relations to other distributions:

- If instead of stopping after the first success occurs, we continue counting until the first r successes occur, the waiting time will be described by the distribution `NegativeBinomial`(r, p).

Applications

You can model many situations with trials that are repeated until the first success occurs using the geometric distribution.

For example, if the probability of success for some difficult task is p, then $f_X(k)$ represents the probability of succeeding on the k^{th} attempt. Persistence is the key my friends!

Alternatively you can use the geometric distribution to compute the probability of failure after repeated successes. Suppose each time you turn on a light bulb it has a probability p of burning out, then you can $f_X(k)$ represents the probability of burning out on the k^{th} use, after $k-1$ successes.

In baseball you can model the probability of a batter with average hit probability p of getting a hit on one of the first three attempts. In business, you could model the number of interviews you'll need to perform in order to hire a competent candidate as a geometric distribution, assuming each hiring interview has probability of success p.

23.1.5 Negative binomial

The geometric distribution describes repeated Bernoulli trials until the first success outcome. The *negative binomial distribution* is a generalization of a geometric distribution where we wait to obtain r successes. The probability mass function a random variable $X \sim$ `NegativeBinomial`(r, p) is

$$f_X(k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r,$$

where p describes the probability of success, and k takes on values in the set $\{r, r+1, r+2, \ldots\}$. The minimum value of k is r you need to run at least r trials to obtain r successes. Note the last trial is necessarily a success so the coefficient $\binom{k-1}{r-1}$ counts for the number ways of choosing the remaining $r-1$ successes among the $k-1$ trials before the last.

The name of this distribution comes from the fact that we can rewrite $\binom{k-1}{r-1}$ as as an expression that involves the negative binomial coefficient $(-1)^m\binom{-r}{m}$, where $m = k - r$. See E23.4 for the calculation.

The mean and variance of the negative binomial distribution with parameters r and p are

$$\mu = E(X) = \frac{r}{p} \quad \text{and} \quad \sigma^2 = V(X) = \frac{r(1-p)}{p^2}.$$

See E23.3 and P23.3 for the derivations.

Computer functions:

- In Excel use the function `NEGBINOM.DIST(k-r,r,p,FALSE)` to obtain the values from the pmf and `NEGBINOM.DIST(k-r,r,p,TRUE)` to obtain the values from the cdf.
- In R use `dnbinom(k-r,r,p)` for the pmf and `pnbinom(k-r,r,p)` for the cdf.
- In Python use `nbinom.pmf(k-r,r,p)` for the pmf and `nbinom.cdf(k-r,r,p` for the cdf.

Note the computer functions above take the number of failures as their first argument, which corresponds to the difference $k-r$—if the r^{th} success occurs on the k^{th} trial, then there must have been $k-r$ failures before it.

Relations to other distributions:

- The geometrics distribution is a special case of the negative binomial with $r = 1$.

Applications

One possible real-world situation where the negative binomial distribution would be needed Consider a distributed storage system in which information is stored on multiples hosts for redundancy. When adding a file to the system, the software needs to store r copies of the file. Suppose the probability of connecting and completing the transfer to any one of the peers is given by p, then the probability of successfully publishing the file after k attempts is given the negative binomial distribution.

a server must When a server wants to publish a piece of content,

23.1.6 Hypergeometric

Consider a bucket that contains a total of n balls, of which k balls are labelled "success" and the remaining $n - k$ balls are labeled as "failure." We choose a sample of n balls randomly and count how many of them are successes. If the sampling is performed without replacement, meaning we choose one of the balls at random, record it's value then put it back in the bucket, then the probability distribution for the number of successes is described by the hypergeometric distribution `Hypergeometric`(N, n, K).

The number of successes in the sampling scenario described above is described by a random variable $X \sim$ `Hypergeometric`(N, n, K) with probability mass function

$$f_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

where k lies between $\max\{0, n + K - N\}$ and $\min\{K, n\}$.

The mean and the variance of the hypergeometric distribution with parameters n, N, and k are

$$\mu_X = n \cdot \tfrac{k}{N} \quad \text{and} \quad \sigma_X^2 = np(1-p)\frac{N-n}{N-1}.$$

where we've defined $p = K/N$. See E23.5 and P23.4 for the derivations.

