Modeling Mindsets

The Many Cultures of Learning From Data

Christoph Molnar

2022-06-24

Modeling Mindsets for Data Scientists

The Many Cultures of Learning From Data

christophmolnar.com

2022, First Edition

Christoph Molnar, c/o Mucbook Clubhouse, Westendstraße 78, 80339 München, Germany

Contents

Summary

We use data to advance science, make businesses more profitable, automate annoying tasks, and develop smart products. But there is a middleman between data and its usefulness: the **model**. The model represents a simplified aspect of the world; it's the glue that connects data and world.

Statistics versus machine learning, frequentist versus Bayesian inference, causation or association, … There are many mindsets to consider for building models from data. Each of these modeling mindsets has its own assumptions, strengths, and limitations.

The best modelers, researchers, and data scientists don't stubbornly stick to just one mindset. The best modelers mix and match the mindsets.

It can take years to truly grasp a new mindset. Most books and courses jump right into math and methods instead of discussing the fundamental mindset. But learning a new mindset doesn't have to be this difficult. **The Modeling Mindset book introduces many cultures of learning models from data**. Each of them enhances your own mind and makes you a better modeler:

- Frequentist inference: learning about nature's "true" values.
- Bayesian inference: updating your beliefs about the world.
- Supervised machine learning: predicting new data well.
- Causal inference: taking causality seriously.
- Deep learning: embedding the world into a neural network.
- And many more.

Modeling Mindsets opens the door to all these different ways of thinking. The book is packed with **intuitive explanations and illustrations**. Quickly get an overview of the strengths and the limitations of each modeling mindset. Expand your mind when it comes to modeling the world using data.

What This Book is About

The book is about all the different mindsets that allow you to model the world with data. Each mindset represents a different perspective on how to see the world through data. In this book, you will find for each mindset the **assumptions, central ideas, their relationship to other mindsets, and their strengths and limitations**. Modeling mindsets is not about history. Modeling mindsets is a mixture of lightweight methodological introduction and philosophy. That said, this book is not and **cannot be a full introduction to each mindset**. There are entire books about, for example, Bayesian inference, or supervised machine learning. After reading this book, you will not automatically become a frequentist statistician, or a causal inference expert. Sorry to disappoint this early in the book. However, reading Modeling Mindsets can open doors to new ways of thinking about modeling. But there are other resources to explore what's behind each door – an online course on machine learning, blog posts about about causal inference, a book about design-based inference, … In each chapter, I refer to **useful resources to deepen the particular mindset**.

1 Who This Book is For

The book is for data scientists, statisticians, machine learner, quantitative researchers, … In short, for anyone who already has experience with modeling data. This means you should probably **know at least one of the mindsets**. Perhaps, like me, you studied frequentist statistics. Or you may be a researcher who has learned to use Bayesian inference to analyze your data. Or maybe you are a self-taught machine learner.

That said, it's crucial that you don't cling to the mindset you already know. Let go of the rigid assumptions you've learned. Open your mind to fundamentally new ways of modeling data.

A little math shouldn't scare you either. But I can promise you that the book is not too heavy on the math side.

1 Models

You gaze at the screen. The screen shows a table with some data. Based on this data, you are to answer some questions. These questions could be:

- Which patients might get side effects from a certain drug?
- How do bee hives react to a change in climate?
- Which supermarket products are often out-of-stock?

In the data you can see in detail what happened: patients with ID 124 and 22 got acne; 2/3 of bee colonies had trouble during drought in 2018; on that one Tuesday the flour was sold out; But with data, you can't immediately see general rules and relationships. Is flour generally in low supply at the beginning of the week? It would be even better if these rules and relationships applied not only to your specific data sample, but to a more general population of patients/hives/supermarkets. To move from data to generalizable relationships, we have to simplify the world and make assumptions. The end result is a model of the world based on data.

A model is a simplified representation of some aspect of the world. For example, how bee colonies depend on climate. With a model we can answer questions and make predictions that we couldn't with the raw data.

In this book, we talk about certain types of models: Models must be computational or mathematical models. This excludes, for example, physical models, like the tiny houses that architects build. The second restriction: The models are learned from data. This excludes "designed" models such cellular automata.

There is no philosophical consensus on what makes a model. For our purpose, let's say that **a mathematical model consists of three ingredients: variables, relations and parameters**. A mathematical model contains mathematical structures that represent *variables* and put them in *relation* (Figure 1.1).[1] The relations are often expressed as parameterized functions of the variables. The model **parameters** make the mathematical structure adaptable. When the model is learned from data, in the learning process the parameters and sometimes relations (functions) are adapted to the data. If you want to interpret models instead of the world, you have to make assumptions about the relationship between the model and aspects of the world. But more about this in the chapter Mindsets.

The aspects of the world are represented within the model as *variables*. The blood pressure of a patient is represented with a numerical value. Images, for example, are represented as tensors of pixels. Variables can also represent a latent, hidden or abstract aspect. Like happiness or introversion. There are different names for variables: Random variables, covariates, predictors, latent variables, features, target, outcome, … These names sometimes reveal the role of a

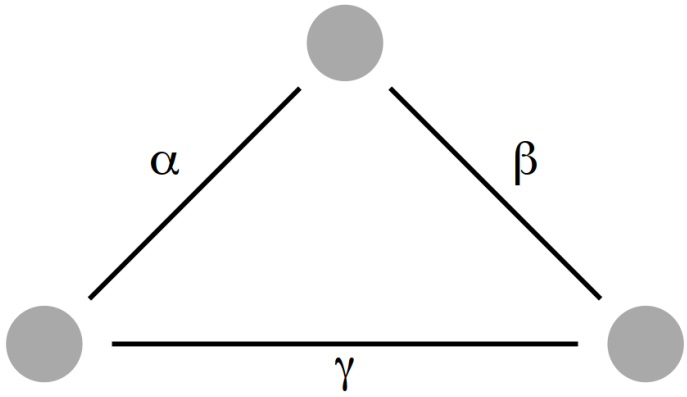

Figure 1.1: A mathematical model sets variables (dots) into relation (lines) using parameters.

variable in the model. For example, the "target" is the variable that we want to predict. In different mindsets, variables have different names and roles: In machine learning, for example, the terms feature and target are used. In statistics, people might say dependent and independent variable or covariates and outcome instead.

Within the model, the components are mathematically or computationally set in *relation* to each other. These relations are usually expressed as parameterized functions. For example:

- Causal models represent relations between variables in a directed acyclic graph that can be translated into conditional (in-)dependencies.
- The joint distribution of variables describes the occurrence of certain variable values in a probabilist sense.
- A predictive model represents the output variable as a parameterized function of the input variables.
- In the case of a linear regression model, the output variable is a weighted sum of the input variables.

The expressive power of such relationships really depends on the class of the model. A relation can be a simple linear equation like $Y = 5 \cdot X$ involving two or more variables. For example we might model the probability of a stroke as a function of blood pressure and age. A relation can also be a long chain of mathematical operations involving thousands of variables. Deep neural networks are an example of such a complicated relation.

We don't know the relations between variables in advance, so we use data to learn them. For some models, learning these relationships is a matter of finding the right *parameters*. This is true for neural networks and generalized additive models, for example. For other models, the model structure is "grown", as in decision trees or support vector machines. Growing the structure means not only learning parameters, but also learning the structure of the mathematical function.

You can think of a model as having an uninstantiated state and an instantiated state. An uninstantiated model is not yet fitted to the data. Uninstantiated models form families of models. For example the neural network ResNet architecture, or the family of Poisson regression models. An instantiated model is trained / learned / fitted using data: It's parameterized and/or the structure has been learned.

I can buy carrots with money. How many grams of carrots can I get for 1 euro? Let's call this unknown parameter in our equation β: 1 EUR $= \beta$ Carrots. I could figure out the β by going to the supermarket and checking the price. Maybe $\beta = 500$, so I get half a kilogram of carrots for 1 euro. But that's only for one supermarket! Maybe I have to add more variables and relations to the model. Maybe I need to consider the supermarket chain, special deals, organic / non-organic, … All these choices add variables, relations and parameters to the model.

What we can do with the model depends on the modeling mindset. In supervised machine learning, we take advantage of the modeled relations to make predictions. In causal inference, we use our model to estimate causal effects. In likelihoodism, we can compare the likelihoods between models.

2 Mindsets

A model is only a bunch of variables, relations, and parameters. A model alone can't tell us how to interpret the world. The use and interpretation of the model depends on the mindset from which the model arose. In order to derive knowledge about the world from the model, we need to make further assumptions. Consider a linear regression model that predicts regional rice yield as a function of rainfall, temperature, and fertilizer use. It's a model, not a mindset. How may we interpret the model? Can we interpret the effect of fertilizer as causal to rice yield? Or is it just a correlation? Would we trust the model to make accurate predictions for the future as well? Can we say anything about the statistical significance of the effect of the fertilizer? Or have we just updated prior information about the effect of fertilizer with new data?

Welcome to **Modeling Mindsets**.

A modeling mindset is a specification of how to model the world using data. Modeling means investigating a real world phenomenon indirectly using a model.[2] Modeling mindsets are like different lenses. All lenses show us the world, but with a different focus. Some lenses magnify things that are close, some that are far away. Some glasses are tinted so you can see in bright environments. When you look through a lens, you see the world in a certain way. With different modeling mindsets, you can look at the modeling task, but the model will focus on different things. Bayesianism, frequentism, supervised machine learning, generative models, … these are all different mindsets when it comes to building models from data. Mindsets differ in how they interpret probabilities – or whether probabilities are part of the language at all. While mindsets cover many different modeling tasks, they have some tasks where they really shine. Each mindset invites you to ask different questions, and so shapes the way you view the world through your model. In supervised machine learning, for example, everything becomes a prediction or classification problem, while in Bayesian inference, the goal is to update our beliefs about the world using probability theory.

Within a mindset there are two worlds: the model world and the real world (Figure 2.1. Or, as McElreath called them in his book "Statistical Rethinking": the "small" and the "large" world.[3] All modeling results are first and foremost statements about the model world, and to be interpreted within the simplified model world.. How and if model results may be transferred from the model world to the real world, depends on the mindset.

Modeling mindsets are normative: A modeling mindset distinguishes between good and bad models. Even though model evaluations are partly based on objective criteria, the choice of a criterion is subjective. Each mindset has their own set of accepted models and evaluation procedures.

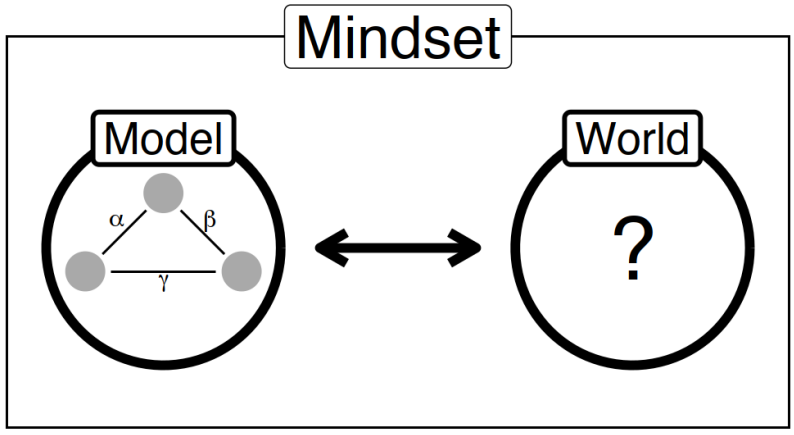

Figure 2.1: Only when a model is embedded in a mindset can we put it into context with the world.

For a frequentist statistician, a good model of the world is probabilistic and has a high goodness-of-fit to the data. The residual errors of the model also pass diagnostic checks. The frequentist rejects the use of prior probabilities as they appear to be subjective. The frequentist would also not switch to a random forest just because it has a lower mean squared error on test data. And why would the statistician switch? From their point of view, the random forest is a poor model of the world. We learn nothing about the probability distribution of our variables. We can't do frequentist hypothesis testing of the effects of variables. The performance of the frequentist's model on unseen test data is not as important to the frequentist.

A modeling mindset limits the questions that can be asked. Consequently, some questions or tasks are out of scope of the mindset. Often the questions are out of scope because they just don't make sense in a particular modeling mindset. Supervised machine learners formulate tasks as prediction or classification problems. Questions about probability distributions are out of reach since the mindset is: choose the model that has the lowest generalization error given new data. So the best model could be any function, such as the average prediction of a random forest, a neural network, and a linear regression model. If the best model can be any function, questions that a statistician would normally ask (hypothesis testing, parameter estimation, …) become irrelevant to the machine learner. In machine learning, the best models are usually not classical statistical models. If the machine learner started asking questions a statistician would ask, they would have to choose a suboptimal model, which is a violation of the mindset.

Modeling mindsets are cultural. Modeling mindsets are not just theories; they shape and are shaped by communities of people who model the world based on the mindset. In many scientific communities, the frequentist mindset is very common. I once consulted a medical student for his doctoral thesis. I helped him visualize some data. A few days later he came back, "I need p-values with this visualization." His advisor told him that any data visualization needed p-values. His advisor's advice was a bit extreme, and not advice that a real statistician would have given. However, it serves as a good example of how a community perpetuates a certain mindset. Likewise, if you were trying to publish a machine learning model in a journal that publishes mostly Bayesian analysis, I would wish you good luck. And I'd bet 100 bucks that the paper would be rejected.

The people within a shared mindset also accept the assumptions of that mindset. And these assumptions are usually not challenged, but mutually agreed upon At least implicitly. If you work in a team that has been using Bayesian statistics for some time, you won't be questioning each model anew about whether using priors is good or whether the Bayesian interpretation of probability is legit. In machine learning competitions, the model with the lowest prediction error on new data wins. You will have a hard time arguing that your model should have won because it's the only causal model. If you believe that causality is important, you would simply not participate. You can only thrive in machine learning competitions if you have accepted the supervised machine learning mindset.

The modeling mindsets as I present them in this book are archetypes: pure forms of these mindsets. In reality, the boundaries between mindsets are much more fluid. These archetypes of mindsets intermingle within individuals, communities and approaches: A data scientist who primarily builds machine learning models might also use some simple regression models with hypothesis tests – without cross-validating the models' generalization error. A research community could accept analyses that use both frequentist and Bayesian statistics. A machine learning competition could include a human jury who would award additional points if the model is interpretable and includes causal reasoning.

Have you ever met anyone who is really into supervised machine learning? The first question they ask is "Where is the labeled data?". The supervised machine learner turns every problem into a prediction or classification problem. Or perhaps you've worked with a statistician who always wants to run hypothesis tests and regression models? Or you had intense discussion about probability with a hardcore Bayesian? Some people really are walking archetypes. 100% of one archetype. But I would say that most people learned one or two mindsets when they start out. And later they got glimpses of other mindsets here and there. Most people's mindset is already a mixture of multiple modeling mindsets. And that's a good thing. Having an open mind about modeling ultimately makes you a better modeler.

3 Statistical Modeling

- Goal: Changing your mind under uncertainty.
- Assumes the world is best described by probability distributions.
- Requires additional assumptions for drawing conclusions. See Frequentism Inference, Bayesian Inference and Likelihoodism.
- Machine learning is an alternative mindset.

Do you become more productive when you drink a lot of water? Is there a fixed "law", like a mathematical function, that expresses productivity as a function of water intake? No. No, because productivity depends on many factors: sleep duration and quality, distractions, noise pollution, … Because of all these factors and other contingencies, we won't get a perfect equation relating water intake to productivity. Uncertainty remains. Even the water intake varies from day to day and from person to person.

Statistics is all about changing your mind under uncertainty. One way to deal with the uncertainty of the world is to abstract aspects of the world as random variables and ascribe a probability distribution to them. Daily water intake could be a random variable. For productivity we would need some clever idea on how to best measure this abstract concept. A not so clever example: Daily time spent in front of the computer. To further relate these variables to each other, we can make assumptions on how the data were generated and connect the random variables in a statistical model.

Welcome to the **statistical modeling** mindset.

A statistical model consists of a set of probability distributions that are fitted to data. A probability distribution describes the frequency with which we expect to see certain values of random variables. The second ingredient to a statistical model is the data, from which we estimate those probability distributions. But let's start with the most elementary unit: the random variable.

3.1 Random Variables

A random variable is a mathematical object that carries uncertainty. Daily water intake can be a random variable. Data are seen as **observations** or realizations of random variables. If someone drank 2.5 liters of water, that is a realization of the random variable "daily water intake".

Other random variables:

- Outcome of a dice roll.
- Monthly sales of an ice cream truck.
- Whether or not a customer has canceled their contract last month.
- Daily amount of rain in Europe.
- Pain level of patients arriving at the emergency room.

People with a statistical modeling mindset **think** in random variables. Any problem related to data is translated into a set of random variables and solved based on that representation. A random variable is a construct that we can't observe directly. But we can observe the realizations of a random variable, as shown in Figure 3.1. But the raw data are not that useful to humans. Because humans aren't databases, we need a model that simplifies the noisy data for us. Statisticians came up with probability distributions: a mathematical construct that describes potential outcomes of the random variable.

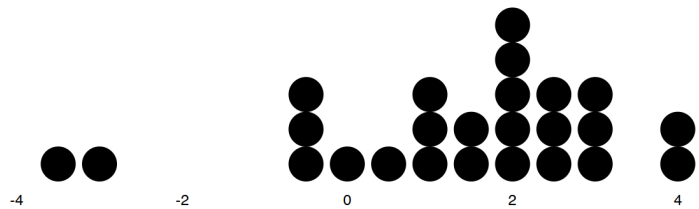

Figure 3.1: Each dot represents a data point, a realizations of a random variable. The x-axis shows the value of the variable. Dots are stacked up for frequently occurring values.

3.2 Probability Distributions

A probability distribution is a function which gives each possible outcome of a random variable a probability. Value in, probability out. Not all functions can be used as probability functions. A probability function must be larger or equal to zero for the entire range of the random variable. For discrete variables such as the number of fish caught, the probability distribution must sum up to 1 over all possible outcomes. And for continuous outcomes such as water intake, the probability distribution, also called density function, must integrate to 1 over the possible range of values.

For the outcome x of a fair dice, we can write the following probability function:

$$P(x) = \begin{cases} 1/6 & x \in \{1, \dots, 6\} \\ 0 & x \notin \{1, \dots, 6\} \end{cases}$$

The Normal distribution is for continuous variables and defined from minus to plus infinity:

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

In this equation, π and e are the famous constants $pi \approx 3.14$ and $e \approx 2.7$. The distribution has two parameters: mean μ and standard deviation σ. These parameters are sometimes called location (μ) and scale (σ) parameters, since they determine where on the x-axis the center of the Normal distribution is and how flat or sharp the distribution is. The larger the standard deviation σ, the flatter the distribution.

Let's try it out and set $\mu = 0$ and $\sigma = 1$, as visualized in Figure 3.2, leftmost curve. Now we can use this probability distribution function for telling us how probable certain values of our random variable are. We get $f(1) \approx 0.24$ and for $f(2) \approx 0.05$, making $x = 1$ much more likely then $x = 2$. We may not interpret $f(x)$ as probability directly. But we can integrate f over a region of the random variable to get a probability. The probability for $x \in [0.9, 1.1]$ is 4.8%.

There is huge arsenal of probability distributions: The Normal distribution, the Binomial distribution, the Poisson distribution, … And there is an infinite number of probability distributions that you could invent yourself.

