Mathematical Foundations of Psychology, Fall 2013

Timo von Oertzen

November 29, 2013

Contents

Chapter 1

Linear Algebra

This chapter will teach you two concepts, vectors and matrices. Vectors are a convenient way of combining multiple numbers into a single object, the *vector*. For example, assume you collected multiple psychological variables (have your pick which ones) for some participants in a study. The outcome for each participants are multiple numbers (one for each variable), so we store them together into a single vector. This is an abstraction; we do this because everything we learn to analyze data can then be applied to single variables, two variables, or any other number of variables. In other words, we invest in understanding vectors, and our payoff is that we have to learn all analysis techniques only once, and can apply them to as many variables as we want.

Matrices describe changes of vectors, which is typically called a *map*. The maps described by matrices are a special subclass, the *linear maps*. We make these restrictions because linear maps are a simple and yet powerful class of maps; so we have less to learn and still can cope with most problems we encounter in social sciences.

So from a higher level, this chapter teaches two basic principles of effective lazyness: Vectors are a 'detour' to understanding variable analyses, the additional work quickly pays off when we don't have to think about univariate and multivariate analyses separately. Matrices are a restriction (and often a simplification) of the real world, but they give us 90 % effect for 10 % effort.

1.1 Vector Spaces

1.1.1 Goals

Terms: Vector, vector space, scalar multiplication, dimension, basis, scalar product, norm
Skills: Adding vectors, multiplying numbers to vectors, translating real-world entities into vectors
Understanding: Vector spaces are all around us, especially in social science data.

1.1.2 Vector Spaces

Assume a friend is on his way to visit you and calls from his car: 'Man, I'm lost; I see the post office here, where am I?'. Assuming you know where the post office is compared to your place, you'll answer something like 'You are still 2 miles north and 1 mile west of my place' (if your friend is good with terms like 'north' and 'west'). We need two numbers to describe his position (unless (s)he is straight north, south, east or west), so 'a number' is not sufficient to describe a position on a flat surface, we need two numbers; in math, such pair of numbers are called *vectors*. The space of all possible vectors (in our example, every possible position relative to your house) is called a *vector space* [1]. If we are talking about the position of a plane instead of a car, we might even need three numbers, while if we know that the person is already on the same street as our house, a single number (the distance in miles, for example) is sufficient.

The definition of vector spaces looks awfully complex, but is really one of the simplest in this book. Every set can be a vector space (the elements are then called vectors) if we can make up a multiplication with real numbers ($\cdot$) and an addition ($+$). The addition takes two vectors and creates a new vector; in our example '1 mile north' plus '2 miles west' together add up to the actual position of the post office. Addition needs to satisfy some rules we typically expect from $+$: Order does not matter (so $x + y$ is the same as $y + x$, and $(x + y) + z) = x + (y + z)$), and there exists a 'zero', an element such that $0 + x = x$ for all vectors x. Lastly, every element x has a 'negative', an element $(-x)$ such that $(-x) + x = 0$. In our example, if our friend finally arrived at our house, we may upset him by saying 'you are 5 miles north plus 5 miles south of my place'.

The multiplication with numbers is called *scalar multiplication*. In our example, '5 miles north' is short for '5 times one mile north'. We can multiply with non-integer numbers as well; for example, if our friend starts driving in the correct direction and has covered half of the way, we can equivalently say 'you are now 0.5 miles north and 1 mile west', or 'you are at 0.5 the vector you have been previously now'.

Multiplication has less rules: If we multiply a vector with zero, the result is the zero vector. If we multiply with one, the vector does not change. Finally, there is the distributive rule that connects the two operations: For numbers λ and μ and vectors x and y, $\lambda(x + y) = \lambda x + \lambda y$, and $(\lambda + \mu)x = \lambda x + \mu x$.

As you may have observed, the literature typically uses Greek letters for the numbers (which also are called *scalars*) and Latin letters for the vectors. Rules like these are never put in stone, and you'll find different usages in different scientific fields or even within single books.