Compare the equations for the mean and the variance of the hypergeometric with the equations for the mean and the variance of the Binomial distribution (see page 86). The resemblance is not a coincidence. Indeed, the binomial and hypergeometric distributions describe the same model, but the binomial distribution describes sampling with replacement while the hypergeometric distribution describes sampling without replacement. When k and n are large numbers, the effect of sampling without replacement becomes negligible and we can approximate the hypergeometric distribution using a binomial distribution.

Computer functions:

- In Excel use the function `HYPGEOM.DIST(k,n,K,N,FALSE)` to obtain the k^{th} value from the pmf, and `HYPGEOM.DIST(k,n,K,N,TRUE)` to obtain the values from the cdf.
- In R use `dhyper.pmf(k,K,N-K,n)` for the pmf and `phyper.pmf(k,K,N-K,n` for the cdf.
- In Python use `hypergeom.pmf(k,N,n,K)` to get the values of the probability mass function and `hypergeom.cdf(k,N,n,K)` for the values of the cdf.

Relations to other distributions:

- If the draws from the are performed *with* replacements, the number of successes is described by the binomial distribution $\texttt{Binomial}(\frac{K}{N}, n)$.

Example Suppose you have bag containing a total of $N = 7$ tomatoes of which $K = 3$ are good tomatoes and $N - K = 4$ are rotten. You want to choose two tomatoes from this bag to make a salad. What is the probability you will end up with zero, one, and two good tomatoes? The situation is described by the random variable $X \sim \texttt{Hypergeometric}(N = 7, n = 2, K = 3)$ whose probability mass function is $f_X(k) = \frac{1}{\binom{7}{2}} \binom{3}{k} \binom{4}{2-k}$. Intuitively, the distribution counts the number of ways to choose k good tomatoes from the three good ones, times the number of ways to choose the remaining $2 - k$ from the bad ones, and normalization factor describes all possible ways to choose two tomatoes from a bag of seven.

We can compute the probabilities of the three different outcomes by hand. The probability of picking zero good tomatoes is given by $p_X(0) = \frac{4}{7} \cdot \frac{3}{6} = 0.2857$, where $\frac{4}{7}$ is the probability of picking a bad tomato on the first draw, and $\frac{3}{6}$ is the probability of picking a bad tomato on the second draw. There are two possible ways to pick one good tomato in a draw of two: the first one or the second one, so the probability is $p_X(1) = \frac{4}{7} \cdot \frac{3}{6} + \frac{3}{7} \cdot \frac{4}{6} = 0.5714$. The probability of picking two good tomatoes is $p_X(2) = \frac{3}{7} \cdot \frac{2}{6} = 0.1429$.

Use one of the computer functions in Excel, R, or Python to independently compute the values of $p_X(0)$, $p_X(1)$, and $p_X(2)$ stated above and verify they are correct.

23.1.7 Poisson

The Poisson distribution is popular for modelling the number of times an event occurs in an interval of time or space.

Assumptions: When is the Poisson distribution an appropriate model? The Poisson distribution is an appropriate model if the following assumptions are true.

k is the number of times an event occurs in an interval and k can take values 0, 1, 2, ... The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently. The rate at which events occur is constant. The rate cannot be higher in some intervals and lower in other intervals. Two events cannot occur at exactly the same instant; instead, at each very small sub-interval exactly one event either occurs or does not occur. Or The actual probability distribution is given by a binomial distribution and the number of trials is sufficiently bigger than the number of successes one is asking about (see Related distributions). If these conditions are true, then k is a Poisson random variable, and the distribution of k is a Poisson distribution.

The Poisson distribution describes the number of random events that occur during some period of time. We assume the events occur independently of each other at a constant average rate of λ. The probability mass function for the random variable $X \sim \texttt{Poisson}(\lambda)$ is

$$f_X(k) = \frac{(\lambda)^k e^{-\lambda}}{k!},$$

where $k \geqslant 0$. The parameter $\lambda > 0$ describes the average probability of the event during the chosen time period.

The mean and variance for a random variable $X \sim \texttt{Poisson}(\lambda)$ are

$$\mu_X = \mathbb{E}[X] = \lambda \quad \text{and} \quad \sigma_X^2 = \lambda.$$

See E23.6 and P23.5 for the calculations.

The parameter λ depends on the length T of the time intervals indexed by the random variable X. We can compute λ by multiplying the average rate r for the events to occur, and the time period: $\lambda = rT$. For example, if we know some event occurs three times per hour on average $r = 3/hour$, and we want to know the probability distribution for the total number in a given day, the we choose $\lambda = 3 \times 24 = 72/day$.