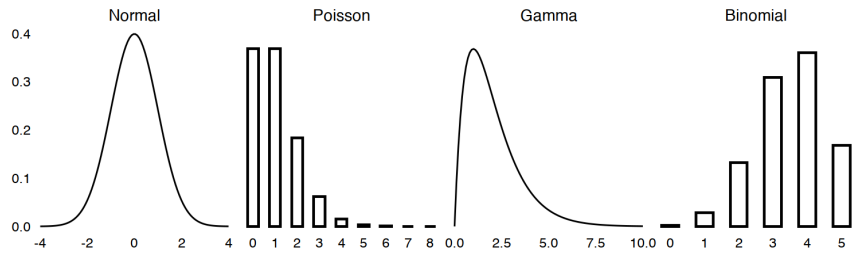

Figure 3.2: Distributions

3.3 Assuming a Distribution

An important step in the statistical modeling mindset is to connect the random variables that we are interested in to distributions. A common approach is to choose a distribution that matches the "nature" of your data:

- A numerical outcome such as IQ? That's Normal distribution.
- A count outcome such as number of fish caught in an hour? The Poisson distribution is a solid choice.
- A binary repeated outcome, like the number of successful free throws in basketball? It follows a Binomial distribution.

These were examples 1-dimensional distributions that only consider a single variable. The world is more complex than that: Random variables can be correlated. The distribution of one variable can depend on the value that another random variable takes. Fortunately, it's possible to define so-called multivariate distributions. A multivariate distribution is a function that takes as input two ore more values and returns a density (still a single number). We might assume that the joint distribution of water intake and productivity (measured in minutes) follows a 2-dimensional Normal distribution.

Another option is conditional distributions. For example, we could assume that productivity, conditional on water intake, follows a Normal distribution.

With all these probability distributions, we are still in the realm of abstraction and assumptions. On one side we have the data: messy and untamed. On the other side, we have the probability distributions: clean and idealized. Via random variables we have at least a theory how the two are connected: The data are the realizations of variables, and distributions summarize the stochastic behaviour of variables. But we still need to mathematically connect observed data and theoretical distributions. How can we connect them?

The answer is **statistical models**.

3.4 Statistical Model

A statistical model connects theoretical distributions with observed data. Statistical models are mathematical models that make assumptions about how the data are generated. With these assumptions in the background, statistical models are then estimated using data. More formally, a statistical model is the combination of the sample space from which the data comes from, and a set of probability distributions on this sample space.

The distributions are "fitted" to the data by changing the parameters. Imagine the distribution as a cat. And your data is a box. Your cat fits it's shape and position to match the box.

How does the cat know which form to take on? Ah pardon, our question now is: How do we find parameter values so that the distribution fits the data well? The density function of the Normal distribution, for example, has mean and variance as parameters. Given the parameters, the density or probability function can tell us how likely certain values of our random variables are.

We can also use the probability or density function to find our parameters – by reversing the point of view and calling it the likelihood function. The value of the random variable is the input of the probability function, and the parameters are seen as given. The likelihood function $L(\theta, X)$ is equal to the probability function. Except that the parameters are now the input, and the values of the random variable are seen as given. They are given, in the sense that we have data that are realizations of the random variable.

We can take an observed value for X from our data and plug it into the likelihood function. Now we have a likelihood function that can tell us, for this one data point, which parameters

would give the highest probability for observed this particular value. But our data consist of multiple realizations of random variables. To get the likelihood for the entire dataset, we multiply the likelihoods of the individual data points. This data likelihood can tell us, for a given parameter value, how likely our data is. For example we could try $\mu = 1$ and $\sigma = 2$ and the likelihood function returns a value. The larger the value, the better the distribution (with these parameters) fits the data. That's very useful, because it helps us in finding the best parameters.

Finding the best parameters is a classical optimization problem: Maximize the data likelihood L with respect to the parameters. We want to maximize the likelihood for all of our data: $L(\theta, \mathbf{x}_1, ..., \mathbf{x}_n) = \prod_{i=1}^{n} L_i(\theta, \mathbf{x}_i)$. Note that $\mathbf{x}_i$ can be a vector with values for multiple variables. And we maximize the data likelihood:

$$\arg\max_{\theta} L(\theta|\mathbf{x}_1, ..., \mathbf{x}_n)$$

Maximum Likelihood Estimation

Maximum likelihood estimation is a classic optimization problem. For simpler cases this can even be solved analytically. For the Normal distribution, we know that the optimal mu is the mean of the data: $\frac{1}{n}\sum_{i=1}^{n} x_i$. When an analytical solution is not possible, other optimization methods like gradient descent, the Newton-Raphson method and Fisher's scoring can be used.

Maximum likelihood estimation is a key element to understanding the statistical modeling mindset. Maximizing the likelihood means bringing together the theoretical probability distributions and the observed data.

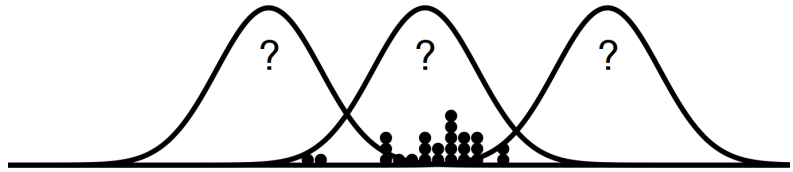

Figure 3.3: Fitting distributions to data

Statistical Hypothesis

A statistical hypothesis is a specification of a probability distribution for (observable) random variables. A statistical model can also embody a statistical hypthesis, because it specifies a probability distribution. For example, saying that the height of people follows a Gaussian distribution is a statistical hypothesis. The effect of a drug on disease progression could be modeled with a logistic regression model. After the model was fitted, the resulting statistical

model also represents a statistical hypothesis. An alternative statistical hypothesis would be that the effect of the drug is zero, which also encodes a statistical model.

As you can see there is a tight relationship between statistical models and statistical hypotheses. All statistical models can be your statistical hypotheses. But a statistical hypothesis is not necessarily fitted with data, or at least parts are manipulated. A statistical hypothesis can come purely from data plus the modeling assumptions. Or it can be purely come from assumptions. Or something inbetween, that some parameters from a statistical model are fitted, the other are assumed to take on a certain value.

What we do with those hyptotheses depends on the mindset. Frequentist inference engages in null hypothesis significance testing: Define an alternative hypothesis to your statistical model (which is a hypothesis); then computed the probability to observe your model under the alternative hypothesis (= p-value). Likelihoodism compares the likelihoods of two statistical hypothesis. In Bayesian inference we can compare the Bayes factor to compare two hypothesis.

Interpreting Parameters

The parameters are not only things to be optimized. Their role goes beyond that of a mere technical controller. The parameters are central to the statistical modeling mindset.

Statisticians interpret the parameters as summary of the data.

Statistical modeling is about understanding our data better. The nice consequence of modeling the data with probability distributions is that we summarize our data with just a few numbers, the parameters. In the case of the Normal distribution, we can describe the distribution of a single variable with only two numbers, the mean and the standard deviation. Also for other types of statistical models, parameters are central to interpretation.

3.5 Joint or Conditional Distribution

All statistical models target an aspect of a probability distribution. For most of the chapter, we talked about simpler cases, like the distribution of a single variable: $P(X)$. But we could have the case of multiple random variables $X_1, \dots, X_p$. To describe the full distribution of more than one variable, we have to work with the joint distribution.

A Gaussian mixture model, for example, requires learning the entire distribution of multiple variables. Gaussian mixture models can be used for identifying clusters in the data, which are assumed to stem from a mixture of Normal distributions. Gaussian mixture models are also an example of using a different optimization algorithm than the maximum likelihood algorithm: the expectation-maximization algorithm, which iteratively jumps between optimizing the model parameters and predicting the "mixture" of cluster centers for each data point.

The joint distribution is not always of interest and can be difficult to estimate. It's often much simpler to work with *conditional distributions*. The conditional distribution tells us

how likely the outcome of one or more random variables is given that we already know the values of some other random variables. For example:

- Which risk factors influence the probability of getting lung cancer?
- How do climatic conditions like temperature and humidity affect the occurrence of bonobos?
- On what days is a hospital likely to be understaffed?

The conditional distribution is the natural form for prediction tasks. And it's usually simpler to estimate than the joint distribution. Models of the conditional distribution are central to statistical modeling. They are also known as regression models.

3.6 Regression Models

Regression models are statistical models of the conditional and not the joint distribution. Let's say we have two variables: Y and X. With the joint distribution $P(Y, X)$ we could answer "How likely is it that $X = x$ and $Y = y$ occur together? But with the conditional distribution we can ask:"Given $X = x$, how likely are certain values for Y"?

For example, we want to know not only how often a disease is successfully treated. We might want to know if a certain drug played a part in the disease outcome. Other factors such as the patient's age, disease progression, etc. may play a role.

Our target is the distribution of outcome variable Y conditional on variables $X_1, ..., X_p$. Within the regression model, the distribution of Y is linked to the variables $X_1, ..., X_p$. Often this means that we express the parameters θ of Y's distribution (such as the mean μ in a Normal distribution) as a function of the other variables. How exactly this link looks like depends on the distribution that the modeler assumed for Y and the link they chose to connect θ and X. The simplest case is a linear regression model. We assume that Y follows a Normal distribution and link the mean of Y's distribution to a weighted sum of the other variables:

$$Y \sim N(\mu, \sigma)$$

$$\mu = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

The linear regression model expresses the mean of the target Y as the weighted sum of the other variables. Y given $\mathbf{X}$ follows a Normal distribution. We only link the mean μ to the variables. A typical assumption is that the standard deviation σ is independent of the value of the other variables.

What can we do now with a regression model? We can make predictions. We just have to fill in values for X and get the expected value of the probability distribution of Y. While the conditional mean of Y is often the goal of the regression model, the statistician can also target other parts of the conditional distribution. For example the median or other quantiles. These

types of regression models are quantile regression models. For survival models the target is often the hazard rate, which is a function of the probability of an event happening in a time interval. See for example the Cox proportional odds model.

Interpretation of parameters is usually even more important than prediction. The parameters contain information about the relationship between the target and the other variables. For regression models, we usually speak of the coefficients. In the linear regression model, a positive coefficient β_j means that increasing the value of variable X_j increases the expected value of Y.

A large part of statistical modeling consists of regression models in different flavors.

- Generalized linear models (GLMs)
- Generalized additive models (GAMs)
- Quantile regression
- Mixed effect models
- ...

Most of these models were "invented" to address a shortcoming of another regression model. GLMs can model targets that don't follow a Normal distribution but some other distribution. GAMs can model non-linear relationships.

Regression models also have a place in supervised machine learning. But in supervised machine learning, the goal is to achieve good predictive performance on unseen data. And if a random forest performs better than a GAM, then the GAM gets thrown out, even if it has nicer statistical properties.

3.7 Model Evaluation

Good statisticians evaluate their models. Understanding how this evaluation works will help you better understand the statistical modeling mindset. It's particularly interesting to see how much the evaluation in the statistical modeling mindset differs from that of supervised machine learning. The evaluation of statistical models consists of two parts: **model diagnostics and goodness-of-fit**.

The role of model diagnostics is to check whether the modeling assumptions were reasonable. If we have assumed that a random variable follows a Normal distribution, we can check this assumption visually. For example, with a Q-Q plot. Another typical assumption is homoscedasticity: The variance of the target is independent of other variables. Homoscedasticity can be checked visually by plotting the residuals (y_i minus its predicted value) against each of the other variables.

A model that passes these diagnostics is not automatically a good model. Statisticians use goodness-of-fit measures to compare models and evaluate modeling choices, such as which variables to have in the model.
Typical goodness-of-fit measures are the (adjusted) R-squared, Akaikes Information Criterion (AIC), the Bayes factor, and likelihood ratios. Goodness-of-fit is literally a measure of how

well the model fits the data. The goodness-of-fit can guide the model building process and decide which model is chosen in the end.

Goodness-of-fit is typically computed with the same data that were used for fitting the statistical models. This choice may look like a minor detail, but it says a lot about the statistical modeling mindset. The critical factor here is overfitting: The more flexible a model is, the better it adapts to ALL the randomness in the data instead of learning patterns that generalize. Many goodness-of-fit metrics therefore account for model complexity, like the AIC or adjusted R-squared. Compare this to supervised machine learning: in this mindset, there is a strict requirement to always use "fresh" data for evaluation.

3.8 Data-Generating Process (DGP)

A quite central, but fuzzy also topic of the statistical modeling mindset is the data-generating process. The statistical modeler thinks about the data-generating process all the time. The data-generating process is a construct, an unknowable ideal of how the data were generated. The data-generating process produces the unknown distributions that then produce the data. We observe the data-generating process indirectly through the data. With data and reflections on the DGP, sometimes together with experts, the statistician tries to decipher the DGP. Statisticians talk about the DGP all the time, but it remains more of a mental model than a clearly defined concept. When I was studying statistics, the DGP was mentioned in most lectures. But I never had a formal lecture on what the DGP is. It's also difficult to find textbooks, lectures, or blog posts that define what the DGP really is. It seems to me that the mental model of a DGP is a natural consequence of viewing the world in terms of probability distributions.

The DGP is a powerful idea, even if it's not well defined. Assuming a DGP encourages you to intellectually dive deep into your data. Having this image of a DGP in your mind let's you take on the mindset of a detective: Statisticians are like detectives reconstructing a crime. You can't observe the crime directly. But the crime has generated a lot of "data" at the scene. The stats detective then tries to uncover the data-generating process by making assumptions and learning from the observed data.

There is no definition of the data-generating process, so I'll give you a some examples:

- Rolling dice is a data-generating process. A die is symmetric, making each side equally likely. We could factor in the angle of the throw, surface roughness and so on. But the chaotic behaviour of the dice bouncing and spinning across the table makes us disregard all these factors.
- We study the income of computer scientists via a survey. Instead of mindlessly reporting the income distribution, we think about the entire data-generating process: For example, some income values are missing. Are they missing at random? Are higher-income individuals more likely to leave the income answer empty? Are some businesses overrepresented in the survey? Is the sample truly random?

- A research team has collected chest x-ray images of patients with and without COVID-19 for building a COVID-19 prediction model. A closer look at the data-generating process shows: the images not only differ in COVID-19 status, but they come from different data-generating processes. COVID-19 images are often taken with patients lying down because of exhaustion. One of the non-COVID-19 datasets are even children x-ray images.[1] Considering the DGP casts doubt on whether such data can be used to build a COVID-19 classifier or whether, in reality, it classifies data into child/adult, standing/lying down, … I chose this example because it's a paper from the supervised machine learning mindset. A good statisticians would check the DGP much more carefully, making it more likely to detect such problems.

If these examples of data-generating processes sound like common sense to you, it's because they are. But it's surprisingly uncommon among non-statistician mindsets. For example, for supervised machine learning, considerations of the data-generating process play a subordinate . For most machine learning competitions the winner is determined solely by the lowest prediction error on new data. It doesn't matter whether the model meaningfully reflects the data-generating process.

3.9 Drawing Conclusions About the World

In most cases, statistical models are built for practical reasons: To make a decision, to better understand some property of the world, or to make a prediction. These goals require the interpretation of the model instead of the world. But using the model as a representation of the world isn't for free. The statistician must consider the representativeness of the data and choose a mindset that allows the findings of the model to be applied to the world.

Considering the data-generating process also means thinking about the representativeness of the data, and thus the model. Are the data a good sample and representative of the quantity of the world the statistician is interested in? Let's say a statistical modeler analyzes data on whether a sepsis screening tool successfully reduced the incidence of sepsis in a hospital. They conclude that the sepsis screening tool has helped reduce sepsis-related deaths at that hospital. Are the data representative of all hospitals in the region, the country, or even the world? Are the data even representative of all patients at the hospital, or are data only available from patients in intensive care unit? A good statistical modeler define and discuss the "population" from which the data are a sample of. Design-based inference fully embraces this mindset that the data are sampled from a larger population.

More philosophical is the modeler's attitude toward causality, the nature of probability, and the likelihood principle.

[1] Wynants, Laure, Ben Van Calster, Gary S. Collins, Richard D. Riley, Georg Heinze, Ewoud Schuit, Marc MJ Bonten et al. "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal." bmj 369 (2020).

> "It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel."

– Leonard Savage, 1972[2]

Statistical modeling is the foundation for learning from data. But we need another mindset on top of that to make the models useful.

Frequentist inference is the most prominent mindset for inferring properties about the world from statistical models. Frequentist statistics sees probability as relative frequencies of events in long-run experiments.

Bayesian inference is based on an interpretation of probability as a degree belief about the world. Bayesianism states that the model parameters also have a (prior) distribution. And the goal of the statistician is to update the prior distribution by learning from data. The resulting posteriori distributions of the parameters express our belief about the world.

Likelihoodism is a lesser known modeling mindset. Like Bayesianism, it adheres to the likelihood principle, which states that the likelihood function captures all evidence from the data (which frequentist inference violates). However, it does not require prior probabilities.

Causal inference adds causality to statistical modeling. It can be superimposed onto any of the other three mindsets.

A different but complimentary approach is design-based inference, which focuses on data sampling and experiments instead of models.

3.10 Strengths

- The statistical modeling mindset is a *language to see the world*. Even when not used for inference, random variables and probability distributions are useful mental models for perceiving the world.
- Statistical modeling has a long tradition and extensive theoretical foundation, from measurement theory as the basis of probability theory to convergence properties of statistical estimators.
- The data-generating process is an underestimated mental model. But it's a powerful mental model that encourages mindful modeling and asking the right questions.
- Conditional probability models can be used not only to learn about the parameters of interest, but also to make predictions
- Probability distributions give us a language to express uncertainty. Bayesianism arguably has the most principled focus on formalizing and modeling uncertainty.
- Can naturally handle different types of variables: Categorical, ordinal, numerical, …

[2]Savage, Leonard J. The foundations of statistics. Courier Corporation, 1972.

3.11 Limitations

- Statistical modeling reaches its limits whenever defining probability distributions becomes difficult. Images and text don't easily fit into this mindset, and this where supervised machine learning and especially deep learning shine.
- Working within the statistical modeling mindset can be quite "manual" and tedious. It's not easy to always think about the DGP, and sometimes more automatable mindsets such as supervised machine learning are more convenient.
- Statistical models require a lot of assumptions. Sometimes more, sometimes less. Just to name a few common assumptions: homoscedasticity, independent and identically distributed data (IID), linearity, independence of errors, lack of (perfect) multicollinearity, … For most violations, there is a special version of a statistical model that doesn't require the critical assumption. The price is more complex and less interpretable models.
- Statistical modeling, when used for prediction, is often outperformed by supervised machine learning. To be fair, outperforming here requires an evaluation based on the supervised learning mindset. However, this means that goodness-of-fit and diagnosis are no guarantee that a model will performs well on all metrics.

4 Frequentist Inference

- Popular modeling mindset in science.
- The world consists of probability distributions with fixed parameters that have to be uncovered.
- Interprets probability as long-run relative frequencies from which hypothesis tests, confidence intervals and p-values are derived.
- A statistical mindset, with Bayesian inference and likelihoodism as alternatives.

Drinking alcohol is associated with a higher risk of diabetes in middle-aged men. At least this is what a study claims.[4]

The study modeled type II diabetes as a function of various risk factors. The researchers found out that alcohol significantly increases the diabetes risk for middle-aged men by a factor of 1.81.

"Significant" and "associated with" are familiar terms when reading about scientific results. The researchers in the study used a popular modeling mindset to draw conclusions from the data: frequentist inference. There is no particular reason why I chose this study other than it is not exceptional. When someone thinks in significance levels, p-values, hypothesis tests, null hypotheses, and confidence intervals, they are probably frequentist.

In many scientific fields, such as medicine and psychology, frequentist inference is the dominant mindset. All frequentist papers follow similar patterns, make similar assumptions, and contain similar tables and figures. Knowing how to interpret model coefficients, confidence intervals and p-values is like a key to contemporary scientific progress. Or at least a good part of it. Frequentism not only dominates science, but has a firm foothold in industry as well: Statisticians, data scientists, and whatever the role will be called in the future, use frequentist inference to create value for businesses: From analyzing A/B tests for a website to calculating portfolio risk to monitoring quality on production lines.