The following gives the formal definition of vector spaces, which summarizes all the above rules. It appears overwhelming at first glance, but don't allow it to do that to you; what you should keep in mind is that vectors can be added

[1] In fact, vector spaces are even more general: Instead of being 'pairs of numbers', they can be 'pairs of something', where the something comes from any algebraic field. In this book, however, we are only concerned with vector spaces over the real numbers $\mathbb{R}$, so if we say vector space, we always refer to these vector spaces

and multiplied with numbers, in the way one would expect. Everything else is details.

Definition 1.1.1. An $\mathbb{R}$-vectorspace is a set V with two operations,

$$\begin{aligned} + &: V \times V \to V \\ \cdot &: \mathbb{R} \times V \to V \end{aligned}$$

such that

- $+$ has an element 0 such that $0 + x = x$ for all $x \in V$,
- every element x has a negative $-x$ such that $x + (-x) = 0$
- $+$ is commutative $(x + y = y + x)$
- $+$ is associative $(x + (y + z) = (x + y) + z)$.
- For all elements $x \in V$, $0 \cdot x = 0$
- For all elements $x \in V$, $1 \cdot x = x$
- For all $\lambda, \mu \in \mathbb{R}$ and $x, y \in V$, $\lambda(x+y) = \lambda x + \lambda y$ and $(\lambda + \mu)x = \lambda x + \mu x$.

1.1.3 Examples for Vector Spaces

Vector spaces are everywhere you look. Prices in the grocery store are a simple example: Say an apple costs \$1.50. The price for a basket with two apples is a multiplication with 2 (or three or four or whatever). If a peach costs \$2, then a basket with an apple and a peach is the sum of the individual prices, which is, \$3.50. Observe that all rules above are satisfied: An empty basket is the zero vector, 0 times any given basket of fruits is an empty basket, it doesn't matter for the price in which order we put the fruits in the basket, and so on. You think that is easy? Here you are, it's 90 % of what you need to know about vector spaces.

There is an even more trivial example: A set with a single element (let's call it '0'). We set $0 + 0 = 0$, $\lambda 0 = 0$, and you may check that all conditions are satisfied. This is the easiest vector space possible.

Positions on a plane are vector spaces, as we have seen in our initial example. You could walk from any point we consider zero (your house in the example above) one mile north and 2 miles west. We can 'double' this vector to make it 2 miles north and 4 miles west. We can add two movements, say 2 miles north and 4 west to another 3 miles west, to reach a new vector which is '2 miles north and 7 miles west'. We can easily check that all rules above are satisfied, so again, we have a vector space. Note that we could have negative numbers just as well: Negative two times '3 miles west' would be '-3 miles west', which we may also call '3 miles east' if it pleases us.

Let's use a data analysis example. Say we have taken the IQ and the body height from a number of participants, and let's say for simplicity that we write

them as deviations from the population average. Then again, we can add up vectors (for example, we could add up all males to get the sum of IQ points and body heights for males). Equally easily, we can multiply the score with something if it makes sense for whatever reason. So again, scores on our test are a vector space. Observe that this vector space is very similar to the positions on the map in that every vector in both cases represents two numbers. Both vector spaces have the same number of *dimensions*; let's talk about that concept a little more.

1.1.4 Bases and Dimension

A basis of a vector space is a minimal subset of vectors from which you can create all other vectors by addition and multiplication, or formally,

Definition 1.1.2. A *basis* of a vector space V is a minimal subset of elements $u_1, ..., u_n \in V$ such that for all $x \in V$, x can be written as

$$x = \lambda_1 u_1 + ... + \lambda_n u_n$$

Definition 1.1.3. The minimal number of elements needed for a basis in a vector space is called the *dimension* of the vector space.

The horizontal dots in the sum require the reader to guess some missing part, which is something we don't like too much; so in math, we usually write

$$\lambda_1 u_1 + ... + \lambda_n u_n = \sum_{i=1}^{n} \lambda_i u_i$$

In the example with grocery prices, the basis is trivial: The price '\$1' is already a basis, because obviously any dollar amount can be represented as λ times one dollar. Observe that using a single basis vector is not possible for the example using directions on a map: No matter with which direction we start, we can not reach a direction that is not on the same line. If we start with 'one mile north' as the first direction, we can not multiply it with any number to get the vector 'one mile west'. But as soon as we take two directions which are not on the same line, we can reach any other point (you may want to test that these can be any two directions). So the plane is 'two-dimensional', which coincides with our intuition of the word.