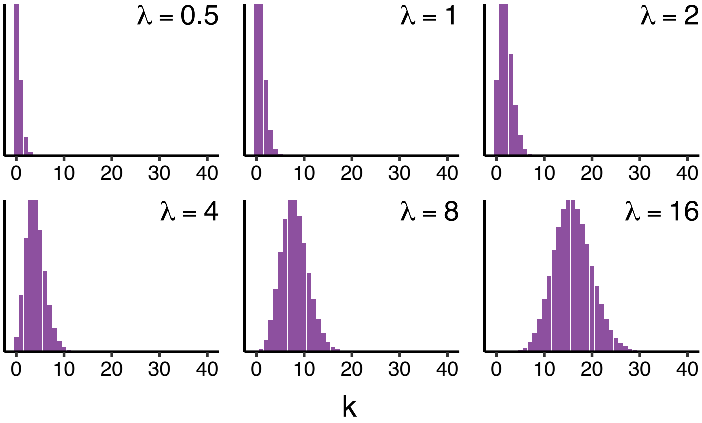

Figure 23.2: Histogram showing $n = 20$ draws from the Poisson distribution with for different values of λ.

Figure 23.2 shows the probability mass function of the Poisson distribution for different values of λ. As λ increases the mean of the distributions shifts to larger values and also becomes more spread out.

Computer functions:

- In Excel use the function `POISSON(k,lambda,FALSE)` to obtain the values from the pmf and `POISSON(k,lambda,TRUE)` to obtain the values from the cdf.
- In R use `dpois(k,lambda)` for the pmf and `ppois(k,lambda)` for the cdf.
- In Python use `poisson.pmf(k,lambda)` for the pmf and `poisson.cdf(k,lambda)` for the cdf.

Relations to other distributions:

- alskj

Relation to the binomial distribution

We can obtain the Poisson distribution from the binomial distribution by taking the limit $n \to \infty$ and $p \to 0$.

Consider a random variable X

The construction of this distribution is for a random variable X as the number of "successes" in the interval $[0, T]$, which has length T. We assume, a priori, that the number of successes per unit length has an average of r so the expected value of successes over the time period of length T is $\mathbb{E}[X] = rT$, which we'll give a new name λ.

We know from the formula for expected value of a binomial distribution should also be equal to np. Thus, we may solve for p as

$$np = \lambda \qquad \Rightarrow p = \frac{\lambda}{n}.$$

Now, the probability that there are k successes given n trials is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

Now, we seek to take the limit as $n \to \infty$. That is

$$\begin{aligned} P(X = k) &= \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \lim_{n\to\infty} \underbrace{\frac{n}{n}}_{=1} \cdot \underbrace{\frac{n-1}{n}}_{\to 1} \cdots \underbrace{\frac{n-k+1}{n}}_{\to 1} \frac{(\lambda)^k}{k!} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\to e^{-\lambda}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{\to 1} \\ &= \frac{(\lambda)^k e^{-\lambda}}{k!} \end{aligned}$$

Geometrically, taking the limit n to infinity represents dividing the interval $[0, T]$ into n equal segments where, as n increases, the number of multiple successes in each equal segment necessarily becomes zero. In general, given the partition of $[0, T]$ into n small subintervals (tending to zero), we have

- the probability of more than one success in a sub interval tends to 0
- the probability of one success in a subinterval tends to $\frac{\lambda}{n}$
- events in each subinterval are independent of each other.

Applications

The Poisson distribution is an important model for many phenomena we can observe in the real world:

- The number of earthquakes occurring in a fixed period of time.
- The number of phone calls arriving at a particular point in a telephone network in a fixed time period.
- The number of visitors to a website per minute.
- The number of customers arriving at a ticket window.

The commonality is that all these phenomena is that we're counting the total number of statistically independent events that occur at a constant rate λ (the expected number of the events per unit time).

For the multinomial distribution see `dmultinom`.

23.2 Plotting distributions

commands to plot pmf/pdf and cmf functions in Excel, R, and Python

mention useful for understanding parameters — hands on

23.3 Modelling real-world data using probability

Let's try to connect probability distributions we discussed in this chapter with the data sets that we used as examples in the data chapters. I want you to see that probability modelling skills you developed can help you better understand datasets.