As much as frequentism dominates the world of data, it's also criticized. Frequentist inference has been the analysis method for scientific "findings" that turned out to be a waste of research time. You may have heard about the replication crisis.[5] Many scientific findings in psychology, medicine, social sciences and other fields could not be replicated. The problem with that is that replication is at the center of the scientific method. Many causes have contributed to the replication crisis But frequentist statistics is right in the middle of it. The frequentist mindset enables practices such as multiple testing and p-hacking. Mix this with the pressure on academics to "publish or perish". The result is a community that is incentivized to squeeze out "significant" results at a high rate. Frequentism is a decision-focused mindset and can give seemingly simple yes/no answers. Humans are lazy. So we tend to forget all the footnotes and remarks that come with the model.

Frequentist inference is a statistical modeling mindset: It depends on random variables, probability distributions, and statistical models. But as mentioned in the chapter Statistical Modeling, these ingredients are not sufficient to make statements about the world.

Frequentism comes with a specific interpretation of probability: Probability is seen as the relative frequency of an event in infinitely repeated trials. That's why it's called frequentism: frequentist inference emphasizes the (relative) frequency of events. But how do these long-run frequencies help to gain insights from the model?

Let's go back to the 1.81 increase in diabetes risk among men who drink a lot of alcohol. 1.81 is larger than 1, so there seems to be a difference between men who drink alcohol and the ones who don't. But how can the researchers be sure that the 1.81 is not a random result? For fair dice, the average eyes in the long-run series of experiments is 3.5. If I roll a die 10 times and the average is 4, would you say it's an unfair die? No? Would you say it's unfair if the average is 4.5? 5? Or if a 6 shows up 10 times?

The researchers applied frequentist thinking to decide between randomness and true effects. The parameter of interest is a coefficient in a logistic regression model. The logistic regression model links variables such as alcohol to diabetes. In the diabetes study, a 95% confidence interval for the alcohol coefficient was reported: The interval goes from 1.14 to 2.92. This interval settles the question of randomness versus signal: The interval doesn't contain 1, and so the researchers concluded that alcohol is a risk factor for diabetes (in men). This confidence interval describes uncertainty regarding the alcohol coefficient. If we were to repeat the analysis many times with new samples, the respective 95% confidence interval would cover the "true" parameter 95% of the time. Always under the condition that the model assumptions were correct.

4.1 Frequentist probability

The interpretation of the confidence interval reveals the **core philosophy of frequentism**:

- The world can be described by probability distributions;
- The parameters of the probability distributions are constant and unknown;
- Repeated measurements/experiments reveal the true parameter values in the long-run.

In contrast, Bayesianism assumes that the parameters of the distributions are themselves random variables. As the frequentists collect more and more data ($n \to \infty$), their parameter estimation gets closer and closer to the true parameter (if the estimator is unbiased). With each additional data point, the uncertainty of the estimated parameter shrinks and the confidence interval becomes narrower.

The frequentist interpretation of probability requires imagination. Frequentists start with a population in mind. The population can be adults between 20 and 29 living in Iceland, daily measurements of water quality of the Nile River, or 1-inch wood screws manufactured in a factory in the U.S. state of Texas. These populations can be described by finding out their probability distributions. Going back to the initial example: What's the probability

that a middle-aged man will develop diabetes in the next 12 months? Frequentists would say: There is an unknown and fixed probability for diabetes. The more people we observe, the more accurate our estimate of the probability of diabetes becomes. We estimate the probability of diabetes as the relative frequency of diabetes in the population. Probabilities are frequencies in the long-run:

$$P(X = 1) = \lim_{n \mapsto \infty} \frac{1}{n} \sum_{i=1}^{n} I(x_i = 1)$$

.

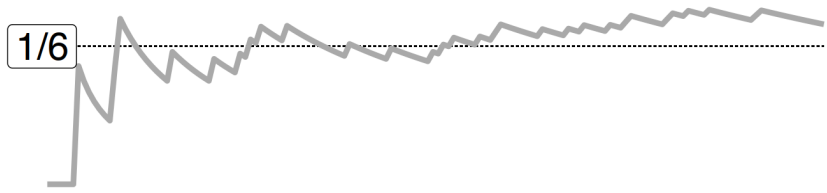

Figure 4.1: The line shows how the relative frequency of 6 eyes changes as the number of dice roles increases from 1 to 100 (left to right).

Imagining experiments that will never take place is essential to the frequentist mindset. By defining probability in terms of long-run frequencies, frequentism requires imagining that the sampling and experiment will be done many times. These "imagined" experiments are central to the interpretation of confidence intervals, p-values, and hypothesis tests. And *every* interpretation of probability in the frequentist mindset has to be connected to frequencies of events in long-run samples / experiments.

These imagined experiments have a curious implication for frequentism. Frequentism violates the likelihood principle, which says that all evidence about the data is contained in the likelihood function. But with frequentism, it's important to know what experiments we are further imagining. You can find a simple example involving coin tosses in the chapter on Likelihoodism. Likelihoodism and Bayesianism adhere to the likelihood principle.

4.2 Estimators are Random Variables

We can learn a lot about frequentist inference, especially in contrast to Bayesian inference, by understanding which "things" are random variables and which are not. In the frequentist mindset, the estimand, the "true" but unknown parameter is assumed to be fixed. Mean, variance and other distribution parameters, model coefficients, nuisance parameters, all are seen as having some unknown but fixed value. And the values can be uncovered with frequentist inference. Bayesians, in contrast, view all these parameters as random variables.

Since the quantities of interest are seen as fixed but unknown, the frequentist's job is to estimate them from data. The estimation is done with a statistical estimator: A mathematical procedure for inferring the estimand from data. The estimator is a function of the

data. And data are realizations of random variables. As a consequence, the estimators themselves become random variables. Let's compare this with the Bayesian mindset: Bayesians assume that the parameters are random variables. Bayesian inference updates the (prior) probability distributions of the parameters, which results in the posterior distributions of the parameters.

Typical frequentist constructs like confidence intervals, test statistics and p-values are also random variables. Mix this with the long-run frequencies and you get a special interpretation, for example, for confidence intervals.

Let's say you want to know how many teas you drink on average per day. If you are a frequentist, you would assume that there is a true but unknown daily number of teas. Let's call this estimand λ. The frequentist might assume that the daily number of teas follows a Poisson distribution. The Poisson distribution can handle count data well, and is described by the "intensity" λ with which events happen. The intensity parameter λ is also the expected number of events. Teas in our case. We could estimate the tea intensity using the maximum likelihood estimator: $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} k_i$, where k_i is the number of teas on day i. Our estimator $\hat{\lambda}$ is a random variable. If the model assumptions are correct and if the world is truly frequentist, then the estimator $\hat{\lambda}$ will get closer and closer to the true λ as n increases. The estimator $\hat{\lambda}$ approximately follows a Normal distribution.

Frequentist inference builds on the fact that the estimators are random variables. Combined with the idea of fixed true parameters, it becomes possible to connect the analysis results to the world. A commonly used tool to derive insights about the world is null hypothesis significance testing.

4.3 Null Hypothesis Significance Testing

Let's say, your estimator $\hat{\lambda}$ says that you drink 2.5 teas per day on average. Initially you had the hypothesis that you drink at least 3.0 teas per day. Obviously, $2.5 \neq 3.0$, so the initial hypothesis is incorrect. Case closed. But that would be too simple an answer, wouldn't it? You also wouldn't say that a coin is unfair if heads come up in 51/100 tosses just because $51 \neq 50$. But when would a frequentist reject the initial hypothesis of 3.0 teas? Would we reject the hypothesis if we get $\hat{\lambda} < 2.9$, or $\hat{\lambda} < 2.5$ or maybe must it be much lower, like $\hat{\lambda} < 1.5$? With the statistical modeling mindset alone, we can't answer this question.

The frequentist mindset has an answer to this question of whether to accept or reject a hypothesis. The frequentist estimator for the number of teas is a random variable that is supposed to approximate the true number of teas. We can make (frequentist) probabilistic statements about this estimator. And while the true value for λ is unknown, we can study the hypothesis of $\lambda = 3.0$ by examining the random variable $\hat{\lambda}$.

This idea of proposing a hypothesis, and then accepting or rejecting it based on a statistical model or test is called null hypothesis significance testing. [1] Hypothesis testing is a central

[1] There are two other "main" approaches for hypothesis testing: The Fisher approach, and the Neyman-Pearson approach. Null hypothesis significance testing is a combination of the two.[6]

method in the frequentist modeling mindset. Hypothesis tests simplify decisions: The frequentist accepts or rejects the so-called null hypothesis based on the results of the statistical model. A statistical model can be very simple: It can be as simple as assuming that the data follow a Normal distribution and comparing two means with a Student t-test.

How does a hypothesis test work?

- Start with a hypothesis.
- Formulate the **alternative or null hypothesis**.
- Decide which statistical test to use. This step includes modeling the data. In fact, all statistical tests are statistical models[3]
- Calculate the distribution of the parameter estimates under the null hypothesis (or rather, the test statistic T).
- Choose the significance level α: the probability threshold at which to reject the null hypothesis assuming it would be true. Often $\alpha = 0.05$.
- Calculate the p-value: Assume that the null hypothesis is correct. Then p is the probability of getting a more extreme test statistic T than was actually observed. See figure 4.2.
- If p-value $< \alpha$, then the null hypothesis is rejected.

Some examples of tests and test statistics:

- Comparing the means of two distributions. Do Germans consume more pretzels than U.S. Americans? Hypothesis: Germans eat more pretzels. The "model" of the data simply assumes a Normal distribution for average pretzel consumption per person and year. The null hypothesis would be that Germans and U.S. Americans consume the same amount. Then we would run a t-test. The test statistic in the t-test is the (scaled) difference of the two means.
- Estimating the effect of one variable on another. Is surgery better than physiotherapy for treating a torn meniscus in your knee? The statistical model could be a linear regression model. The model could predict knee pain dependent on whether a patient had physiotherapy or surgery. The null hypothesis would be that there is no difference in pain, so a model coefficient of zero for surgery/physiotherapy. The test statistic T would be the coefficient divided by its standard deviation.

The p-value has a frequentist interpretation because it's based on long-run frequencies. To interpret the p-value, we have to pretend that the null hypothesis is true. Then the p-value is the probability of observing the outcome of our analysis or a more extreme one. Again, the frequentist interprets probability with imagined future experiments. A p-value of 0.03 for an estimated average of 3.0 daily teas would mean the following: If we repeat the analysis many times and the null hypothesis $\lambda = 2.5$ is correct, 3% of the time we would observe an estimate of $\hat{\lambda} \geq 3$. If $\alpha = 0.05$ was chosen, the null hypothesis would be rejected.

Null hypothesis testing is very weird. It's like answering the question around two corners. Let's say a researcher wants to prove that a drug prevents migraines. They test the drug because they expect it to work, so the hypothesis they assume to be true is that patients that take the drug have fewer migraines. But the null hypothesis is formulated the other way around: The null hypothesis assumes that the drug has no effect. Suddenly the goal of

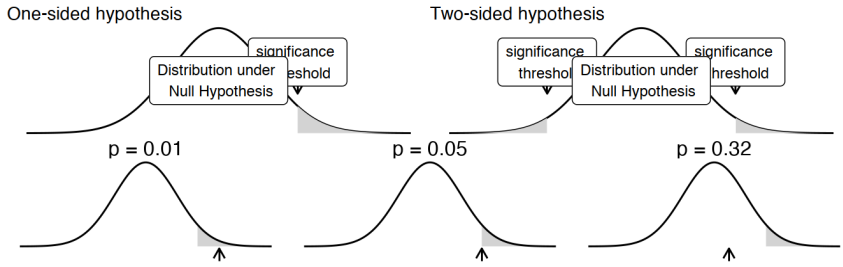

Figure 4.2: Frequentists make binary decisions based on hypothesis tests. Assuming the null distribution, the null hypothesis is rejected if the observed estimand is extreme.

the researcher becomes to show that the null hypothesis is false, rather than showing that their hypothesis is correct. The problem with statistical models is: We can't prove that they are true because our assumptions are not testable. With frequentist inference, however, we can tell how likely a model result is under a given hypothesis and given the data. That's why hypothesis tests work by rejection. The likelihoodist mindset navigates this issue: two statistical models are compared in terms of the evidence through the likelihood.

Null hypothesis tests are even more troublesome.

- The choice of the null hypothesis is arbitrary.
- If the null hypothesis is accepted, it's not evidence that it's true. It just means that the data that were observed are not in conflict with the null hypothesis. But there is still an infinite number of models that could have produced the same results.
- If the null hypothesis is rejected, it doesn't mean that the hypothesis of interest is true.
- A significant result doesn't mean that the deviation from the null hypothesis is relevant. The drug has a significant effect on the disease progression? The difference might be too small to be relevant.
- Especially the larger the data sample, the more likely the null hypothesis is rejected, because the tiniest differences to the null hypothesis are enough to produce significant results when n becomes large.

4.4 Confidence Intervals

Frequentists use confidence intervals as an alternative to statistical tests. Hypothesis tests and confidence intervals ultimately lead to the same decisions, but confidence intervals are more informative. Many statisticians prefer confidence intervals over mere p-values.

Remember that estimators, such as model parameters, are random variables? That means that estimators have probability distributions. A confidence interval describes where the mass of that distribution lies. The interval consists of the estimator, and the lower and upper bounds for the mass of the distribution. The modeler decides the percentage of the

distribution in the confidence interval through the α-level. If $\alpha = 0.05$, then we get a 95%-confidence interval. The construction of the confidence interval depends on the probability distribution we have derived for the quantity of interest (coefficient, mean estimate, ...).

How are the confidence intervals to be interpreted? Well, in a frequentist manner, of course! The "true" parameter value is fixed, so it's not a random variable. To say that the true parameter is in the confidence interval with a 95% probability would be false. The true parameter is either in the interval or it's not, we just don't know. The confidence itself is a random variable since it's derived from data and therefore from other random variables. So the interpretation of a 95% confidence interval is: If we were to repeat the analysis many times, the confidence interval would cover the true value of the quantity of interest 95% of the time. Only given that the model assumptions are correct. As you can see, this is a very frequentist point of view: the confidence interval is interpreted in the context of repeated experiments.

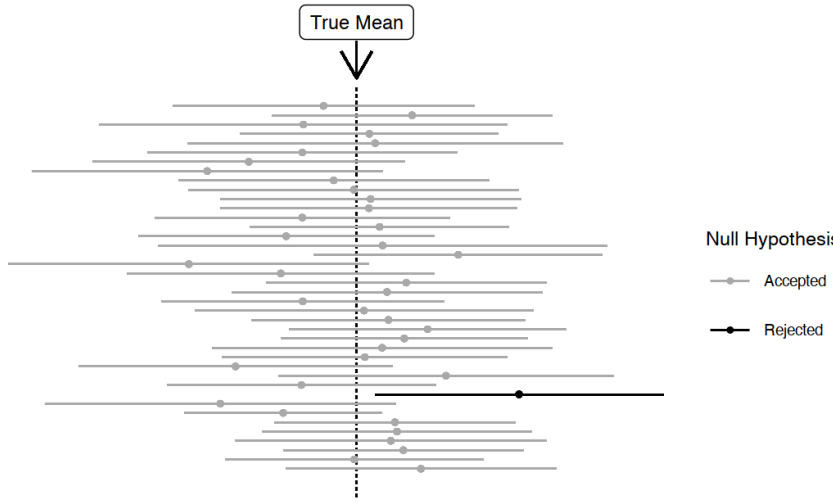

Figure 4.3: 100 95% confidence intervals and the true value.

4.5 Strengths

- Once you understand frequentist inference, you have the key to understanding most modern research findings. I studied statistics and can now quickly grasp many research papers. For example, to figure out whether I should have knee surgery for my torn meniscus, I read papers comparing knee surgery and physiotherapy alone. All of those

papers used frequentist methods, and although I didn't understand everything, I was able to quickly get an idea of their analyses and results.
- Frequentist methods are generally faster to compute than methods from Bayesian inference or machine learning.
- Compared to Bayesianism, no prior information about the parameters is required. This makes frequentism more objective.
- Frequentism allows binary decisions (accept/reject hypothesis). This simplicity is one of the reasons why frequentism is popular for both scientific publications and business decisions.
- Frequentism has all advantages of statistical models in general: a solid theoretical foundation and an appreciation of the data-generating process.
- When the underlying process is a long-run, repeated experiment, frequentist inference shines. Casino games, model benchmarks, …
- The scientific method requires that scientific experiments be repeatable. The frequentist idea that truth lies in long-run frequencies of experiments is therefore well compatible with a core idea of science.

4.6 Limitations

- Frequentism makes it easy to **over-simplify questions** into yes/no-questions. Reducing models to binary decisions obscures critical model assumptions and the difficult trade-offs that had to be made for the analysis.
- Focusing on p-values encourages **p-hacking**: the either conscious or unconscious search for "positive" results. Guided by the lure of a significant result, researchers and data scientists may play around with their analysis until the p-value in question is small enough. With α-level of 0.05, 1 in 20 null hypotheses are falsely rejected. P-hacking increases this percentage of false positive findings.
- Similarly, if the analysis is exploratory rather than hypothesis-driven, a naive frequentist approach may produce many false positive findings. Look again at figure 4.3: Imagine these were confidence intervals for different variables. Again, for $\alpha = 0.05$, we would expect 1 in 20 hypothesis tests to yield false positives. Now imagine a data scientist testing hundreds of hypothesis tests. This problem is called the multiple testing problem. There are solutions, but they are not always used and multiple testing can be very subtle.
- The frequentist interpretation of probability is very awkward when it comes to confidence intervals and p-values. They are commonly misinterpreted. Arguably, frequentist confidence intervals are not what practitioners want. Bayesian credibility intervals are more aligned with the natural interpretation of uncertainty.
- Frequentist analysis depends not only on the data, but also on the experimental design. This is a violation of the likelihood principle that says that all information about the data must be contained in the likelihood, see also the example in the Likelihoodism chapter.
- Frequentist probability can fail in the simplest scenarios: Imagine you are modeling the probability of rain in August. The data only has 20 August days, all of which are without rain. The frequentist answer is that there is absolutely no chance that it will

ever rain in August. The frequentist recommendation is that to collect more data if we want a better answer. Bayesianism offers a solution to involve prior information for such estimates.

- There is an "off-the-shelf"-mentality among users of frequentist inference. Instead of carefully adapting a probability model to the data, an off-the-shelf statistical test or statistical model is chosen. The choice is based on just a few properties of the data. For example, there are popular flow charts of choosing an appropriate statistical test.
- Frequentist statistics says nothing about causality except that "correlation does not imply causation".
- Weird interpretation of probability: Often it does not make any sense to interpret every probability with imagined experiments. For example, the probability for a party to get the majority vote in an election requires to imagine multiple elections, yet under the same circumstances, like same year, same candidates, and so on.

5 Bayesian Inference

- Probability is a degree of belief, and learning from data means updating belief.
- Model parameters are random variables with prior and posterior distributions.
- A statistical modeling mindset with frequentism and likelihoodism as alternatives.

While frequentists analyze data to answer the question "What should I do?", Bayesians analyze data to answer "What should I believe?".

If you haven't read the chapter on Statistical Modeling, I recommend that you do so first, since Bayesian inference is easier to understand if you have a good understanding of statistical inference. Bayesian inference is based on probability distributions, interpreting parameters, and learning from data through the likelihood. The twist: distribution parameters are also random variables. Random variables that have a prior distribution. Prior means before encountering the data. Learning about the parameters means updating the prior probabilities to the posterior probabilities.

In Bayesian statistics, probability can be interpreted as the plausibility of an event or our belief about it. That's different from the more objective interpretation of probability in frequentist inference. In Bayesian inference, it's not necessary to imagine repeated experiments. One can even apply Bayesian inference to a single data point. This is not possible with frequentist inference. A Bayesian model even works without data, by just using the prior distributions. [1]

5.1 Bayes Theorem

Bayesians want to learn the distribution of the model parameters from the data: $P(\theta|X)$. $P(\theta|X)$ is a strange way of looking at the probabilities involved. The data-generating process generates data as a function of the parameters: $P(X|\theta)$. So Bayesians look for the inverse of what the DGP would naturally do.