By agreement, the sum over no summands is zero (so $\sum_{i=1}^{n} x_i = x_1 + ... + x_n$ for $n = 0$ is always zero). If we use this, then the extreme example of a vector space which only contains a single element needs no elements in the basis and is hence a zero-dimensional vector space.

Observe that in the direction example, the set {1m North, 1m West, 1m South, 1m East} with four elements also represents all directions. However, the set is not minimal (since 'South' and 'East' are not needed), so it is not a basis.

The zero-dimensional vector space is the only vector space that has finitely many elements. As soon as we have one dimension, the vector space looks like the real numbers, so it has infinitely many vectors.

As a child, were you sometimes considering what larger number is there than infinity? If so, two-dimensional vector spaces are the answer, because we now have two entries which both have infinitely many elements.

1.1.5 Notation of Vectors

There is a theorem that tells us that if $u_1, ..., u_n$ is a basis of a vector space V, then every element x in V has exactly one representation in this basis. That means if $x = 5u_1 + 3u_2$, then there are no other λ_1, λ_2 such that $x = \lambda_1 u_1 + \lambda_2 u_2$. Think about this a moment, it's really fairly simple; if I'm 5 feet north and 2 feet west of you, then by using the words 'north' and 'west' there is no other description of my position relative to you. We say that every vector is *uniquely described* by its combination of basis vectors. This gives rise to a shorthand notation of a vector as a column of numbers between parentheses, which we will use in this book; so if you see

$$x = \begin{pmatrix} 5 \\ 3 \\ 2 \end{pmatrix}$$

then this really means $x = 5u_1 + 3u_2 + 2u_3$ for a basis $\{u_1, u_2, u_3\}$ of the vector space.

Note that addition of vectors can be done by adding each element separately, that is,

$$\begin{pmatrix} 5 \\ 3 \\ 2 \end{pmatrix} + \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 4 \\ 4 \\ 2 \end{pmatrix}$$

and every scalar multiplication is a multiplication of the scalar with each entry, that is,

$$5 \cdot \begin{pmatrix} 5 \\ 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 25 \\ 15 \\ 10 \end{pmatrix}$$

This shorthand notation makes addition and multiplication easy; if we would write the addition above with the basis vectors, we would use much more lines:

$$\begin{aligned} (5u_1 + 3u_2 + 2u_3) + (-1u_1 + 1u_2 + 0u_3) &= 5u_1 + 3u_2 + 2u_3 - u_1 + u_2 \\ \text{(commutative law)} &= 5u_1 - u_1 + 3u_2 + u_2 + 2u_3 \\ \text{(distributive law)} &= (5-1)u_1 + (3+1)u_2 + 2u_3 \\ \text{(simple addition)} &= 4u_1 + 4u_2 + 2u_3 \end{aligned}$$

This is much bulkier. However, it should not be forgotten that the vector notation with a column of numbers is just a shorthand. For example, whenever it is used, we assume that the reader knows which basis we are talking about.

1.1.6 Scalar Products and Norm

We have discussed scalar multiplication, which is the multiplication of a real number with a vector, which gives another vector. The scalar product, on the other side, takes two vectors with same dimensions and returns a number. Corresponding numbers in the vectors are multiplied, and the results are added up:

Definition 1.1.4. Let

$$x = \begin{pmatrix} x_1 \\ \dots \\ x_n \end{pmatrix} \quad \text{and} \quad y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}$$

then, the scalar product is

$$< x, y >= \sum_{i=1}^{n} x_i y_i$$

You can think of the vector product as the 'common strength' of the two vector. If all entries of the vectors have the same sign, then the scalar product will be high. Geometrically, this corresponds to the two vectors showing in similar directions. If both vectors have an ankle of 90°, however, then the scalar product is zero. This situation is called *orthogonal*:

Definition 1.1.5. If for two vectors $< x, y >= 0$, then x and y are *orthogonal*.

Higher entries of the two vectors will typically mean a higher scalar product (in fact, if we multiply a vector by 2, the scalar product doubles). So the scalar product of a vector with itself is a measure for the 'overall absolute value' of the vector. The *norm* of a vector is the square root of its scalar product with itself,

Definition 1.1.6. The norm of a vector x is given by

$$|x| = \sqrt{< x, x >}$$

Observe that the scalar product and the norm depend on the basis we choose. This is not surprising in the real world (of course, the absolute distance is larger if we talk in feet instead of miles), in the abstraction, however, this sometimes is forgotten.