We'll plot the histogram of the data and and a plot of the pmf in the same graph, then tweak the model parameters interactively until the two curves start to look the same.

Student grades data

Recall the students' grades data set we covered in Part I, let's try to find the parameters μ and σ that describe the shape of the data.

Hard disk failure rates

number of failures in a data centre

RECIPE: - repeat definitions - show mini plots with different values of the parameters - plot histogram of dataset - show steps: - choice of distribution - choose mean to match data (by eyeballing and trial and error) - choose variance to match data (by eyeballing and trial and error)

Assume five examples of modelling real-world distributions by "eyeballing" the parameters. Drill and repeat. No theory, only visual explanations of fitting parameters to match real-world datasets.

23.4 Exercises

E23.1 Compute the mean of the random variable $X \sim$ Binomial(n, p) whose distribution is $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$, for $k \in \{0, 1, \ldots, n\}$.

E23.2 Compute the mean of the random variable $X \sim$ Geometric(p) with probability mass function $p_X(k) = (1-p)^{k-1}p$, for $k \in \{1, 2, \ldots\}$.

Hint: You can use the formula $\sum_{k=1}^{\infty} kar^{k-1} = -\frac{a}{(1-r)^2}$, which is obtained by taking the derivative of $\sum_{n=0}^{\infty} ar^n = \frac{a}{1-r}$ with respect to r.

E23.3 Compute the mean of $X \sim$ NegativeBinomial(r, p) whose distribution is $p_X(k) = \binom{k-1}{r-1}(1-p)^{k-r}p^r$, for $k \in \{r, r+1, r+2, \ldots\}$.

E23.4 Show that $\binom{k-1}{r-1}$ equals $(-1)^m \binom{-r}{m}$ where $m = k - r$.

Hint: Expand both expressions separately to show they are equal.

E23.5 Find the mean of the random variable $X \sim$ Hypergeometric(n, K, N) with probability mass function $f_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$, where k is between $\max\{0, n+K-N\}$ and $\min\{K, n\}$.

E23.6 Compute the mean of the random variable $X \sim$ Poisson(λ) with probability mass function $p_X(k) = \frac{(\lambda)^k e^{-\lambda}}{k!}$, for $k \in \{0, 1, 2, \ldots\}$.

E23.7 Use the probability functions `dpois(k,lambda)`, `ppois(k,lambda)`, and `qpois(q, lambda)` to reproduce the calculations of the the number-of-hard-disk-failures scenario from page 22.4.1. Compute **a)** the probability of exactly 20, 21, and 22 failures, **b)** the probability of the event $\{16 \leqslant Z \leqslant 24\}$, and **c)** the 95 percentile $F_Z^{-1}(0.95)$ for the random variable $Z \sim \text{Poisson}(\lambda = 20)$.

E23.8 Use the probability functions in Excel to reproduce the calculations of the the number-of-hard-disk-failures scenario from page 22.4.1. Compute **a)** the probability of exactly 20, 21, and 22 failures, **b)** the probability of the event $\{16 \leqslant Z \leqslant 24\}$, and **c)** the 95 percentile $F_Z^{-1}(0.95)$ for the random variable $Z \sim \text{Poisson}(\lambda = 20)$.

23.5 Chapter summary

You should now be able to

1. Determine probabilities from mass functions and vice versa.
2. Determine probabilities and mass functions from cumulative distributions and vice versa.
3. Compute mean and variance of discrete random variable.
4. Understand assumptions for common discrete distributions.
5. Select appropriate distribution in specific applications.
6. Compute probabilities and determine mean and variance of common discrete distributions.

23.6 Discrete distributions problems

Intro/motivational text...

P23.1 Compute the variance of the random variable $X \sim \text{Binomial}(n, p)$ whose distribution is $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$, for $k \in \{0, 1, \ldots, n\}$.

Hint: Start from the formula $\sigma^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, then add and subtract $\mathbb{E}[X]$ and rewrite as $\sigma^2 = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2$.

P23.2 Compute the variance of the random variable $X \sim \text{Geometric}(p)$ with probability mass function $p_X(k) = (1-p)^{k-1} p$, for $k \in \{1, 2, \ldots\}$.

Hint: Start from the formula $\sigma^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

P23.3 Compute the variance of $X \sim \text{NegativeBinomial}(r, p)$ whose distribution is $p_X(k) = \binom{k-1}{r-1}(1-p)^{k-r} p^r$, for $k \in \{r, r+1, r+2, \ldots\}$.