To make $P(\theta|X)$ computable, Bayesians use a trick that earned them their name: the Bayes' theorem. Bayes' theorem can be used to invert the conditional probability:

[1] This is called the prior predictive simulation and is used to check that the chosen priors produce reasonable data. The modeler simulates by first sampling parameters from the priors and using those parameters to generate data. Repeating this several times results in a distribution of data.

$$\underbrace{P(\theta|X)}_{\text{posterior}} = \frac{\overbrace{P(X|\theta)}^{\text{likelihood}} \cdot \overbrace{P(\theta)}^{\text{prior}}}{\underbrace{P(X)}_{\text{evidence}}}$$

$P(\theta)$, also called prior, is the probability distribution of θ before we have collected any data. The probability distribution is updated by multiplying the prior by the data likelihood $P(X|\theta)$. [2] This product is scaled by the probability of the data $P(X)$, also called evidence. The result is the posterior probability distribution, an updated belief about the parameters θ.

Bayes' theorem is a generally useful equation for working with probabilities, but we focus on its use for Bayesian inference. The theorem is not just a simple rearrangement of probabilities, but a powerful mental model: Bayesian updating. Remember that the data likelihood is the product of the likelihoods for each data point: $\$P(X|\theta) = \prod_{i=1}^{n} P(X^{(i)}|\theta)$. Here, $X^{(i)}$ is the vector of random variables of the i-th outcome. For example, in a drug trial, the variables $\$X\^{}\{(i)\}$ could belong to the i-th (not yet observed) patient and include pain level, blood pressure and iron level. Plugging this version of the likelihood into Bayes' theorem, we can see how it relates to updating ones belief with new data. For simplicity, I have removed the evidence $P(X)$ which only serves as a normalization so we can interpret the results as probability:

$$\begin{aligned} P(\theta|X) &\propto P(X|\theta) \cdot P(\theta) && (5.1)\\ &= P(\theta) \cdot \prod_{i=1}^{n} P(\theta, X^i) && (5.2)\\ &= P(\theta) \cdot P_1 \cdot ... \cdot P_p, && (5.3) \end{aligned}$$

where $P_i = P(X^{(i)}|\theta)$. Even with just one data point, we can update our belief about the parameters! And each time, this posterior then becomes – in a sense – the prior for the next update. The posterior distribution is the product of prior and likelihood (Figure 5.1).

Next, let's explore the individual terms of Bayes' theorem, so that we can update our own beliefs about Bayesian inference.

5.2 Prior Probability

Bayesians assume that model or distribution parameters a prior probability distribution $P(\theta)$. [3] Let's say we randomly pick a person and want to know how many hours per day they

[2] In the Bayesian mindset θ is a random variable, so the notation can be confusing: $P(X|\theta)$ refers to the likelihood function. The same function that we used in frequentist inference. But in the frequentist inference, we might write $P(\theta|X)$ to emphasize that θ is the input that varies. But this notation refers to the posterior in Bayesian statistics, so we write $P(X|\theta)$.

[3] When I was learning about Bayesian inference, one of my first question was: do the parameters of the priori distributions have priori distributions themselves? Where does it end? The answer: Bayesian stop after

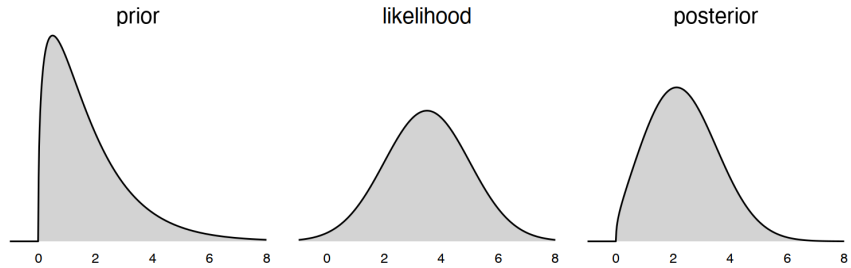

Figure 5.1: The posterior distribution (right) is the scaled product of prior and likelihood: prior $\times$ likelihood $\propto$ posterior.

usually work. The number of hours worked per day, the random variable of interest, follows a probability distribution. For example, the number of hours worked might follow a Gaussian distribution. Bayesians assume that mean and variance of this Gaussian distribution are random variables.

How do priors make sense? How can Bayesians know the distribution of parameters *before* observing any data? Priors are a consequence of saying that parameters are random variables, and a technical requirement for working with the Bayes' theorem. But how can we know anything about the parameters before we see the data?

Picking a Prior

The first consideration in choosing a prior is the *space* the parameter is in. Is the parameter the mean of a continuous distribution? If so, it makes sense for the parameter to follow a continuous distribution as well, such as a Gaussian distribution. Maybe the mean of the data distribution has to be positive. Then the prior distribution should contain only positive values (meaning the probability for negative values should be zero), for example, the Gamma distribution. Furthermore, expert knowledge can be used to choose the prior. Maybe we know from other experiments that the mean parameter should be around 1. So we could assume a Gaussian distribution for θ: $\theta \sim N(1, 1)$. In the case where the data follow a Binomial distribution, the Beta distribution is a good prior (see Figure 5.2). Depending on what the modelers believes about the success probability parameter p of the Binomial distribution, they might choose different parameterizations of the Beta distribution. Maybe the modeler believes that the parameter must be symmetrically distributed around 0.5. Or maybe the parameter is lower, around 0.25? Another Beta prior might put emphasis on p being 1. It's even possible to have a prior that that places the greatest probability symmetrically on 0 and 1.

Without expert knowledge about the parameter, the modeler can use "uninformative" or "objective" priors[7]. Uninformative priors often produce results similar to those of frequentist inference (for example for confidence/credible intervals). Another factor influencing the choice

[7] one level of priori distributions and don't go full inception.

of prior is mathematical convenience. Conjugate priors are convenient choices. Conjugate priors remain in the same family of distributions when multiplied by the right likelihood functions. A Beta prior distribution multiplied by a Bernoulli likelihood, in turn, produces a Beta posterior distribution.

Although there are all these different strategies for choosing a prior, even "objective" ones, the choice remains subjective. And this subjective choice of prior is why many frequentists reject Bayesian inference. While the prior can be seen as a bug, it can also be seen as a feature. Thanks to the prior, Bayesian modeling is very flexible. The prior can be used to constrain and regularize model parameters, especially when data are scarce; the prior can encode results from other experiments and expert knowledge; the prior allows a natural handling of measurement errors and missing data.

To obtain the posterior distribution of the parameters – the ultimate goal of Bayesian inference – we need to update the prior using data, or rather, the likelihood function.

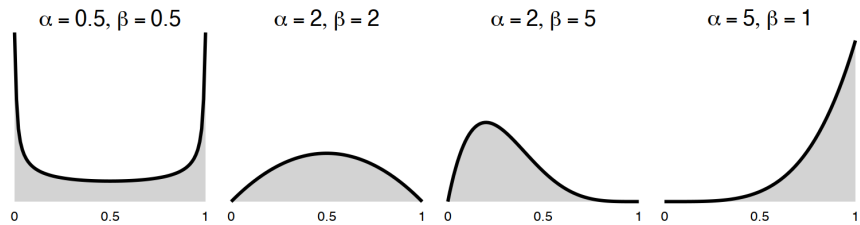

Figure 5.2: Various Beta prior distributions for the success probability p in a Binomial distribution. The parameters α and β influence the shape of the Beta prior.

5.3 Likelihood

If you have read the chapter Statistical Modeling, you should be familiar with the likelihood function $P(\theta, X)$. The likelihood is equivalent to the probability function of the data. Only that the focus is switched: the likelihood is a function of the parameters, while the probability is a function of the data.

The Bayesian makes an assumption about how the data are distributed and forms the likelihood function. This is no different than frequentism and likelihoodism. But for Bayesians, the likelihood is just one part of the equation.

With all the comparisons between frequentist inference and Bayesian inference, it's easy to forget that both approaches are quite similar. Comparisons between the two mindsets often focus on the (lack of) prior distribution. This overlooks the fact that both frequentist and Bayesian inference use the likelihood at the core of their models. Especially in cases with a lot of data, both approaches produce similar results. That's because the more data are available for the model, the less impact the prior has on the Bayesian results.

Let's now turn to the last part of the equation: the evidence $P(X)$.

5.4 Evidence

The evidence is the marginalized probability of the data. Marginalized means that the probability of the data is integrated over all possible parameter values: $P(X) = \int_\Theta P(X|\theta)P(\theta)d\theta$, where Θ are all possible parameter values. Because of this marginalization, $P(X)$ is no longer a function of the parameters θ. $P(X)$ is just a constant factor in terms of maximizing the posterior probability Constant factors don't change *where* the maximum is, just how large it is. In search of the maximum, the evidence $P(X)$ can be ignored. For this reason, the posterior probability is often expressed as proportional to the numerator:

$$\underbrace{P(\theta|D)}_{\text{posterior}} \propto \overbrace{P(D|\theta)}^{\text{likelihood}} \cdot \underbrace{P(\theta)}_{\text{prior}}$$

Just one problem: When throwing away $P(X)$, the posterior probability is not a probability at all, because it doesn't integrate to 1, but to $P(X)$. How can this problem be solved?

5.5 Posterior Probability Estimation

The goal of the Bayesian modelers is to estimate the posterior distributions of the parameters. Once the modelers have the posteriors, they can interpret them, make predictions, and draw conclusions about the world.

But how is the posterior estimated? In the ideal case, the posterior can be written down as a simple formula. But that's only possible for certain combinations of prior and likelihood, for example when conjugate priors are used. For many Bayesian models it's impossible to obtain a closed form for the posterior. The main problem is that $P(X)$ may not be computable.

Sample From the Posterior with MCMC

The good news: We don't have to compute the posterior probability. We can sample from it. Approaches such as Markov Chain Monte Carlo (MCMC) and derivations thereof are used to generate samples from the posterior distribution.

The rough idea of MCMC and similar approaches is as follows:

- Start with some initial values for the parameters θ.
- Repeat the following steps until a stopping criterion is reached (like pre-determined number of samples):
    1. Propose new parameter values. Proposals are based on a proposal function receives as input the previous parameters.
    2. Accept or reject the new values, based on an acceptance function. The acceptance function depends on the prior and the likelihood, but not on the evidence.

3. If the new parameter are accepted, continue with these new values.

A run of MCMC produces a "chain" of samples from the posterior distribution. MCMC can be repeated to produce multiple chains.

MCMC has many variants such as Gibbs sampling and the Metropolis-Hastings algorithm. Each variant differs in proposal and acceptance functions or other algorithmic steps.

MCMC produces a random walk through the posterior distribution of the parameters. Regions where the parameters have a high probability are also "visited" with a higher probability. The samples can be seen as samples from the posterior. But first, some cleaning up needs to happen: Since MCMC has to start somewhere, it's possible that the first samples will be from parameter regions with low probability. So the first 1000 or so samples are "burned", meaning they are not used for estimating the posterior. Another problem is autocorrelation within the chain: Samples that occur one after the other are correlated since the proposal function usually proposes new parameters that are close to the previous values. So the chain is sampled at different points to ensure that there are enough MCMC steps between two samples to make them independent.

MCMC sampling can be complex and can take some time to compute. Fortunately, most probabilistic software runs MCMC automatically. But this can take time. More time than fitting a frequentist model would take. A shorter alternative is variational inference.[8] But while MCMC delivers approximately exact estimates of the posterior probability, variational inference is more of a heuristic.

Frequentists have their parameter estimates. Bayesians have … samples from the posterior distribution? That's not the end of the story. There are two more steps required to get from posterior samples to insights: turning the samples to a distribution and (optionally) summarizing the distribution.

5.6 Summarizing the Posterior Samples

The posterior samples can be visualized with a histogram or a density estimator for a smoother looking curve. Visualizing the entire posterior is the most informative way of reporting the Bayesian results.

People love numbers and tables. The fewer and simpler, the better. You won't get your manager to understand posterior distributions. They demand simple answers! So, let's simplify the posterior. There's advice about not summarizing the posterior[9], but people do it anyways. Summaries of the posterior can be points or intervals. Intervals can be defined via fixed boundaries or fixed probability mass. Some examples:

- Point estimate: The parameter value with the highest posterior probability.
- Interval with fixed boundaries: The interval from 10 to infinity indicates the probability that the parameter is greater than 10.

- Interval with fixed probability mass: The shortest interval containing 50% of the posterior probability mass. Or the interval that ranges from the 2.5% quantile to the 97.5% quantile (called the 95% credible interval).

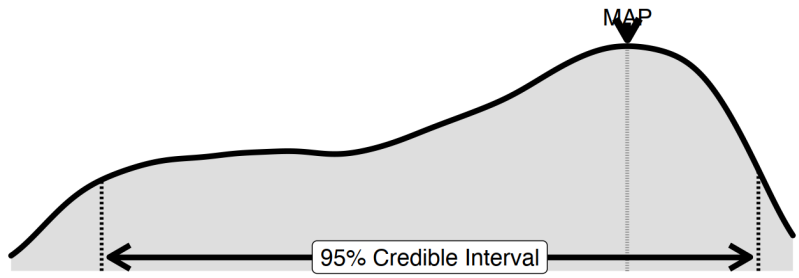

Figure 5.3: Describing the posterior distribution with an interval, for example the 95% credibility interval, or a point estimate, for example the maximum a posteriori estimation (MAP).

5.7 From Model to World

Bayesians build statistical models to approximate the data-generating process with probability distributions. These distributions are parameterized, and Bayesians say that these parameters are themselves random variables. Learning means updating the model parameters. But after the "update" there is no definite answer on what the true parameters are. Instead, the modeler was able to reduce the uncertainty about the parameters whereabouts. The posterior distribution doesn't encode the uncertainty inherent in nature, as in quantum mechanics. Instead, the posterior distribution expresses the uncertainty about **information about the world**. That's different from how frequentists connect their statistical models to the world. Frequentists assume that there are some unknown but fixed parameters that can be approximated with statistical models. Uncertainty is a function of the estimators, and conclusions about the world are derived from how these estimators are expected to behave when samples and experiments are repeated.

5.8 Simulate to Predict

Parameters are random variables. At first glance, that's a problem if we want to make predictions with the model. But it's actually a gift and not a problem. In the Bayesian mindset, predictions are also random variables, not just point estimates. The distribution of a prediction reflects the uncertainty of the parameters. Not getting a definite answer or prediction seems inconvenient at first. But it's much more honest and informative than a point estimate. The modeler in the Bayesian mindset is encouraged to consider the prediction along with its uncertainty.

To predict, the Bayesian must simulate. Prediction means marginalizing the distribution of the prediction for a new data point X_{new} with respect to the distribution of parameters:

$$P(X_{new}|X)_{\Theta} = \int P(X_{new}|X,\theta) \cdot P(\theta|X) d\theta$$

Simulation means that values for the parameters are drawn from the posterior distribution. For each draw of parameters, the modeler can insert these parameters into the probability distribution of the data and then draw the prediction. Repeating this process many times yields the posterior predictive density.

5.9 Strengths

- Bayesianism allows the use of prior information such as expert knowledge.
- Bayesian inference inherits all advantages of statistical models.
- Bayesian inference provides an expressive language to build models that naturally propagate uncertainty. This makes it easy to work with hierarchical data, measurement errors and missing data.
- A general benefit: Bayesian updating is an interesting mental model for how we update our own beliefs about the world.
- Bayesian interpretation of probability is arguably more intuitive than frequentist interpretation: When practitioners misinterpret frequentist confidence intervals, it's often because they interpret them as credible intervals.
- Since Bayesian inference always estimates the full posterior, decision based on the posterior always require another step. As a consequence, inference and decision making are decoupled. In frequentist inference, it's common to design the entire inference process around the decision (hypothesis tests).
- Bayesian statistics adheres to the likelihood principle which states that all the evidence from the data relevant to a quantity of interest must be contained in the likelihood function.

5.10 Limitations

- The choice of prior distributions is subjective.
- The modeler always has to define a prior which can be tedious when many priors are involved.
- Bayesian methods are mathematically demanding and computationally expensive.
- When used exclusively for decisions, all the additional information about the entire posterior may appear as unnecessary overhead.
- No causal interpretations are allowed, just associations are modeled.

6 Likelihoodism

- The likelihood function is all you need and is interpreted as evidence for a statistical hypothesis (law of likelihood)
- Statistical hypotheses are compared by the ratio of their likelihoods.
- A statistical modeling mindset with frequentism and Bayesianism as alternatives.

A frequentist, a Bayesian, and a likelihoodist walk into a bar, a wine bar. The sommelier quickly joins the three. The Bayesian wants to hear the sommelier's opinion first before trying the wines.. The frequentist asks the sommelier about the tasting process: is the number of wines fixed in advance? Is the tasting over when the customer has found a suitable wine? How are subsequent wines selected? The likelihoodist politely tells the sommelier to fuck off.

Frequentist inference has a long list of limitations. But it's still the dominant statistical mindset in science and elsewhere. Bayesian analysis has seen a resurgence thanks to increased computational power for sampling from posteriors with MCMC. But using subjective prior probabilities doesn't sit well with many statisticians. Could there be another way to "reform" the frequentist mindset? A mindset without the flawed hypothesis testing and without priors?

Welcome to the **likelihoodist mindset.**

I studied statistics for 5 years, worked as a statistician and data scientist for 3 years, and then did PhD studies in machine learning for 4.5 years. In those 12 years of statistics, I never learned anything about likelihoodism. It's fair to say that likelihoodism is the underdog. Likelihoodism leads a shadowy existence while Bayesianism and frequentism are engaged in an epic battle.

Likelihoodism is the purist among the statistical modeling mindsets. A mindset that focuses entirely on the likelihood function. Likelihoodism is an attempt to make statistics as objective as possible.

All three mindsets use likelihood functions in different ways. A quick recap: The likelihood function is the same as the data density function, but the roles of data and parameters are reversed. Data X are "fixed" and the likelihood is a function of the parameters θ $P(\theta; X) = P(X = x|\theta)$. The likelihood links observed data to theoretic distributions. Bayesians multiply prior distributions with the likelihood to get the posterior distributions of the parameters. Frequentists use the likelihood to estimate parameters and construct "imagined" experiments that teach us about long-run frequencies (hypothesis tests and confidence intervals). Likelihoodists view the likelihood as evidence derived from data for a statistical hypothesis. Likelihoodists emphasize the likelihood and reject the non-likelihood elements

from frequentism and Bayesianism: Likelihoodists reject priors because they are subjective; Likelihoodists reject the frequentists' reliance on "imagined" experiments because these never-observed experiments violate the likelihood principle.

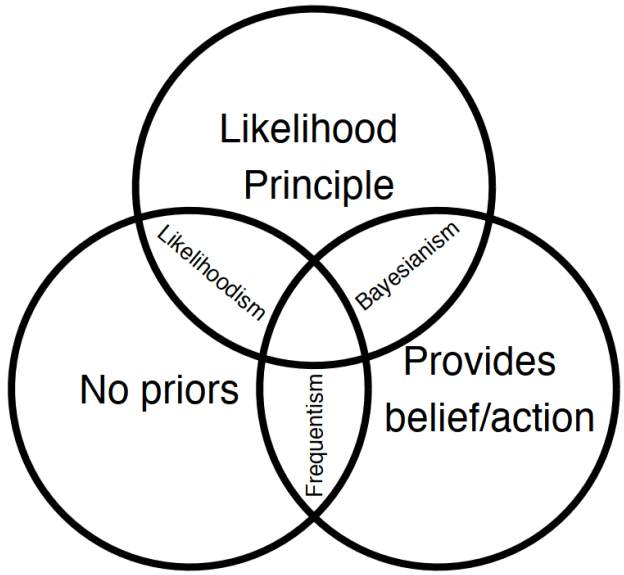

Figure 6.1: How Bayesianism, frequentism, and likelihoodism overlap and differ. Figure inspired by Greg Gandenberger: https://gandenberger.org/2014/07/28/intro-to-statistical-methods-2/.

But what is the likelihood principle that is so central to likelihoodism?

6.1 Likelihood Principle

"The likelihood principle asserts that two observations that generate identical likelihood functions are equivalent as evidence."[10] As a consequence, all evidence that comes from the data about a quantity of interest θ has to be part of the likelihood function $P(\theta; X)$. If we reverse the statement: If information from the data influences the analysis but is not part of the likelihood, then the likelihood principle is violated.

Let's say we want to estimate the average waiting time for a public bus. To model waiting times, the exponential distribution is a good choice. So we could assume that $X \sim Exp(\lambda)$, where λ can be interpreted as the inverse waiting time. The expected waiting time is $\frac{1}{\lambda}$. We have collected n waiting times $x_1, ..., x_n$. The likelihood function is:

$$P(\lambda; x_1, ..., x_n) = \lambda^n exp\left(-\lambda \sum_{i=n}^{n} x^{(i)}\right).$$

In all three mindsets, we could work with this likelihood. Bayesians would, in addition, assume a prior distribution for λ. Whether the likelihood principle is violated depends on what we do after calculating the likelihood. Bayesians obtain a posterior distribution for λ that is interpreted as belief about the parameter. The likelihoodist might report the likelihood region for λ, which can be interpreted as relative evidence for a range of parameter values compared to the maximum likelihood estimate for λ. Both the Bayesian and the likelihoodist approaches adhere to the likelihood principle: All evidence from the data about λ is included in the likelihood. Bayesians use priors, but as long as they don't include any information from the data, it's fine. [1]

The frequentist might test whether λ is significantly smaller than a certain value. When performing such a test, the frequentist has to "imagine" experiments under the null hypothesis distribution. But the null hypothesis is not part of the likelihood. Frequentists choose the distribution under the null hypothesis based on how they "imagine" repetitions of the sample or experiment. This, in turn, depends on how the experiment was conducted or how the data were collected in the first place. You will see later an example of a coin toss where the same data from different experiments lead to different conclusions in the frequentist mindset.

So the big difference between frequentism and likelihoodism is the likelihood principle. The likelihood principle gives the likelihood function the monopoly over data evidence. But the likelihood principle alone is not sufficient to create a coherent modeling mindset. We need the law of likelihood.

6.2 Law of Likelihood

The law of likelihood is the foundation for using the likelihood function alone formaking modeling decisions. The law of likelihood says[11]:

- Given:
  - Hypotheses H_1 and H_2; data $\mathbf{x} = \{x^{(1)}, ..., x^{(n)}\}$.
  - Likelihood for H_1 is larger than for H_2: $P_{H_1}(X = \mathbf{x}) > P_{H_2}(X = \mathbf{x})$.
- Then:
  - Observation $X = \mathbf{x}$ is evidence supporting H_1 over H_2.
  - The likelihood ratio $P_{H_1}(x)/P_{H_2}(x)$ measures the strength of this evidence.

[1] The likelihood principle is violated when data is used to inform the prior. For example, empirical priors which make use of the data violate the likelihood principle.

The hypotheses can be the same statistical model, but with different parameter values θ. Returning to the bus waiting time example, H_1 could be that $\lambda = 1$, and H_2 could be that $\lambda = 0.5$. The resulting likelihood ratio might be:

$$P(\lambda = 1; x_1, ..., x_n)/P(\lambda = 0.5; x_1, ..., x_n) = 4$$

.

The likelihood ratio is the likelihood of one statistical hypothesis divided by the likelihood of another. As a reminder, statistical hypotheses are statistical models where, optionally, some or all of the parameters are assigned by hand rather than learned from the data. The law of likelihood tells us, that to compare hypotheses H_1 and H_2 with their likelihood ratio:

$$\frac{P(H_1; X = \mathbf{x})}{P(H_2; X = \mathbf{x})}$$

In frequentism, likelihood ratios are often used as test statistics for hypothesis tests. In likelihoodism, the likelihood ratio is interpreted as evidence.

Likelihoodists may use a rule of thumb for judging the strength of evidence. For example, a likelihood ratio of 8 is considered fairly strong and 32 or more is considered "strong favoring".[10] In our example, a likelihood ratio of 4 in favor of $H_1 : \lambda = 1$ over $H_2 : \lambda = 0.5$ is not enough to be "fairly strong". H_1 and H_2 can also be more complex hypotheses, such as regression models with different covariates or assumptions.

The law of likelihood is stronger than the likelihood principle: The likelihood principle states that the all evidence from the data must be in the likelihood; **The law of likelihood describes how evidence can be quantified and compared.** And this is where Bayesian inference and likelihoodism differ. Bayesians are not guided by the law of likelihood, but by Bayesian updating and a subjective interpretation of interpretability.

The law of likelihood makes it clear how we can compare statistical hypotheses: Not by hypothesis testing, but by their likelihood ratios. The larger the ratio, the stronger the evidence for one hypothesis over another.

This is also where likelihoodism reaches a dead end. The ratio may only be interpreted as evidential favoring. The likelihoodist mindset doesn't come with guidance on what we should believe about the parameters or what decision/action to take based on the results. The likelihood ratio only tell us which hypothesis is favored.

6.3 Likelihood Intervals

Likelihood intervals are the likelihoodist analogue to frequentist confidence interval and Bayesian credible intervals. Likelihood intervals are interpreted in terms of, you guessed it, relative likelihood. The likelihood interval of a model parameter θ is the set of all θ values that yield a relative likelihood greater than a certain threshold:

$$\left\{\theta : \frac{L(\theta|X)}{L(\hat{\theta}|X)} \geq \frac{p}{100}\right\}$$

The $\hat{\theta}$ is the optimal θ after fitting the model using maximum likelihood estimation or another optimization method. Let's say for a logistic regression model coefficient β_j: $\hat{\beta}_j = 1.1$. Then an interval could be $[0.9; 1.3]$. The role of the constant p is similar to the one of the α-level for confidence and credible intervals: It specifies the size of the interval. See figure 6.2. Each θ-value within that interval can be seen as constituting a different hypothesis. And these hypotheses are compared with the optimal model $\theta = \hat{\theta}$.

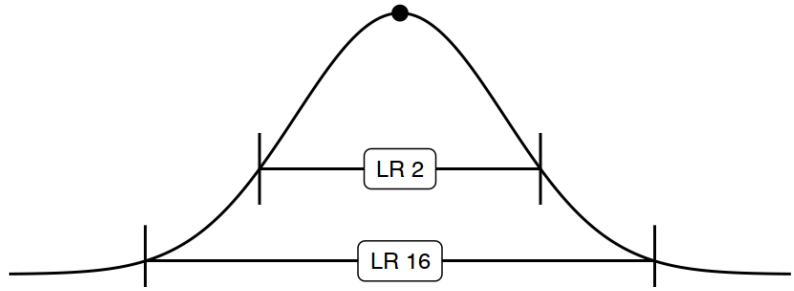

Figure 6.2: 1/2 and 1/16 likelihood ratio intervals.

6.4 Why Frequentism Violates the Likelihood Principle

Many frequentist methods violate the likelihood principle because they require "imagined" experiments. But frequentists need these theoretic distribution to compute p-values, for example. Statistical null hypothesis tests also depend on stopping criteria for data collection, which is in conflict with the likelihood principle.

The following example shows how frequentism violates the likelihood principle because of different stopping criteria. Suppose we have a coin. We want to find out whether it's fair, or whether maybe head turns up too often. I still don't know why statistician's are so upset about unfair coins, but that's the way it is. We call θ the probability of head. We have two hypotheses:

$$H_0 : \theta = 0.5 \text{ and } H_1 : \theta > 0.5$$

H_0 means that the coin is fair. H_1 claims that head comes up more often than tail. We define two random variables: the number of heads X, and the number of coin tosses Y.

We perform two experiments with a different setup but with the same results:

1. Toss the coin 12 times. We observe head 9 out of 12 times.

2. Toss the coin until tail is observed 3 times. The third tail appears on the 12th toss.

Both experiments have the exact same outcome, but we defined the stopping criteria for the experiments differently: In experiment 1), we fixed the number of tosses Y. In experiment 2), we fixed the number of heads X. Should we reach different conclusions about the fairness of the coin? What do you think?

Both experiments give the same likelihood, up to a constant factor. Experiment 1):

$$L(\theta|X=3) = \binom{12}{3}\theta^3(1-\theta)^9 = 220\theta^3(1-\theta)^9$$

And experiment 2):

$$L(\theta|Y=12) = \binom{11}{2}\theta^3(1-\theta)^9 = 55\theta^3(1-\theta)^9$$

So the likelihoodists say that both experiments carry the same evidence. The likelihood intervals would be the same for both experiments.

Frequentists would come to different conclusions depending on the experiment. Frequentists include results that have not occurred but depend on how the experiments are conducted. They assume that H_0 is true and infer how the test statistic is distributed in future experiments under H_0. Then frequentists place the estimated value of $\hat{\theta}$ within this distribution of imagined experiments and see how extreme the result is. In experiment 1), where the number of tosses is fixed, experiment outcomes of 9, 10, 11, or 12 heads are more extreme than the actual experiment outcome. In the other experiment, the number of tails is fixed. More extreme outcomes in experiment 2) are all possible experiments where we observe more than 12 tosses. This includes, for example, the experiment where the third tail only comes up after 1108 tosses.

When we test H_0 vs. H_1 in experiment 1), we get:

$$P_{H_0}(X \geq 9) = \sum_{x=9}^{12}\binom{x}{12}0.5^x(1-0.5)^{12-x} = 0.073$$

At a significance level of $\alpha = 0.05$, we would not reject the fair coin hypothesis.

For experiment 2), we assume a negative binomial distribution:

$$P_{H_0}(Y \geq 12) = \sum_{y=12}^{\infty}\binom{3+y-1}{2}0.5^y0.5^3 = 0.0327$$

The p-value is now smaller than 0.05, and with that the coin is significantly unfair.

In frequentist inference, the way data are collected and the way experiments are designed affect the results.

This has much more subtle consequences than I've illustrated so far. Imagine a domain expert asks you to perform an analysis with 1000 data points. As a frequentist, you need to know how those 1000 data points were collected. What was the stopping criterion for data collection? If the domain expert only planned to collect 1000 data points, that's fine. But if the expert says she would collect more data depending on the outcome of the analysis, then that changes the analysis, which is a violation of the likelihood principle.

6.5 Strengths

- Likelihoodism inherits all the strengths of statistical models.
- It's a coherent modeling approach: all information is contained in the likelihood. Frequentism, in contrast, is more fragmented with long lists of differently motivated statistical tests and confidence intervals.
- Like Bayesian inference, likelihoodist inference is also consistent with the likelihood principle. Therefore neither is affected by experimental design, as is the case with frequentism.
- Likelihoodism is arguably the most objective of the statistical modeling mindsets. No priors, no imagined experiments.
- Likelihoodist ideas can improve the reporting of Bayesian results. For example, Bayesians can additionally report likelihood ratios as evidence.
- A significance test might reject H_0, even if the evidence for H_0 is greater than for H_1. Likelihoodism doesn't have this problem.

6.6 Limitations

- Likelihoodism doesn't provide guidance in the form of belief or decision. Evidence is less practical, and the statistician has no certainty about which the final model is and how to work with it. This is the strongest argument against likelihoodism, and maybe the reason why we don't see it in practice.
- To be more specific: There is no theoretical underpinning for saying when there is enough evidence to choose one hypothesis over another.
- With a likelihoodist mindset, we can can only compare simple hypotheses where all parameters are specified. Composite hypotheses for ranges of parameters are impossible. Likelihoodism can't compare $\theta > 0$ versus $\theta \leq 0$. Only, for example $\theta = 1$ against $\theta = 0$.
- Likelihoodism allows only relative statements. It can't state the probability that a statistical hypothesis is true – only how its evidence compares to another hypothesis.

6.7 Resources

- The book "Statistical Evidence: A Likelihood Paradigm" is good introduction to likelihoodism, if you have some background as a statistician.
- I've found the Greg Gandenberger's blog[2] super helpful in learning about likelihoodism. He takes a more philosophical viewpoint and argues against likelihoodism and for Bayesianism. This critique is most detailed in his essay "Why I am not a likelihoodist".[12]

[2] http://gandenberger.org/

7 Causal Inference

- A model is a good generalization of the world if it encodes causality.
- Causal models connect random variables through directed cause and effect relationships.
- Causal inference is a two-step process in which first a causal model is assumed and constructed and then translated into a statistical or machine learning model.

"Thank you so much for this statistical model," the ecologist says to the statistician, and continues, "Nice p-values, and insightful findings! Would it be correct to say that the droughts **caused** the crop failures?" The statistician looks at the ecologist, a hint of concern at the corner of the eyes. As if practiced, the statistician says: "Correlation does not imply causation". Unsatisfied, the ecologist responds: "But it would make so much sense to conclude that the drought was the cause!" The statistician grimaces and clenches the teeth as if in pain. "Correlation does not imply causation", the statistician repeats, the words having a weird melody, as if in prayer. "But without a causal interpretation, how can the results be applied to advance science? I want to understand **why** the crop failures happened!", the ecologist insists. "Correlation does not imply causation. Correlation does not imply causation. Correlation ...", the statistician now chants, eyes closed shut, face twisted as if in great pain. The ecologist slowly retreats, shocked by the strong reactions of the statistician. Sometimes at night, when the wind howls outside, the ecologist hears the statistician's mantra in the wind.

7.1 Does The Drug Help?

A while back, I worked with a rheumatologist on an important medical question: Do TNF-alpha blockers reduce the long-term symptoms of patients with axial spondyloarthritis, a chronic disease that is associated with inflammation of the spine. In the long-term, the joints in the spine can fuse to due to new bone formation (ossification). TNF-alpha blockers, given regularly as injection or infusion, work really well to reduce inflammation. To understand whether TNF-alpha blockers also help against the ossification, a clinical trial might have given the best evidence. But withholding TNF-alpha blockers would be unethical due to their proven efficacy, and also the study would require a long-term observation. The next best option was to use observational data from hospitals and medical practices. The registry for rheumatic patients I was working for maintained a huge data base of patients with axial spondyloarthritis, holding insightful data about the patients health-related history: doctor visits, blood values, x-ray images, and so on. In collaboration with the rheumatologist I built a statistical model to answer whether TNF-alpha blockers help against ossification. For these patients, we had x-ray images two years apart, from which radiologists had scored the

progression of the new bone formation in the spine. To predict the progression, the model included various variables that were measured at the time of the first x-ray: the age of the patients, the disease duration, inflammation levels, medication used, and so on. The result of the analysis was that the drug didn't reduce ossification. The lead statistician of the patient registry happened to participate in a course on causal inference around the same time. She had the epiphany that we approached the modeling question the wrong way. She drew a diagram visualizing how the drug, the inflammation and the ossification might be related. She drew the graph like this (Figure 7.1:

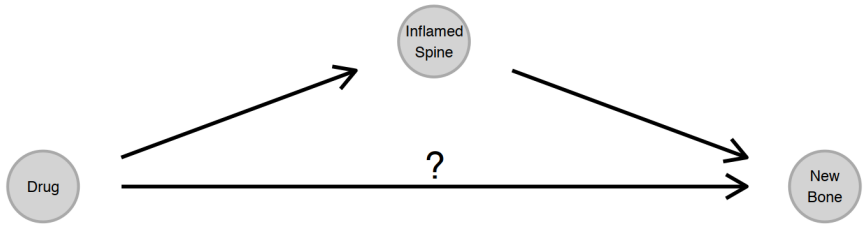

Figure 7.1: The drug was known to influence inflammation (reduce it). Inflammation was thought to cause ossification (new bone formation). More inflammation, more ossification. The drug has, potentially, two pathways to reduce ossification: directly, or indirectly via reducing inflammation.

It became immediately clear what the problem with our current model was: Inflammation was a potential mediator of the effect of TNF-alpha blockers on long-term ossification. Figure 7.1 shows that the effect of the drug can be split into a direct effect and an indirect effect. By having the inflammation level as part of the model, the indirect effect of the treatment was "swallowed". Any changes to the ossification based on reducing inflammation was reflected in the coefficient for inflammation levels. The way we set up the initial model, the causal interpretation of the coefficient for TNF-alpha blockers concerned only the direct effect. But we were actually interested in the total effect. The total effect is the direct effect of the medication plus all indirect effects, in this case indirectly via reduction of inflammation. So we adjusted the model accordingly, removing the inflammation variable. [1], so that the coefficient for TNF-alpha blockers would contain the total and not only the direct effect. Now the model clearly showed that TNF-alpha blockers reduce ossification by decreasing the inflammation levels. It sounds like common sense in hindsight, but coming from a frequentist mindset, my mind was blown. This moment was such a revelation and got me interested in causal inference.

[1]The attentive reader might object that I referred to inflammation now as both confounder and mediator. Both is correct, if we distinguish different time points. The initial model had inflammation after treatment begin as variable, so it acted as mediator. We later also adjusted the model for inflammation before treatment, when it acts a confounder.

7.2 Causality

We all have an intuition about causality. Rain is a possible cause for a wet lawn. A drug can be a cause of getting healthy. An environment policy can be a cause of reduced CO2 emissions. In terms of random variable, we can express causality in terms of distributions and interventions: If you **force** a random variable to have take on a certain value, how would the distribution of another random variable change? A cause is different from association: An association is only a statement about observation. We know that having a wet lawn does not *cause* your neighbours lawn to be wet. How do we know it? Try watering your lawn for one year, every day, and see whether the probability for your neighbors lawn being wet has changed. But the two lawns two are associated: When you observe that your lawn is wet, the probability that your neighbors lawn is wet is high. The reason for association is, of course, the rain. Such shared causes are called confounders.

The archetypal statistician avoids speaking of causality. At least that's my experience after getting a Bachelor and Master in statistics. What I learned about causality in those 5 years can be summarized in two statements: 1) Always add all confounders when building a statistical model, and 2) correlation does not imply causation. We were taught not to causally interpret statistical models. We were taught to ignore the elephant in the room. Causality was presented as an unreachable goal that should not even be attempted with statistical modeling. We were taught how to dance around the topic. A random variable can only be associated with the outcome, but we may not speak about the variable causing the outcome.

"Correlation does not imply causation" truly is a mantra that you hear multiple times when you learn about statistics. I find that very curious, especially given that statistical modeling is supposed to be THE research tool of our times. Isn't research all about detecting how the world works? The "how", at least for me, implies that scientists are supposed to uncover causal structures. And the truth is that, in the end, the results are, very often, interpreted causally, by the domain experts, by lay persons, and by the media. So shouldn't everyone at least attempt to make the model reflect causality as much as possible? Fortunately, some people think that we should put causality first.

Welcome to the **causal inference** mindset.

7.3 The Causal Mindset

The causal inference mindset puts causality in the center of modeling. The goal of causal inference is to identify and quantify the **causal** effect a random variables had on the outcome of interest.

I would say that causal inference is a statistical modeling mindset, because it relies on probability distributions and random variables. Causal inference could also be seen as an "add-on" to other mindsets like frequentist or Bayesian inference, but also for machine learning. But it would be wrong to think causal inference as just a cherry on top of other mindsets. It's much more than just adding a new type of method to another mindset, like adding support

vector machines to supervised learning. Causal inference challenges the culture of statistical modeling. It requires the modelers to think more about the data-generating process, to be explicit about causes and effects.

It's kind of surprising just how many models are "broken" because they ignore causal thinking. A lack of causal considerations can mean that the analysis of a research paper is invalid or that a machine learning model in a product is vulnerable to changes in the data distribution or adversarial attacks. Take Google Flu prediction model as an example. Google predicted outbreaks of the flu based on frequencies of certain search terms. Clearly, the prediction model was not a causal model. If it were causal, it would mean that you can cause flu outbreaks by searching on Google for certain terms. The flu detection model missed, for example, the nonseasonal 2009 flu.[13] The machine learning model quickly declined in performance because the search patterns changed over time. No causalist would have signed off on such a non-causal model. A model that relies on only associations is as ephemeral as a fruit fly. A model only generalizes well when it encodes causal relationships. A causal flu model might rely on the virulence of the current flu strains, the number of vaccinated people, predictions of how cold the winter would be and so on. But never based on search terms.