Hint: Use the formula $\sigma^2 = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2$ as in P23.1.

P23.4 Compute the variance of the random variable $X \sim \text{Hypergeometric}(n, K, N)$ with probability mass function $f_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$, for

Hint: Use the formula $\sigma^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

P23.5 Compute the variance of the random variable $X \sim \text{Poisson}(\lambda)$ with probability mass function $p_X(k) = \frac{(\lambda)^k e^{-\lambda}}{k!}$, for $k \in \{0, 1, 2, \ldots\}$.

Hint: Use the formula $\sigma^2 = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2$.

P23.6 Suppose $Z = \sum_{i=1}^{n} a_i Z_i$ is a linear combination of independent random variables each having means μ_i. Show that $\mathbb{E}[Z] = \sum_{i=1} a_i \mathbb{E}[Z_i]$ and $\mathbb{V}[Z] = \sum_{i=1}^{n} a_i^2 \mathbb{V}[Z_i]$.

Chapter 24

Continuous probability distributions

By the end of this chapter, you should now be able to

1. Determining probabilities for events given a probability density function
2. Determining pdf from the cdf and and vice-versa
3. Calculating means and variances for continuous random variables
4. Understanding the assumptions for some common continuous probability distributions
5. Selecting an appropriate continuous distribution model for specific applications
6. Standardizing normal random variables
7. Using the table for the cumulative distribution function of a standard normal distribution to calculate probabilities

The sample space of the a *continuous random variable* is some interval of the real numbers. Instead of the random events taking on

24.1 Probability density functions

A continuous random variable X is described by a *probability density function* $f_X(x)$ that satisfies

- $f_X(x) \geqslant 0$ for all $x \in X$
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$

The probability of the random variable falling between $\{a \leqslant X \leqslant b\}$ is obtained by computing the integral of the probability density function between a and b:

$$\Pr\left(\{a \leqslant X \leqslant b\}\right) = \int_a^b f_X(x)dx.$$

The analogy with the probability mass function of discrete

Due to the properties of integration, the limits of

$$\Pr(a \leqslant X \leqslant b) = \Pr(a < X \leqslant b) = \Pr(a \leqslant X < b) = \Pr(a < X < b).$$

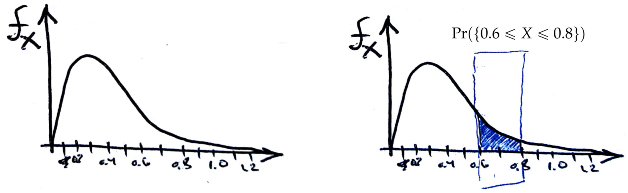

Figure 24.1: Illustration of the probability density function f_X for some random variable X. The area highlighted in the left half of the figure shows the probability of the event $\{0.6 \leqslant X \leqslant 0.8\}$, which is computed as the integral $\Pr(\{0.6 \leqslant X \leqslant 0.8\}) = \int_{x=0.6}^{x=0.8} f_X(x)\,dx$.

The *cumulative distribution function* of a random variable X is defined as

$$F(x) = \Pr(X \leqslant x) = \int_{-\infty}^{x} f(t)dt.$$

It follows that $\Pr(X > x) = 1 - F(x)$.
The probability density can be obtained from the cumulative distribution using differentiation

$$f_X(x) = \frac{dF(x)}{dx}.$$

The mean and variance of a continuous random variable are given by

$$\mu = \mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx,$$

and

$$\sigma^2 = V[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2.$$

Recall that probability theory concept is inspired by physics. Similar to the case for discrete distributions, we imagine the probability density function to represent the weight density of some one-dimensional solid object. The calculation for the mean of a probability distribution μ corresponds to the physics calculation of the centre of mass of the object x_{cm}. The centre of mass of an object is the point along the length of the object where it is balanced. Half the mass of the object is to the left and half is to the right of x_{cm}, so if you place your finger at x_{cm} you can balance the objet using a single point contact.