You can look as hard at the data as you want to, but it won't reveal the causal structures that produced it. You can automatically infer associations from the data, but even the simplest causal structures are ambiguous. The amount of sunshine on a given day can be considered causal for the number of park visitors. In a dataset, both features would appear as columns with numbers in it. And if we would compute the correlation, we would find out that sunshine and park visitors are positively correlated. The more sun, the more people. The more people, the more sun. The causal relationship is clear: The sun couldn't care less about park visitors. Instead, the sun is the cause of park visits. But this causal direction is not clear for your computer. No matter what choose as a target, the computer will oblige and fit the model. Breaking news: The government forbid visits to the park, in an effort to cool down the current heat wave. The causal mindset requires thinking even more about the data-generating process, making assumptions about causal relationships. These assumptions are an attack surface to criticize the mindset. But on the other hand, making these assumptions explicit allows to address and discuss deviating opinions about causal directions.

And the best way to make causal structures explicit is the directed acyclic graph.

7.4 Directed Acyclic Graph

Causal inference comes with a tool to visualize causal relationships: Directed Acyclic Graphs, or short, DAGs. A DAG, such as the one in Figure 7.2 makes it simple to understand which variable is a cause to another variable. Variables are visualized as nodes and the causal direction is visualized with an arrow. DAGs have to be acyclic, meaning arrows are not allowed to go in circles. For example, adding an arrow from Y to X_1 in Figure 7.2 would make the DAG cyclic and most causal frameworks can't handle that.

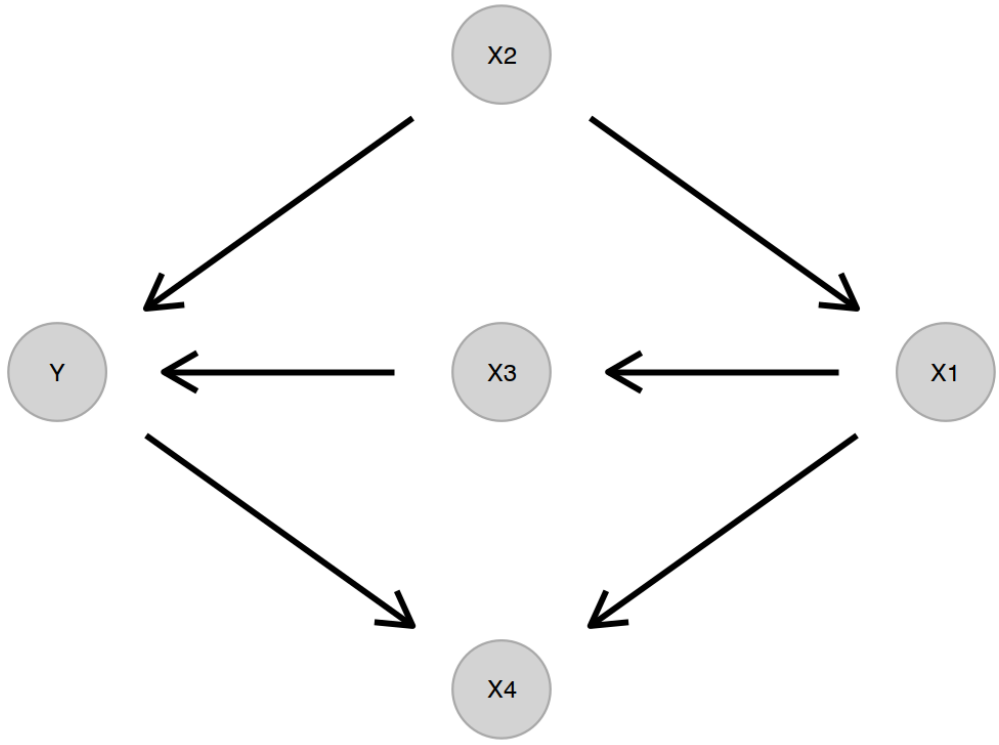

Figure 7.2: A directed acyclic graph (DAG) with 5 variables.

What can we see from the DAG in Figure 7.2? Variables X_2 and X_3 are direct causes for target Y. X_1 only indirectly influence Y via X_3. And X_4 is not a cause of Y, but instead Y, together with X_1 causes X_4.

But how do we know where to put arrows, and in which direction to point them?

- Good old common sense, such as knowing that park visitors can't be the cause of more or less sunshine.
- Domain experts can often tell you in more specific cases what the most up to date knowledge about causal directions is.
- Direction of time: We know that the elevator comes because you pushed the button and not the other way around.
- Causal structure learning: To some degree, we can automatically learn causal structures. But usually this results in sets of DAGs that might be plausible, leaving the user still with putting in some assumptions.

Building a causal model, be it in the form of a DAG or otherwise, requires to make assumptions. These assumptions might not be testable, and therefore a subjective choice of the modeler. That's what causal inference gets criticized for as well, even to the degree that

causality can't be known. But even if we can't know for sure if we got the right causal model, we can at least sit around a DAG, and point the fingers at the arrows we disagree with.

7.5 Many Frameworks For Causality

There are many "schools" or frameworks of causal inference, each with their own notation and approaches.[14] That's what I found to be the biggest entry barrier to the causal inference mindset. If you want to get into, say Bayesian inference, you can take any Bayesian introduction book and they will share a canon, a common set of tools and a shared language. But for causal inference, the field is much more split. So don't despair too much, it's not you, it's the causal inference field. Anyhow, here is a short overview of approaches. The overview is far from exhaustive, but should give you a better impression of what's out there:

- A huge part of causal inference is more about designing experiments rather than causal model for observational data, such as clinical trials or A/B-tests. Claims to causality are derived from randomization and intervention instead of causal modeling.
- Sometimes observational data can also have the character of an experiment, which is often called "natural experiments". When John Snow investigated cholera, he had access to data from a natural experiment. John Snow identified contaminated drinking water as the source of cholera, because the customers of one water company got sick of cholera much more often than customers from the other. The association of households to one company or the other served as a natural experiment, randomizing factors such as age, comorbidities, education, and so on.
- Propensity score matching attempt to to estimate the effect of an invention, such as a treatment, by matching data points to account for differences in other variables.
- Probably the most general and coherent framework of causal inference is by the statistician Judea Pearl. This "school" includes the do-calculus[15], structural causal models, front- and backdoor criteria and many other tools for causal inference.[16]
- The **potential outcomes framework**[17] is another larger causal "school", mostly used for studying causal effects of binary variables.
- Causal discovery or structure identification is a subset of causal inference that aims to discover causal relationships from merely observational data.
- Then there are so many individual methods that aim to provide causal modeling. One example is "honest causal forests", which are based on random forests and designed to model heterogeneity in treatment effects.[17]
- …

All approaches have in common that they assume a causal model. This causal model can be very explicit, for example if it involves drawing a DAG. But it could also be more hidden within the assumptions of some method about which variable to include in the model and so on. The final estimate, however, is always a plain statistical estimator, or a machine learning model or so. But how do we get from a causal model to a statistical estimator?

7.6 From Causal Model to Statistical Estimator

For this section, we assume that, for some reason, we can't do an experiment. Instead, we have observational data, for which we want to do causal inference. With observational data, the first casualty is causality – at least from the point of view of non-causalists. Observational data is when causalists get excited and start stretch their hand-wrists to warm up for all the DAG-drawing and modeling.

Causalists claim that you can estimate causal effects, even for merely observational data. I am willing to reveal their secret: Causalists use large hadron colliders to collide particles with high energy – producing black holes in the process. Each black hole contains a parallel universe that lets them study what if scenarios. Joke aside, there is no magical ingredient for estimating causal effects. Causal modeling is mostly a recipe to translate causal models into statistical estimators:[16]

1. Formulate causal estimand, like: What is the causal effect of a medication on disease cure?
2. Build a causal model: More concretely, identify relevant variables, draw a DAG and so on.
3. Identify whether causal effect can be estimated.
4. Estimate the (translated) target quantity.

Let's have a look at the individual steps. The first question is, what causal relationship we want to study. This can then be expressed as some causal estimand, for example how would the air pollution be reduced when we ban car traffic in the inner city. By choosing the target and the possible cause, we can now proceed to building the causal model.

The causal model can be build using visual tools like the DAG. Besides the target and the potential causal variable, all other variables that are relevant to both should be included in the causal model. It's kind of collecting all the nodes for the DAG. But we also need the arrows that connect the variables, with the direction of the arrows showing the causal direction. Identifying between which variables an arrow should be and in which direction it should point can be narrowed down, in parts, automatically. But in the end, a lot of those causal directions will be based on domain knowledge and subjectivity. But in the end, we do have a DAG. Not all approaches and frameworks will necessarily encourage or require to draw such a DAG, but you always have to decide on what the confounders are and so on.

In the identification step the causalists find out whether the causal estimand can be even answered with the observational data at hand. This means that the causalist has to check whether the assumptions of the causal inference hold, given the causal model. Not all causal effects can be estimated. For example, we want to measure the causal effect a treatment has on a health outcome. Both the decision to treat and the health outcome might be influence by a third variable, such as the education status of the patient. Education, in this example, is a classic confounder. If we haven't measured education, then we can't reliably estimate the causal effect of the treatment. Identification can be a complicated process. But there are also many "simple" rules that tell you how to turn a causal estimand into a conditional estimand.

Identification boils down to a selection of variables to adjust for in the statistical estimate. To give you an idea, here are a few simple rules:

- Include all confounders. Confounders are variables that are cause to both the variable of interest and the outcome. For example in Figure 7.3 X_2 confounds X_1 and Y.
- Exclude colliders. In DAG in Figure 7.3, X_4 is a collider for Y and X_1. Adding colliders to a model opens an unwanted path.
- Exclude mediators. When we want to measure the causal effect of X_1 on Y, we have make sure not to include X_3. Including X_2 would block the path between X_1 and Y, and we would falsely find that X_1 does not influence Y.

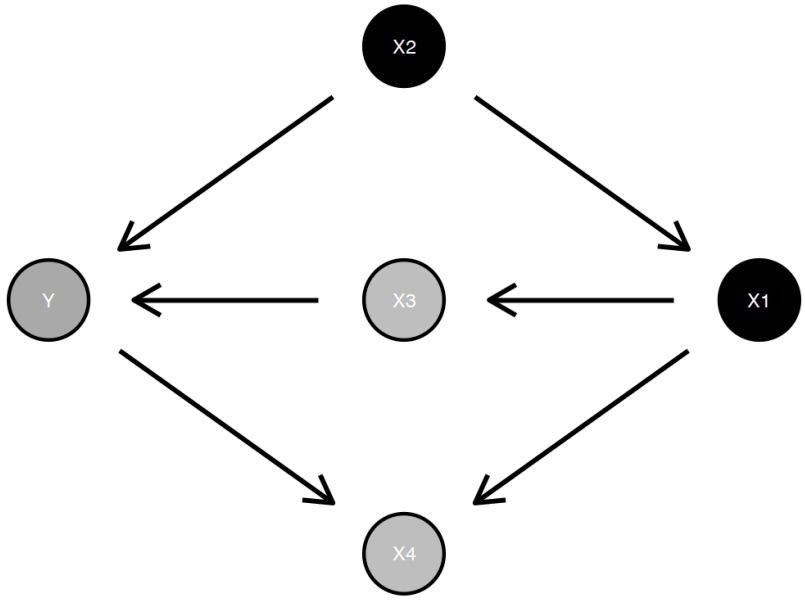

Figure 7.3: To understand the causal effect of X1 on Y, we have to build a regression model with Y as target and X_1 and X_2 as predictor variables. Roles: Y is the target, X_1 the variable of interest, X_2 a confounder of X_1 and X_2, X_3 a mediator to the effect of X_1 on Y, and X_4 is a collider.

For the estimation, we have to make assumptions how the random variable are distributed. But mostly, we can just estimate and interpret our model whatever mindset we are coming from, be it frequentist, Bayesian, likelihoodist inference or machine learning. Also the interpretation of parameters is as usual. But there is this additional benefit that we now may interpret effect as causal.

All of the causal modeling steps have to be potentially repeated if we are interested in the causal effect of another variable. This is because, the identification might lead to a different set of variables that we have to adjust our model for.

7.7 Strengths

- Causality is central to modeling the world, and causal inference is **the** mindset to embrace that fact.
- I believe most modelers actually want causal model. Clearly, scientists want causal explanations to understand the world better. But also in industry, such as marketing, you want to understand how actions causally affect outcomes.
- Only causal models will generalize well, because they are more robust against changes in the environment. Or rather: Non-causal models break more easily, since they are built on associations.
- Causal inference is a rather flexible mindset that enhances many other mindsets such as frequentism, Bayesianism, machine learning.
- DAGs make causal assumptions explicit. If you only have one take-away from this chapter, or from causal inference in general, it should be DAGs as a method to think and communicate.
- You might say that causal modeling with observational data is not possible. The truth is, that models, once out of the hand of the modeler, will in many cases be interpreted causally. Then why not make an effort to introduce some of the best practices from causal inference?

7.8 Limitations

- Many modelers stay away from causal inference for observational data, saying that causal models are either not possible or too tricky.
- Confounders, causes of both variable of interest and target are especially tricky. For a causal interpretation, you have to assume that you found all the confounders. But you can't prove that you have identified all confounders.
- There are many schools and approaches to causal inference. This can be very confusing for people entering the field.
- Causal modeling requires subjective decisions. The causalist can never be sure whether the causal model is correct.
- Using non-causal variables can enhance predictive performance, but make your model non-causal. So causal inference and predictive performance are at odds with each other.

7.9 Further Reading

- Free book: Causal Inference: What If[14]

8 Machine Learning

- Machine learning is a mindset concerned with making a computer "intelligent".
- A good machine learning model solves a task well: external evaluation of task performance is more important than the internal validity of the model.
- Machine learning is an alternative meta mindset to statistical modeling.
- Supervised machine learning, unsupervised machine learning, reinforcement learning, and deep learning are specializations of the machine learning mindset.

It's likely that you've used a machine learning product today. Maybe you have asked your smart assistant to read out your schedule for today, used a navigation app to get from A to B, or checked your mostly spam-free e-mails. In all of these applications, machine learning is used to make the product work: speech recognition, traffic jam prediction, and spam classification are just a few examples of what machine learning can do.

8.1 One or Many Mindsets?

Machine learning is the branch of artificial intelligence that deals with learning models directly from data. The computer improves at a given task through "experience" which means learning from data. The machine learning mindset doesn't tell you **how** the computer should learn from data. For example, machine learner may use random variables, but they don't have to. They can work on a prediction model where it is clearlt of defined when the model is correct, or they can work on clustering where the model is harder to evaluate. The models can be neural networks, decision trees, density estimators, statistical models and many more. Given this wide range of tasks, and without strict guidelines on how the computer must learn: Can we really say that machine learning is a distinct mindset? To answer the question, let's first look at more specific mindsets within machine learning. Machine learning is usually divided into supervised, unsupervised and reinforcement learning. Each of these subsets also represents a distinct modeling mindset: They involve a particular view of the world and of the relationship between the models and the world. The supervised machine learning mindset frames everything as a prediction or classification problem and has a clear definition of what a good model is: when it generalizes well to new data. In unsupervised machine learning, the goal is to find patterns in the data. The reinforcement learning mindset views the world as dynamic, in which an actor interacts with an environment guided by a reward. In deep learning, all models are neural networks. What are the commonalities between all these mindsets? Is there a unified machine learning mindset?

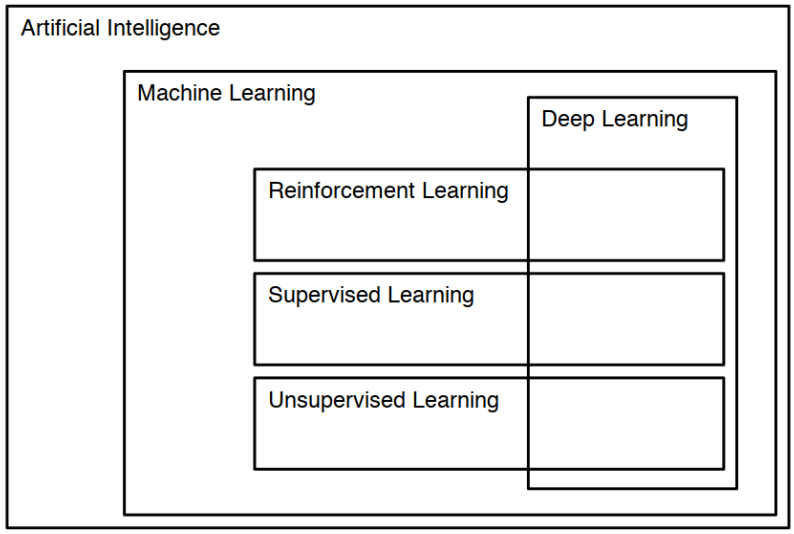

Figure 8.1: Machine Learning is a subfield of artificial intelligence. Within machine learning, there is supervised, unsupervised and reinforcement learning. Deep learning approaches overlap with these three.

The machine learning mindset may not be as unified and principled as statistical modeling. But all machine learning approaches have a few things in common. Let's take a look at what makes a good machine learning model and how these models relate to the real world.

8.2 Computer-Oriented, Task-Driven and Externally Motivated

Like all modeling mindsets in this book, machine learning is based on learning models from data. As the name implies, machine learning focuses on the "machine", meaning the computer. Machine learning is about the computer "learning" to solve tasks such as prediction, recommendation, translation, and clustering. How is this different from the work what statisticians, who also rely on computers? The motivation for using a computer differs between an archetypal statistician and an archetypal machine learner. The statistician uses the computer out of convenience and necessity. Modern statistics wouldn't be possible without the computer. But the computer is not the starting point. The starting point is statistical theory. And the computer is only a tool to apply statistical theory to data. [1] The machine learner,

[1] There is a field called computational statistics, which is computer-oriented. But we are talking about archetypes of mindsets here. You can think of computational statistics as a statistical mindset that is slightly infused with the machine learning mindset.

in contrast, starts with the computer. The machine learner says, "We have this new thing, the computer. How can we get it do intelligent and useful things?"

Machine learning can be understood as a meta-algorithm: An algorithm that uses data to produce machine learning models that are also algorithms. From a programmer's point of view, machine learning is a paradigm shift: Machine learning is just a way of "learning" an algorithm from data, rather than programming it directly. [2]

In contrat to more insight-driven statistical mdoeling, machine learning is typically used to solve a task. The task may be language translation, image captioning, classification, and so on. The success of the model is measured by how well the task was solved using some type of metric. In prediction and classification tasks, the machine learner measures the generalization error for new data. Specifically, for a classification model, this could be the accuracy with which the model assigns classes correctly in a new data set. In clustering tasks, success metrics usually measure how homogeneous the data points in the clusters are and how much the data differs between clusters. This external focus is also reflected in the way machine learning research works: Researchers invent a new machine learning algorithm and show that it works by comparing it to other algorithms in some task benchmarks. The reason why the algorithm works well is often discovered in later scientific publications, if at all.

We can distinguish between external and intrinsic modeling motivation. The motivation and evaluation of a machine learning model is external, based on task performance. It's like judging food based on how it tastes. Statistical modeling is intrinsically motivated. The rationale for constructing the model is important. It's like judging food not only by how it tastes, but also by the cooking process: did the chef use the right ingredients? Was the cooking time appropriate, and so on.

8.3 Strengths

- Task-oriented and very practical.
- A job in machine learning potentially pays you lots of money.
- A computer-oriented mindset in a computer-oriented world.
- Machine learning is predestined for automating tasks and building digital products.

8.4 Limitations

- Not as principled as statistical modeling.
- A confusing amount of approaches with different motivations and technical bases.
- A model that works well in solving a task is not necessarily a good model for insights. A model that predicts diabetes well can be useful, but is less insightful than a statistical model that models diabetes risk explicitly and understandably.

[2] I find it difficult to say that the machine learns by itself. Because machine learning also requires programming. You have to implement the learning part and all the glue code to integrate the final model into the product.

- Often requires a lot of data and is computationally intensive.

9 Supervised Learning

- Prediction-focused mindset that invites automation and competition.
- A good model has low generalization error - it predicts unseen data well.
- A machine learning mindset.

9.1 Competing With the Wrong Mindset

It was 2012, and I had just fitted a statistical model to predict whether a patient would develop type 2 diabetes given some risk factors. And now it was time to test the model. You see, I wasn't the only one modeling diabetes: I was competing with many other data scientists. I uploaded the CSV-file with the prediction results to the competition website. A table with two columns: One with the patient identifier and one with the probability for that patient to develop diabetes. One row per patient. Fingers crossed. But then came the disappointing results. The predictions of my model sucked. What had happened?

At the time, I was a master's student in statistics. I modeled diabetes risk using a generalized additive model, a model often used in statistical modeling. Most importantly, I created the model coming from a frequentist modeling mindset. So I thought a lot about the data-generating process, manually added or removed variables, and evaluated the model based on goodness of fit on the training data. The statistical modeling mindset failed me in this prediction competition. And that what confused me at first. After all, statistical models can be used for prediction and classification, and the same statistical models are also used in machine learning. Heck! Statistical learning is even one of the foundations of machine learning! This overlap of theory and methods may mislead one to believe that statistical modeling and supervised machine learning are interchangeable. But the (archetypal) modeling mindsets are fundamentally different, especially the idea of what makes a good model and how evaluation works. For me, the disappointing model performance was a catalyst for understanding the supervised machine learning mindset. For the diabetes competition, I began to seriously study machine learning models like boosting and random forests, but also how to properly evaluate the performance of machine learning models. While I didn't win any money in the competition (59th place out of 145), I did win something more valuable: With supervised machine learning, I gained a new modeling mindset.

9.2 Predict Everything

In supervised machine learning, everything is a prediction task. Before complaints come rolling in, here is my definition of prediction: The proposition of values that are unknown at a given time, but for which a ground truth exists or will exist. Assigning data points to a cluster is not prediction because there is no ground truth for the clusters. Prediction can mean assigning a classification score, a numerical value (regression), a survival time, etc.
It's amazing how many applications can be formulated as prediction tasks:

- Credit score can be expressed as the probability that someone will repay their loan. Based on information about the person's financial situation, a predictive model assigns a score that indicates how likely it is that the person will pay back the money.
- Predictive maintenance: Many machines require regular inspection and repair. Supervised machine learning models can be used to predict when machines might fail based on current conditions.
- Demand forecasting: using historical sales data to estimate demand for a product.
- Image classification: how should the image be classified? For example, image classification can be used to detect cancer on CT images.

As these examples show, supervised machine learning adopts the "task-oriented" trait of the machine learning mindset. Prediction is a task and can be used to do practical things. A modeling mindset that deals only with prediction tasks seems very narrow. But there is a surprisingly large number of applications for which prediction can be useful. And the type of data that can be used in predictive models can also be quite diverse: The input to the predictive model, usually called features, can be text, an image, a time series, a DNA sequence, a video, a good old Excel spreadsheet, …

9.3 Supervised Machine Learning

Turning any modeling task into a prediction problem is not the only defining trait of the supervised machine learning mindset. A core idea of supervised machine learning is risk minimization. And a good supervised model has a low generalization error, meaning that the prediction for new data points is close to the respective ground truth. To quantify how close a prediction is to the ground truth, the machine learner uses a loss function $L(y, f(x))$. The loss function L takes the ground truth value y and the predicted value $\hat{y}$ and returns a number. The larger the number, the worse the prediction. In the diabetes example, y could be 1 for diabetes and 0 for healthy. Accordingly, $f(x)$ could be the predicted diabetes probability between 0 and 1.

The goal in supervised machine learning is now to find the function f that minimizes the loss across the data:

$$\arg\min_{f} L(y, x, f(y))$$

The focus here is on optimizing the loss, and there are no specific constraints on what the function f may look like. In statistical modeling, f would have to be motivated based on probability distributions, but in machine learning, any form is allowed. This makes supervised learning a true machine learning mindset: The modeling approach is externally motivated by how the model predictions performs on new data. The model is trained using one part of the data (training data) and evaluated on another part (test data).

Is there enough to the claim that supervised machine learning is its own mindset? I believe that supervised machine learning is mindset of its own. The reason sounds very "bureaucratic", but it has strong implications on the mindset: supervised machine learning ALWAYS requires a ground truth. That's also what separates supervised from unsupervised and reinforcement learning. We want the model to predict diabetes? For the training data, we actually need to know if a patients have diabetes. The model is supposed to predict machine failure? We need a data set where we have actually observed many machines, some of which have also failed at some point.

The archetypal supervised learner wouldn't even consider working on unsupervised learning. For example, I know many machine learning researchers who work exclusively with supervised machine learning. There is no ground truth, so what the heck should the model "predict" anyway? And even if we defined something that the model should "predict", without ground truth we wouldn't really know how to evaluate it properly. My observation is that in industry people are more pragmatic and it would be harder to find a pure supervised machine learning because there are many problems without labels.

9.4 Learning Is Searching

We have danced around the question of what the function f is; the function that maps from the features x to the desired values y. Without any restrictions on the form of f, finding the best or at least a good f can seem infeasible. In statistical modeling it's "simple": We can derive estimators for f from the theoretical distributions. This makes the search space much smaller, and searching f is simplified to finding the best parameterization of a statistical model. In cases such as the linear regression model, we can even be sure that we have the optimal parameterization. In supervised machine learning, the loss L helps us evaluate the f's, but it does not tell us how to search for it.

We have to go where the functions f live: This would be the hypothesis space. It's a big space. I mean, the space has to hold infinitely many functions, even if you have only one feature from which to predict the target y. In order to search within this space, we have to at least put some constraints on what f might look like. And that's where all the different model classes come into play: **decision trees, support vector machines, linear regression models, random forests, boosting, neural networks**, …

For simplicity, let's say we have only one feature x_1 and want to predict y from it. The prediction function would then be $f(x_1)$. If we restrict f to be a linear model, we only have to search all f's of the form $f(x_1) = \beta_0 + \beta_1 x_1$. We have just simplified the search in the vast hypothesis space to the search for the optimal parameters β_0 and β_1. A much simpler

task. Similarly, all other machine learning algorithms make the hypothesis space manageable so that it can be searched. Think of the hypothesis space as a dark forest. Machine learning algorithms illuminate paths through the forest so that we can search for the best f within the these paths. The globally best f might not be within this illuminated path, so we will usually only find a locally optimal f. Machine learning algorithms differ in the form and complexity they allow for f. Decision tree algorithms produce f's that look like step functions, since most trees algorithms only allow discrete jumps in prediction. Neural network are universal function approximators that can, in theory, approximate any function f.[18]

Each machine learning algorithm has its own procedure to search the hypothesis space. Most of the time, this search is about finding the right parameters for a model: Neural networks use gradient descent with backpropagation to adjust the weights, regression models use maximum likelihood to find the ideal values for the coefficients, and so on.

9.5 Overfitting

Supervised machine learning has one major nemesis: overfitting. Remember, the goal is to achieve a low generalization error. But as long as we only use training data, we don't know how well the model will perform with new data. Worse, machine learning models can easily overfit the training data. Think of overfitting as memorization of the training data. When the model perfectly memorizes the training data, it will have have zero loss on the training data, but will likely perform badly with new data.

The opposite of overfitting is underfitting. If the hypothesis space is too constrained, then model may not be flexible enough to represent the true relationship between the input features and the target.

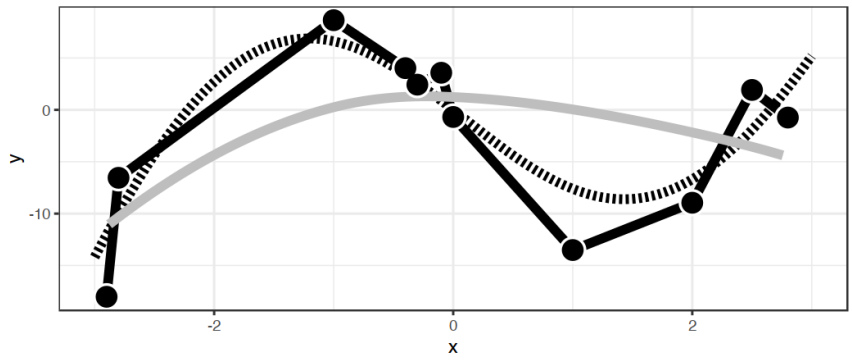

Figure 9.1: The target y is dependent on x through a function f(x) (dotted line). The observed data (points) have an additional random error. One model overfits the randomness in the training data (black line), and the other underfits (grey line).

Fitting a supervised model means walking a fine line between underfitting and overfitting. Model evaluation is central to finding this delicate balance and not ending up on one side of the cliff or the other.

9.6 Evaluation

Let's say you want to enter a cooking competition. A contest with a panel of judges who will evaluate your food and insult you on live TV if it tastes like crap. You've been practicing your cooking skills for a while. Fortunately, you have some "ground truth data" about how well your food is received. You cook for family and friends often, and they've given you feedback on how good your dishes tasted. Over time, your dishes got better and better, and today you consistently get excellent reviews from family and friends.

The jury is the ultimate test of your cooking skills. You have never cooked for these judges before. So this test is about how well your cooking skills generalize to new data points. But are your confident enough about your skills? What if your supposed kitchen prowess is attuned to strange tastes? Your your family might be addicted to salt, for example. And the jurors would be like: "Did you cook this with seawater?", "What is this? Bread? Or a salt lick for goats?". In order not to bring shame to your family and name, you decide to validate your skills before this ultimate test. So you cook for some new people who have never tried your dishes before. This way you can evaluate your skills without having to waste your shot in the contest.

Rigorous evaluation is close to the heart of supervised machine learners. A model generalizes well to the real world if the generalization error is low. A typical recommendation of supervised machine learners is to set up the evaluation pipeline even before training the first model. In supervised machine learning, evaluation means measuring a loss L for unseen data, usually called "test data". The test data is like the judges in a cooking competition. The machine learner may not to use the test data to train the model or test it prematurely. The test data may only be used for the final evaluation. If the test data influences the model training or choice in any way, it's "burned" and does not show the true performance of the model. Rather the evaluation will be too optimistic.

Because of this "burning" of the test data, machine learners need different strategy to guide their model building. The test data are set aside. Whether to compare models or to try different configurations of a model, the machine learner needs unseen data. The trick is to repeat this train/test split within the training data. So we cut off a portion of the training data that can be used to evaluate modeling decisions. This data set is usually referred to as validation data.

In the simplest version, the data is split once before model training into training, validation and test data. In reality, techniques such as cross-validation are used to split the data multiple times and reuse the data intelligently.

Figure 9.2: For evaluation, the data is usually split into training, validation and test data. There are more complex splitting schemes where the data is split multiple times.

9.7 An Automatable Mindset

Supervised machine learning is automatable to a degree that surpasses all other mindsets. Using a well-defined evaluation procedure, the generalization error, the entire process of model building can be automated. Supervised machine learning is essentially an optimization algorithm. In statistical modeling, such as Bayesian and frequentist inference, we have to make all the assumptions, choose the right distributions, decide on the variables to use in the model, look at diagnostic plots, …

There is an entire subfield of machine learning, AutoML, that deals with automating the entire training pipeline. This can include feature engineering, model training, hyperparameter optimization, evaluation, etc. Automating the supervised machine learning pipeline is computationally intensive, so there is a lot of research on how to automate everything in a smart way. As a result of this automation capability, there is an entire industry with hundreds of web services and products that automate the training process for you.

But automation is also problematic. It creates distance between the modelers and the underlying modeling task. Automation makes modelers less aware of the shortcomings of the data. On paper, the model may look very good, because the generalization error is small. But under the surface, the model may be a garbage because it uses features that are not available at the time of the prediction, or the data are terribly biased, or missing data were not handled correctly to, name just a few possible errors.

9.8 A Competitive Mindset

Another consequence of the one-dimensional evaluation is that supervised learning is a competitive mindset. Modeling becomes a sport: which is best model for a task? It also invites competition between people. Entire websites are dedicated to hosting machine learning competitions where the best modelers can win money. Sometimes a lot of money. Your skills as a modeler are reduced to your ability to optimize a single metric. That metric puts you on the leaderboard, which ranks modelers. A ranking that ignores many things, such as domain expertise, model interpretability, coding skills, runtime, … The idea of competition has also taken hold of machine learning research itself. Scientific progress, in large parts, has become

a sport. Progress in machine learning research is when a new machine learning algorithm beats other algorithms in benchmarks.

9.9 Nature, Statistics and Supervised Learning

As we have seen, the mindsets of statistical modeling and supervised machine learning can be quite different. At their core, the two mindsets involve different ideas of how to model aspects of the world. The following comparison is more or less a summary of Leo Breiman's famous article "Statistical Modeling: The Two Cultures".[19]

In the context of prediction, we can think of nature as a mechanism that takes features X and produces output Y. This mechanism is unknown and we want to learn about it using models.

Figure 9.3: Nature

Statistical modelers fill this box with a statistical model. The statistical model is supposed to represent nature. It is supposed to reproduce the inner workings of nature. If we are somewhat convinced that we have found the mechanism, we can then take the model parameters and interpret them as if it was the true mechanism in nature. Nature's true mechanism is unknown and not fully specified by the data, we have to make some assumptions about the forms of this mechanism, which we represent with the function f.

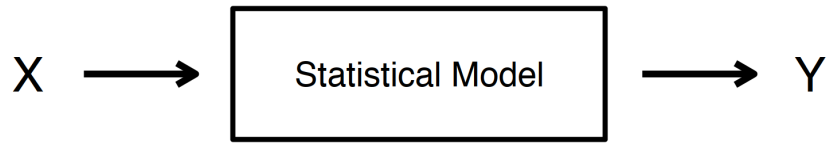

Figure 9.4: Statistical Model

In supervised machine learning, nature is seen as unknowable, or at least no attempt is even made to reverse-engineer the inner mechanisms of nature. Instead of the intrinsic approach, supervised machine learning takes an extrinsic approach. The supervised model is supposed

to mimic nature. It should show the same behaviour as nature, but it doesn't matter if it achieves this behaviour in the same way.

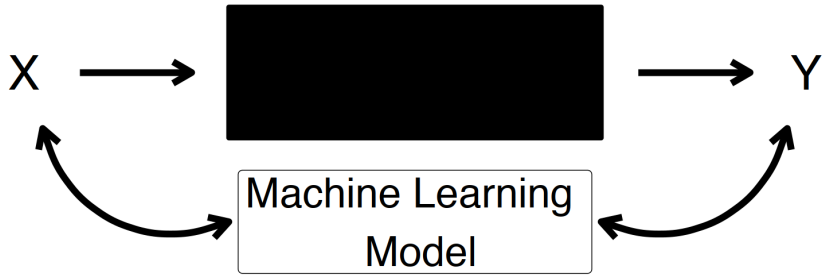

Figure 9.5: Supervised Machine Learning Model

Again, a cooking analogy: Suppose you want to recreate a dish that you ate in a restaurant. A statistician cook would try to find a plausible recipe, even if the end result is not perfect. The supervised machine learner cook would only be interested in the end result; it doesn't matter whether it's was exactly the same recipe.

No one mindset is inherently better or more useful than another. They are different mindsets with different strengths and limitations. If a task involves evaluating unseen data against a well-defined performance metric, the best approach to that task is probably supervised machine learning. If your task requires a model with a strong theory that can explain the relationships in the data, statistical modeling is the way to go.

9.10 Strengths

- The most straightforward mindset when it comes to making predictions.
- Loss function L allows the model to be adapted quite well to the task at hand.
- Supervised machine learning is highly automatable.
- Supervised learning has a very coherent evaluation approach that I personally find very convincing, though quite one-sided. Measuring how well the model predicts new data is a very compelling way to define a good model.

9.11 Limitations

- Supervised learning without constraints does not lead to interpretable models and is therefore not as well suited to for gaining insights.
- Supervised learning is not as theoretically sound as statistical modeling.
- Making decisions based on only the most likely outcome ignores tail risks from less likely, but possible extreme outcomes.
- Uncertainty quantification is not a first class citizen as it is in, for example, Bayesian inference. The modeler has to rely on a subset of machine learning algorithms that quantify uncertainty (for example Gaussian processes) or they have to use additional tools such as conformal prediction.
- Automation can lead to overlooking issues with the data and the task formulation.
- Generalization error is a good way to quantify generalization, relying solely on this metric will fail in the dumbest ways. There are many examples, such as using asthma as a predictor of lower risk of pneumonia[20], classifying based on watermarks[21], and misclassifying dogs as wolfs based on snow in the background[22].
- Feedback loops can break the models. Deployed into the wild, supervised learning models influences and even creates data that might end up training a future version of the same model. But this feedback loop is not well understood and difficult to respect in the modeling process.

9.12 References

- Statistical Modeling: The Two Cultures by Leo Breiman.[19] Highly recommended to understand differences between statistical modeling and supervised machine learning.
- I can recommend the book "Elements of Statistical Learning"[23] which covers not only supervised learning but also other machine learning topics. The book has a strong influence from the statistical modeling mindset.

10 Unsupervised Learning

- A more open and diverse mindset focused on uncovering hidden patterns in the data.
- Typical tasks: Clustering, anomaly detection, dimensionality reduction, and association rule learning.
- One of the three machine learning mindsets along with supervised learning and reinforcement learning.

A group of supervised learners and one unsupervised learner decide to climb a mountain. The trip quickly turns into a race: Who will be the first to reach the top of the mountain, and who will be the first back to the hut? The supervised learners try to outrun each other, one faster than the other. The unsupervised learner quickly falls behind. After an exhausting day, one by one, they return to their hut. To their surprise, the unsupervised learner is already joyfully waiting for them in the hut. Everyone was eager to know how the unsupervised learner managed to climb faster than them. "When you all sprinted off, I took a detour," the unsupervised learner reported,"You won't believe this, but I found a rare mushroom that is not supposed to grow in this area. I have also divided the area around the hut according to the vegetation you find there. But the best part is that …" "Wait!" interrupts one of the supervised learners, "You were not only the first one back, but you also did all these other things?" "I guess so", the unsupervised learner admits, a little puzzled. "How long did it take you to climb the mountain? Did you find a shortcut? We haven't seen you all day.", asks another supervised learner. "Mountain? I didn't see any mountains."

10.1 What Type of Traveler Are You?

Tip Top Travel, a travel agency I just made up, offers a wide range of trips, from all-inclusive holidays in Spain to hiking trips in Norway and weekend city trips to Paris. They have a huge database on the booking history of their customers: The date of the trip, the destination, group size, cost of the trip, and so on. And yet, they know surprisingly little about the general patterns in their data: Are there certain prototypes of customers? Do customers who book trips to Norway, for example, also book trips to Sweden? Our imaginary travel company's big data is a dream for unsupervised learners. They might start with a cluster analysis to get an overview Tip Top Travel's different customer types. A cluster is a group of customers with similar travel patterns. Similar means that the customers in a cluster are "close" in terms of their feature values, such as booking frequency, travel locations, and average cost of the trip. If Tom books 4 trips per year, he is more similar to Tina, who books 5 trips per year than Philipp, who books 1 per year. It gets more tricky because the measure of distance between travelers has to combine features that were measured at very different scales, such as amount

of money, geographical location, counts, and so on. But I'll rant about that problem later. With such a distance measure, there are many clustering algorithms to choose from that can detect clusters. For example, the k-means algorithm.