The variance of a probability distribution σ^2 corresponds to the moment of inertia of the solid object I. For those of you who don't remember physics, the *moment of inertia* of an object tells you how difficult it is to make the object turn around it's centre of mass. The quantity I plays the same role in angular motion as the mass m plays in linear motion, and it appears in the formula for the torque, angular momentum, and angular kinetic energy of an object. Roughly speaking, the more the weight of an object is spread out, the more difficult it will be to make it turn, and the relationship is nonlinear. The contribution to the moment of inertia of a piece of mass dm at a distance r from the centre of rotation is proportional to the *square* of the distance $dI = r^2\,dm$. In other words, if you want an object that is easy to turn—easy to set into rotational motion—then should put all the mass of the object near the centre of rotation. This same notion of "squared distance from the centre" turns out to be useful for describing probability distributions.

More generally, the expected value of any function $h(X)$ in our random variable is determined as

$$E[h(X)] = \int_{-\infty}^{\infty} h(x)f(x)dx.$$

24.2 Mathematical prerequisites

As you just saw, the derivative and integral operation are important for continuous random variables, so it's worth reviewing the concepts right now.

24.3 Continuous distributions reference

The following computer models for continuous distributions are defined in the module `scipy.stats.distributions`: `uniform`, `norm`, `gamma`, `expon`, `t`. We'll describe these distributions in more detail in Section 24.3.

For the uniform distribution see dunif. For the normal distribution see dnorm. For the log-normal distribution see dlnorm. For the beta distribution see dbeta.

For the exponential distribution see dexp.

For the gamma distribution see dgamma.

For the F distribution see df. For the chi-squared distribution see dchisq. For the Cauchy distribution see dcauchy. For the Weibull distribution see dweibull.

24.3.1 Uniform

The *continuous uniform distribution* on $[a, b]$ is

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

Its mean and variance are

$$\mu = \frac{a+b}{2} \quad \text{and } \sigma^2 = \frac{(b-a)^2}{12}$$

respectively. Indeed, assuming we have already determined the mean (exercise!), the variance is found as

$$\begin{aligned}
\sigma^2 &= \int_a^b \frac{x^2}{b-a}dx - \frac{(a+b)^2}{4} \\
&= \frac{b^3-a^3}{3(b-a)} - \frac{a^2+2ab+b^2}{4} \\
&= \frac{b^2+ab+a^2}{3} - \frac{a^2+2ab+b^2}{4} \\
&= \frac{4(a^2+ab+b^2)-3(a^2+2ab+b^2)}{12} = \frac{(b-a)^2}{12}.
\end{aligned}$$

```
a + (b-a)*RAND()
```

24.3.2 Exponential

Along with knowing the number of "successes" on an interval (whose probability is measured via the Poisson distribution) come a different random variable measuring the distance between successes. The link here is that distance between two successes is greater than x if and only if the number of success is 0 on the interval $[0, x]$. That is if D represents the distance between successes and N the number of successes on on the interval $[0, x]$ then

$$\Pr(D > x) = \Pr(N = 0) = \frac{e^{-\lambda x}(\lambda x)^0}{0!} = e^{-\lambda x}.$$

The cumulative distribution here is

$$F(x) = \Pr(D \leqslant x) = 1 - e^{-\lambda x}$$

whose derivative, and hence the probability mass function, is then

$$f(x) = F'(x) = -e^{-\lambda x}$$

for all $x > 0$.

The mean and variance are both calculated via integration by parts as

$$\mu = E[D] = \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

and

$$\sigma^2 = V[D] = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = \frac{1}{\lambda^2}$$

Memoryless property

The exponential distribution is *memoryless*, which comes from the conditional probability argument as follows:

$$\Pr(D < t + \epsilon | D > t) = \Pr(t < D < t + \epsilon) / \Pr(D > t)$$

with

$$\begin{aligned} \Pr(t < D < t + \epsilon) &= F(t + \epsilon) - F(t) \\ &= 1 - e^{-\lambda(t+\epsilon)} - (1 - e^{-\lambda t}) \\ &= e^{-\lambda t}(1 - e^{-\lambda \epsilon}) \end{aligned}$$

and

$$\Pr(D > t) = e^{-\lambda x}$$

so that

$$\Pr(D < t + \epsilon | D > t) = 1 - e^{-\lambda \epsilon} = \Pr(D \leqslant \epsilon).$$

More concretely this says, that the probability that a success will occur in the next ϵ given that it has not already in the first t is the same as the probability of success in the first epsilon.

For example, in the waiting room at the doctor's office with the time waited until your turn as the random variable. If this scenario is modelled by an exponential distribution then this property says that the probability it will be your turn in the next 30 seconds given that you have already waited five minutes is the same as if you had just arrived.

The *memoryless property* is expressed as

$$\Pr(X < t + \epsilon | X > t) = \Pr(X \leqslant \epsilon)$$

or

$$\Pr(X > t + \epsilon | X > t) = \Pr(X > \epsilon)$$

and the exponential distribution is the only continuous distribution with this property. In fact, the exponential distribution can be defined as the continuous random variable satisfying the memoryless property.

24.3.3 Normal

The *normal distribution* on $\mathbb{R}$ is given by

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

having mean μ and variance σ. Notation for this extremely common distribution is $N(\mu, \sigma^2)$.

24.3.4 The standard normal distribution

The *standard normal distribution* is $\mathcal{N}(0,1)$ and the cumulative distribution function (CDF) for $\mathcal{N}(0,1)$ is denoted $F_Z(z) \equiv \Pr(Z \leqslant z)$. All normal distributions have essentially the same "shape." Consider the random variable X which is normally distributed with mean μ and variance σ^2:

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

The probability density function for X has the same shape as the standard normal $Z \sim \mathcal{N}(0,1)$, which has mean 0 and variance 1. The two random variables are related by the following equation:

$$Z = \frac{X - \mu}{\sigma},$$

which involves subtracting the mean and dividing by the standard deviation. Every gaussian random variable can be transformed to the standard normal distribution using this transformation.

Probability calculations

For every calculation you might want to do with the random variable X, there is an equivalent calculation you can carry out using the random variable Z:

$$F_X(a) = \Pr\big(X \leqslant a\big) = \Pr\left(Z \leqslant \frac{a - \mu_X}{\sigma_X}\right) = F_Z\left(\frac{a - \mu_X}{\sigma_X}\right),$$

where $F_X(a) = \int_{-\infty}^{a} f_X(x)\ dx$ is cumulative distribution function (CDF) of the random variable X, and F_Z is the CDF of the standard normal. This means it suffices to know the values of the CDF for the standard normal distribution F_Z, where $Z \sim \mathcal{N}(0,1)$, and the calculations for all other normal distributions can be obtained after a suitable transformation.

The function $F_Z : \mathbb{R} \to [0,1]$ can be used in two directions. Either we have a given value of z and we want to calculate $F_Z(z)$ (the cumulative probability of the random variable Z taking on this or any smaller value), or we start with some probability value q and want to compute the corresponding z-value z_q such that $F(z_q) = q$, in other words $z_q \equiv F^{-1}(q)$.

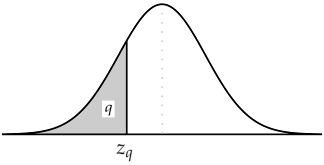

Figure 24.2: Illustration of the cumulative probability density calculations of the standard normal. The value z_q is such that a total probability of q is enclosed in the left tail of the distribution.

Numerical calculations

The values of the CDF of the standard normal distribution ($z_q \equiv F^{-1}(q)$) are used very often in statistics, so it is a good idea for you to know them really well. If you look at Table XX in Appendix YY you'll see the values of z_q for different values of q. From now on, when you see z_q written somewhere in the text, know that this refers to the value of the inverse CDF for the standard normal. You can compute it by looking up the value in Table XX, or using the formula z_q `=NORM.INV(q,0,1)` in spreadsheet software, or calling z_q =?? in R or z_q `=norm.ppf(q,0,1)` in Python.

24.3.5 Student's *t* distribution

Similar to normal but heavier tails. Used in statistical analysis of samples where the population variance is estimated from the sample variance.

TODO: define ν degrees of freedom

TODO: plots for different values of ν

The probability density of Student's t distribution with ν degrees of freedom is $f_T(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$ and its cumulative distribution function is $F_{T,\nu}(t) = \int_{-\infty}^{t} f_{\chi^2,\nu}(x)\,dx$.

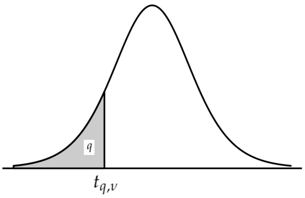

Figure 24.3: Illustration of the cumulative probability density calculations of the t distribution with ν degrees of freedom. The value $t_{q,\nu}$ is such that a total probability of q is enclosed in the left tail of the distribution.