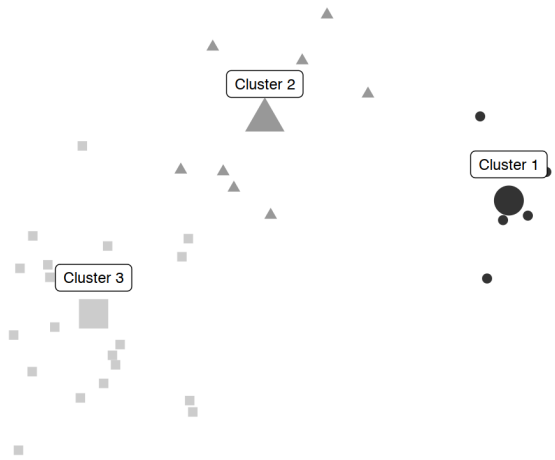

Figure 10.1: Three clusters for two features (x-axis and y-axis) based on k-means clustering.

The k-means algorithm divides our customers into k clusters. The number of clusters k must be chosen by the modeler. The algorithm start with random cluster centers. These cluster centers are iteratively optimized so that all customers in the same cluster are close to each other. Each customer is assigned to the cluster with the closest center. The objective of the k-means algorithm is to find cluster centers that minimize the distances between centers and assigned data points, averaged over all customers.

Great! We have clusters. Now what? We wanted to know what kind of customers Tip Top Travel has. To do this, we can interpret the cluster centers as prototypes of customers. One cluster could be "families looking for relaxation". The cluster center is a customer that books about twice a year for 1.9 adults and 2.1 children. The most common destination is all-inclusive hotels with pool and beach access, where parents can drop off their kids for an all-day entertainment program and then the adults can have their first beer at the pool bar 11 AM. Cluster numero dos is "luxury explorer". Customers in this cluster book an average of 0.7 trips per year for 1.7 adults, with varying destinations in nature with some adventure aspect: hiking, canoeing and camping. Depending on the k chosen by the modelers, there might be more clusters, and they might not always map to some easily interpretable group of travelers. The more interpretable the clusters are, the more the marketing department will love the results. Cluster analysis can provide data-driven insights about customers and offers a narrative angle on which to build marketing campaigns.

10.2 The Unsupervised Learning Mindset

Unsupervised learning is like a journey of discovery. A dataset suddenly becomes a treasure chest potentially filled with valuable insights. The supervised learner can only watch from the sidelines: sipping an energy drink, preparing for the next race; the only excitement is what the stopwatch will show this time.

Unsupervised learning is a machine learning mindset: task-driven, computer-oriented, and externally motivated. Task-driven: We use unsupervised learning to solve specific tasks, such as clustering, anomaly detection, or finding a better representation of the data. Computer-oriented: Like supervised learning, its unsupervised counterpart is motivated by the premise of having a computer, rather than by the premise of a theory where it's simply convenient to have a computer. Unsupervised learning is externally motivated: While measuring performance is more difficult than in other machine learning mindsets, successfully completing the task is more important than following a particular "recipe" (such as using probability theory).

Unsupervised learning is a less coherent mindset than supervised learning with its very rigorous evaluation and Zen-like mindset of optimization. Unsupervised learning is about discovering patterns in the data, which sounds a bit fuzzy. Fortunately, we can use the language of probability theory to make it more understandable. Unsupervised learning is about finding a more compact representation of the joint distribution $P(X)$ or revealing some aspects of $P(X)$. Unsupervised learning includes a broad range of tasks:

- Clustering finds modes of the distribution.
- Anomaly detection finds extreme data points.
- Association rule learning finds modes in binary feature distributions.
- Dimensionality reduction finds lower-dimensional descriptions of the data.
- …

Why do we need unsupervised learning anyway? Can't we just hire a statistician to estimate $P(X)$ and derive all these interesting aspects from that estimate? Well, estimating the joint distribution is extremely difficult with high-dimensional data. The difficulties in estimating $P(X)$ become difficult even with more than a handful of features when the distribution is complex, let alone image or text data.

We can also express supervised learning as learning a distribution. But it's "only" the conditional distribution $P(Y|X)$, which is much easier to learn than the full joint distribution $P(X)$. Supervised learning is about selecting one feature and making it "special", which we also express by giving it a different letter (Y). In unsupervised learning, on the other hand, all features are treated the same. Of course, each algorithm can give different weights to the features depending on the task.

The lack of a target Y to predict also means we have no ground truth to compare our results to. It's more like, "Here are n data points, please find something interesting." Then you say "Here's something interesting: I found these 10 clusters." But you'll never get any feedback on whether these were the "right" clusters. There is no one to pat you on the back and say "You did a great job." Your strength as an unsupervised learner must come from within! That's why unsupervised learning is sometimes called learning without a teacher: There is not

teacher to correct the model. This is also why we can clearly distinguish supervised learning as its own mindset and why it's not just a special case of unsupervised learning.

To be more cheerful about unsupervised learning: It's, in many ways, an open mindset. Unsupervised learning means being open to surprises and discovering hidden patterns. The word "pattern" hides a potpourri of meanings: clusters, outliers, feature representations, association rules, … The mindset is also open in the sense that the range of methods is huge, even for a machine learning mindset. For clustering alone, there are so many different approaches. If I had to pick one modeling mindset that is the most inclusive, I would choose unsupervised learning (in terms of methods, not necessarily people). Next to this hippie community, supervised learners look like dull optimizers who sit in their offices with fine suits trying to increase sales for the second quarter.

Full disclosure: unsupervised learning also involves optimization. But there is much more freedom in the optimization objective because there is no ground truth. The same is true for performance metrics and benchmarks: It's part of the mindset to evaluate models, but there's a lot of ambiguity about how to evaluate unsupervised solutions. For example, in cluster analysis, we could measure cluster purity, the silhouette score, various indexes, look at elbow plots, and so on. One can also create a long list of metrics for supervised learning, but at least they agree on when they become zero (or reach their minimum): When the target is accurately predicted. A luxury that doesn't exist in unsupervised learning.

10.3 Many Tasks

To get a better understanding of unsupervised learning, let's take a look at some of the tasks that are typical of the mindset.

10.3.1 Clustering and Outlier Detection

Both clustering and outlier or anomaly detection are two opposites sides of the same coin. In both cases, the question is where the mass of the data lies.

Clusters are regions in the feature space with a high concentration of data points. In terms of $P(X)$, these regions are modes of the distribution. Outliers or anomalies are in regions where $P(X)$ is small, which are regions in the feature space with almost no data points. Clusters are usually defined such that all the data points within a cluster a similar in their feature values. There are many different approaches for finding clusters: hierarchical clustering, k-means, k-medoids, DB-SCAN, PRIM, Gaussian mixture models, self-organizing maps, … These clustering methods have various motivations, ranging from statistical to more algorithmic, again showing that unsupervised learning is externally motivated: It isn't particularly important *how* the clusters are detected. Different clustering methods can find very different clusters. Let's take a look at a another solution for cluster analysis from Figure 10.1.

The clusters in Figure 10.3 are clearly a poor solution based on Euclidean distance. But this assumes that both features matter are equally important for computing similarity. Perhaps

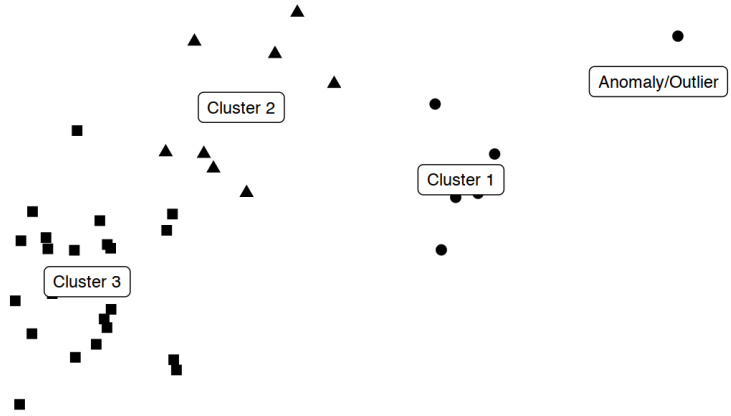

Figure 10.2: Example of three clusters in a two-dimensional feature space. The data point on the top right could be called an outlier.

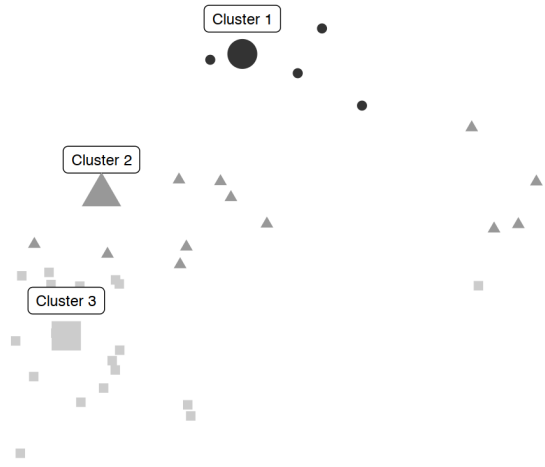

Figure 10.3: Bad clustering solution, measured by Euclidean distance. But good solution if we are only interested in the horizontal direction, giving a weight of zero to the feature that distributes the data in the vertical direction.

the feature that spreads the data in the horizontal direction is rather irrelevant, and we want to give it less weight in the distance computation? Domain experts can tell you which features are important and which aren't. Usually, all features are weighted equally, which seems to be a sensible decision at first glance. But in truth, any other weighting can an equally good choice. For example, when clustering fruits, two of the features might be highly correlated: We could use the features "volume" and "volume after peeling". In practice, we now have two features with almost the same information. If we use both of the feature for clustering, it's as if we give twice as much weight to the volume of the fruit as to any other feature.

And it gets worse. What do we do when features are measured at different scales, such as a weight, a length, and a sum of money? How can we even combine these features into a single distance measure, especially when the features even have different cardinalities, like numerical and categorical features?

"Excuse me, how do I get from point A to point B?" "You go down this street and turn right after 10 meters. Then you go another $7 and turn left until red becomes blue!". And also, what's closer to a banana? An apple or a lemon? Sounds almost like an obscure interview question for a data scientist position at large web search company. But it's a question you better have some answers to if you do cluster analysis.

So no one can tell you what's right. As an unsupervised machine learner, you have to live with ambiguity. Going to bed every night, questioning the fabric of space and time. Supervised machine learners don't have this problem, at least not to this extent. The weighting of features is "supervised" by the relationship between the features and the target for which we know the ground truth.

10.3.2 Anomaly Detection

Cybersecurity specialists monitor events in the intranet. Their job is to protect the company from cyber threats: Trade secret theft, malware, digital blackmail, you name it. But there are thousands of employees who leave their digital footprint every day. A mass, or rather, mess of data. What does an attack look like? The specialist has rules to detect some forms of attacks: If someone tries to brute force a password to log into a service, that's a red flag. But what about all the behaviors that don't follow such simple patterns, perhaps even unknown type of attacks? Fortunately for the cybersecurity specialist, there is unsupervised learning. Anomaly detection, an unsupervised learning task, is concerned with finding extreme data points. Typical applications include financial fraud detection and cybersecurity. Isolation forests, a popular anomaly detection tool, instead work by isolating data points that are extreme. Other algorithms are directly motivated by probability distributions and flag data points as anomalies if they have a low probability. Isolation forests, statistical tests, but also one-class support vector machines and hidden Markov models – the variety of methods shows yet again that machine learning, in this case unsupervised, is a very pragmatic modeling mindset.

10.3.3 Association Rule Learning

I love grocery shopping. Many people hate it, but only because they are unaware of it's magnificence: Supermarkets are incredible places that deserve awe and wonder and embody the progress and ingenuity of humanity. Supermarkets are like the land of milk and honey. It's incredible what you can get in the supermarket: exotic fruits, spices from all over the world, products that take months or even years to make, like soy sauce, wine and cheese. But I digress. Let's talk about association rule learning, which is usually introduced with shopping baskets as example. When you go shopping, you can think of your shopping basket as a binary dataset. Either you buy a certain item (1) or you don't buy it (0). Other people also go shopping and generate their own data of 0's and 1's.

The baskets might look like this: $\{yeast, flour, salt\}$, $\{beer, chips\}$, $\{sandwich, lemonade, chips\}$, $\{cheese, onions, tomatoes, potatoes, flour, oliveoil, chocolate, beer, chips\}$. The goal of association rule learning is to detect the patterns of items. Do people who buy flour often buy yeast? Association rule mining is again a case of describing $P(X)$. An association rule might be $\{beer\} \Rightarrow \{chips\}$ and would mean that people who buy beer frequently buy chips. In more formal terms, association rules are short descriptions that use conjunctive rules to describe high density regions in a binary feature space. A well-known algorithm is the Apriori algorithm, but again, there are many option to choose from. The next time you go to the supermarket, please take a moment. Take it all in. The fact that you have so many choices. Many bad things are happening in the world, but when you stand in a supermarket (with enough money), you are living what humanity must have dreamed of for thousands of years and what we take for granted. I hope you appreciate this as much as I do.

10.3.4 Dimensionality Reduction

Unfortunately, there is the "curse of dimensionality". The curse is that data density decreases exponentially with each additional features. If the number of data points remains constant, adding more features makes any modeling task more difficult, regardless of the mindset. Dimensionality reduction can be used to break this curse, or at least reduce its burden. If unsupervised learning regards all features, on what basis can we reduce the dimensionality? Not all features contribute towards $P(X)$. Some features may almost have no variance. Other features may be highly correlated with other features. In either case, we can select a subset of the features, and $P(X)$ will mostly look the same. There are several methods for feature selection based on information-theoretic measures such as statistical correlation.

Or we can take our data and map it into a lower-dimensional space. Those dimensionality reduction techniques usually make you wish you had paid better attention in linear algebra. They can usually be represented as matrix multiplication of your original feature matrix: principal component analysis (PCA), ICA, non-negative matrix factorization, multidimensional scaling, t-SNE, and so on. If each of your data point represents a fruit, features like height, width and weight could be mapped to a new feature / dimension that represents the volume of the fruit.

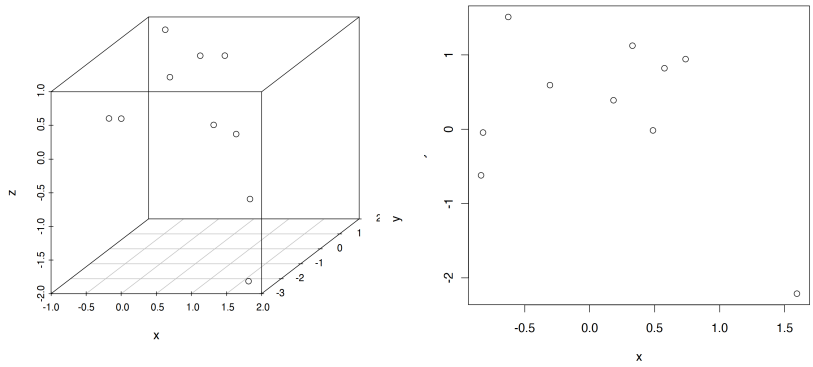

Figure 10.4: Dimensionalty reduction

Unsupervised learning includes more tasks than just clustering, outlier detection, association rule learning and dimensionality reduction. For example, archetypal analysis, latent variables, factor analysis, and more. Unsupervised learning is truly the colorful bird among the modeling mindsets.

10.4 Strengths

- Find hidden pattern in the data that modelers using a supervised learning mindset would likely miss.
- The overall mindset is very open in terms of the range of tasks included, how success can be evaluated, and openness to results and new discoveries.
- The openness for discovery can also mean discovering new business opportunities, learning something new, or gaining scientific insights.
- The world is messy. One often has data, and a sense that the data might be potentially insightful. In this case, the unsupervised learning mindset is wonderful because it gives you the ability to just dive in.
- As a more exploratory mindset, unsupervised learning can be a good starting point for further analysis of the data. These further steps can be done with a different mindset.
- Unlike any other mindset, unsupervised machine learning is able to sift through high-dimensional and complex data, partially automatically.
- Unsupervised learning works without a ground truth. This means no effort is required to label data.

10.5 Limitations

- One of the major limitations is the lack of ground truth, which comes with the difficulty of evaluating the resulting models. As a result, there are many methods with often very different results.[23]
- Unsupervised machine learning is a good approach to the curse of dimensionality, but even so, unsupervised machine learning can suffer greatly. For example, the more features there are, the more meaningless and difficult to interpret the clusters become.
- In order for the modeling results to make sense, to patterns usually need to be interpreted. This is especially true for clustering and association rule learning. This requires expertise and human intervention.
- There is no guarantee that meaningful patterns will be uncovered. And if no patterns are uncovered, there is no guarantee that another unsupervised would have uncovered something interesting.

10.6 Resources

- The book Machine learning: a probabilistic perspective by Keven Murphy.[24]

11 Reinforcement Learning

- The world is dynamic: The model is about an actor acting in an environment. The actions of the agent are only rewarded at the end.
- Reinforcement learning balances exploration and exploitation,
- A sub-mindset of machine learning.

This chapter is under construction! Stay tuned.

12 Deep Learning

- Neural networks based on stacking various layers of neurons.
- World of feature embeddings, transfer learning and foundation models.
- A machine learning mindset, can also be used for supervised learning.

This chapter is under construction! Stay tuned.

13 Interpretable Machine Learning

- Models are built with a machine learning mindset: The best approximation of the task at hand.
- The model is interpreted with additional tools.
- Often takes an external view of the model: Describing how it behaves instead of analyzing the components.
- A machine learning mindset.

This chapter is under construction! Stay tuned.

14 Design-based Inference

- Focused on sampling data representatively rather than complex modeling.
- Data population is finite; uncertainty about results is only due to sampling.
- Often combined with Statistical Modeling, especially Frequentist Inference.

This chapter is under construction! Stay tuned.

References

1. Weisberg M. Simulation and similarity: Using models to understand the world. Oxford University Press; 2012.
2. Weisberg M. Who is a modeler? The British journal for the philosophy of science. 2007;58(2):207–33.
3. McElreath R. Statistical rethinking: A bayesian course with examples in r and stan. Chapman; Hall/CRC; 2020.
4. Kao WL, Puddey IB, Boland LL, Watson RL, Brancati FL. Alcohol consumption and the risk of type 2 diabetes mellitus: Atherosclerosis risk in communities study. American journal of epidemiology. 2001;154(8):748–57.
5. Ioannidis JP. Why most published research findings are false. PLoS medicine. 2005;2(8):e124.
6. Perezgonzalez JD. Fisher, neyman-pearson or NHST? A tutorial for teaching data testing. Frontiers in psychology. 2015;223.
7. Yang R, Berger JO. A catalog of noninformative priors. Institute of Statistics; Decision Sciences, Duke University Durham, NC, USA; 1996.
8. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. Journal of the American statistical Association. 2017;112(518):859–77.
9. Tiao GC, Box GE. Some comments on "bayes" estimators. The American Statistician. 1973;27(1):12–4.
10. Richard R. Statistical evidence: A likelihood paradigm. Routledge; 2017.
11. Hacking I. Logic of statistical inference. 1965;
12. Gandenberger G. Why i am not a likelihoodist. Ann Arbor, MI: Michigan Publishing, University of Michigan Library; 2016.
13. Lazer D, Kennedy R, King G, Vespignani A. The parable of google flu: Traps in big data analysis. Science. 2014;343(6176):1203–5.
14. Hernán MA, Robins JM. Causal inference. CRC Boca Raton, FL; 2010.
15. Pearl J. The do-calculus revisited. arXiv preprint arXiv:12104852. 2012;
16. Pearl J. Causal inference in statistics: An overview. Statistics surveys. 2009;3:96–146.
17. Athey S, Imbens G. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences. 2016;113(27):7353–60.

18. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural networks. 1989;2(5):359–66.

19. Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science. 2001;16(3):199–231.

20. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. 2015. p. 1721–30.

21. Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. Unmasking clever hans predictors and assessing what machines really learn. Nature communications. 2019;10(1):1–8.

22. Ribeiro MT, Singh S, Guestrin C. " why should i trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. p. 1135–44.

23. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: Data mining, inference, and prediction. Vol. 2. Springer; 2009.

24. Murphy KP. Machine learning: A probabilistic perspective. MIT press; 2012.