You can lookup values of the inverse CDF of the different t distributions in Table WW in Appendix YY. You can also use the formula $t_{q,\texttt{df}}$ `=T.INV(q, df)` in Excel, or calling $t_{q,\texttt{df}}$ `=??` in R, or call $t_{q,\texttt{df}}$ `=t.ppf(q, df)` in Python.

24.3.6 The χ^2 distribution

The probability density of the χ^2 distribution with ν degrees of freedom is $f_{\chi^2,\nu}(x) = \frac{x^{\frac{k}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{k}{2}}\,\Gamma(\nu/2)}$ and its cumulative distribution function is $F_{\chi^2,\nu}(\chi^2) = \int_{-\infty}^{\chi^2} f_{\chi^2,\nu}(x)\,dx$.

The shape of the distribution is determined by the degrees of freedom parameter ν. Figure 24.4 illustrates three plots for different values of ν.

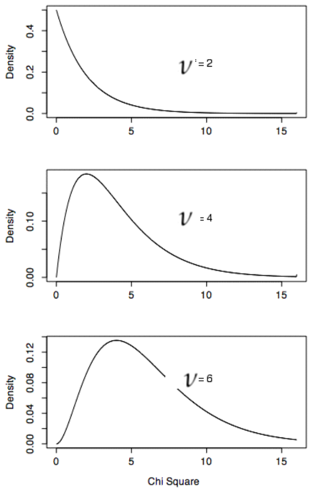

Figure 24.4: The χ^2_ν distribution for the three different values of ν. Note as ν gets larger the peak of the distribution moves to the right.

The mean of a Chi Square distribution is its degrees of freedom. Chi Square distributions are positively skewed, with the degree of skew decreasing with increasing degrees of freedom.

You can lookup values of the inverse CDF of the different χ^2 distributions in Table ZZ in Appendix YY. You can also use the formula $\chi^2_{q,\mathtt{df}}$ `=CHISQ.INV(q, df)` in Excel, or calling $\chi^2_{q,\mathtt{df}}$ `=??` in R, or call $\chi^2_{q,\mathtt{df}}$ `=chi2.ppf(q,df)` in Python.

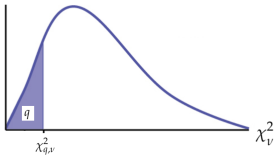

Figure 24.5: Illustration of the probability calculations for χ^2 distribution with ν degrees of freedom. The value $\chi^2_{q,\nu}$ is such that a total probability of q is enclosed in the left tail of the distribution.

24.4 Discussion

24.4.1 Other distributions

The *Erlang, Gamma, Weibull, Lognormal and Beta* distributions are beyond the scope of this course.

24.4.2 Relationships between functions

TODO: insert simplified concept map from `http://www.stat.rice.edu/~dobelman/courses/texts/leemis.distributions.2008amstat.pdf#page=3` or `https://pdfs.semanticscholar.org/c0db/71a4101347404d69`
`pdf#page=2`

Normal approximation to the binomial distribution

If X is a binomial random variable with parameters n and p, then

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

is approximately equal to the standard normal. To approximate a binomial probability with a normal distribution, a continuity correction is given by

$$\Pr(X = k) = \Pr(k - 0.5 \leqslant X \leqslant k + 0.5) \approx \Pr\left(\frac{k - 0.5 - np}{\sqrt{np(1-p)}} \leqslant Z \leqslant \frac{k + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

$$\Pr(X \leqslant k) = \Pr(X \leqslant k + 0.5) \approx \Pr\left(Z \leqslant \frac{k + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

$$\Pr(X \geqslant k) = \Pr(X \leqslant k - 0.5) \approx \Pr\left(Z \geqslant \frac{k - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

This approximation is good for $np > 5$ and $n(1-p) > 5$.

Normal approximation to the Poisson distribution

If X is a Poisson random variable with $E[X] = \lambda$ and $V[X] = \lambda$, then

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

is approximately a standard normal random variable. The same continuity correction used for the binomial distribution can also be applied. The approximation is good for $\lambda > 5$.

TODO: give other examples of phenomena to show normality emerge for large n

http://efavdb.com/normal-distributions/

24.4.3 Limiting behaviour

LLN = FWD reference to Extra Topics

Central limit theorem = FWD reference to Extra Topics

24.5 Exercises

24.6 Continuous distributions problems

Intro/motivational text...

Part III

Statistics