1.1.7 Translating the Real World into Vector Spaces

Many things around us can be conceived of as vector spaces. Typically, if we do this, we tweak things a little. For example, prices of groceries are not strictly a vector space since you can not multiply a dollar by 0.005 (or at least half a cent is an unusual price). Also, it becomes a philosophical debate whether negative prices exist, although most would accept this (take one of our customer cards so that we know what you are shopping, and you get money from us).

The translation process requires us to pick a basis. The basis is the equivalent to a unit in one-dimensional spaces. We have a lot of freedom which basis to pick, and the same objects may look completely different in different basis. For example, we can describe the grocery prices in cent or in dollar; the actual world does not change, even though all our numbers are different by a factor of 100. When describing directions on the plane, we have even more choice: We may use 'north' and 'west' as basis, or 'north' and 'east' (observe that in real life, we often choose a basis so that we avoid negative numbers). We may use

'north-west' and 'south', if it pleases us. Maybe the two streets our friend has to drive to reach our house go in these two direction, which makes it easier to tell him: 'go down market street south, then turn sharp right into main street that goes north-west until you see me waving'. The strictness of math only strikes once we have picked the basis: Then we are forced to live with it.

1.2 Linear Maps

1.2.1 Goals

Terms: Map, Linear Map, Image, Kernel
Skills: Translate Real-World descriptions of change processes into linear maps
Understanding: Linear maps are simplification, but often good approximations of reality, and at the same time easy to handle.

1.2.2 Linear Maps

Think of a rotating disk on a play ground, which at the moment is at rest, and a child that sits on it. We can describe the position of the child as a vector space; let's use the center of the disk as zero, and 'right' and 'up' as two arbitrary directions on the disk. For better intuition, assume 'up' points up and 'right' points right if you draw the disc on a paper in front of you, and the child sits 'up', that is, at $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$. Assume the disk starts to spin clockwise, with a quarter rotation per second. We can describe 'a quarter rotation' in terms of the vector space by saying that a child at the 'up' vector will be on the 'right' vector, that is,

$$f\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

where 'f' denotes the rotation.

Such change processes are called *maps* in math (or synonymously, functions). A map takes some arguments from a *source space* and maps them into a *target space*. On the rotating disk, these two spaces are the same (the positions on the disk). For another example, if a politician states 'grocery prices have doubled since 1980', then this is a map which takes a price and returns twice that price. In typical math notation, this reads

$$f(x) = 2 \cdot x$$

Observe that the x can be a number or a vector (of any dimensions). For example, if we have collected some cognitive variables, and a treatment doubled the score of all participants, then the above map would describe that as accurately as it describe the fiscal development. Note also that the source space and the target space don't need to be identical, the same notation could be used to map into a different vector space. If chewing gums cost 50 cent each, then f may be used to describe the statement 'One dollar buys two chewing gums'.

These examples have a special property: If we halve the input vector, the output vector is also halved. That is, if the kids sits just between the center of the disc and the 'up' edge (at $\begin{pmatrix} 0.5 \\ 0 \end{pmatrix}$), then after one second it will be just between the center and the 'right' edge (which is, $\begin{pmatrix} 0 \\ 0.5 \end{pmatrix}$). Also, if we add two vectors x and y before we map them with f, then the result is the sum of both individual results. Such maps are called *linear maps*:

Definition 1.2.1. A map f from one vector space V_1 into another vector space V_2 is called linear if for all $x, y \in V_1$ and $\lambda \in \mathbb{R}$

$$\begin{aligned} f(x+y) &= f(x) + f(y) \\ f(\lambda x) &= \lambda f(x) \end{aligned}$$

In other words, for linear maps you can exchange the order of the map and the operators: It's the same whether you apply the map to a sum, or first apply the map to both summands and then sum up the results from both. Linear maps are a strong restrictions, and not all 'changes' we want to describe in social sciences are linear maps; however, we get a lot of bang for our bugs here, since linear maps are (1) very simple and (2) are surprisingly frequent in reality [2].

1.2.3 Image and Kernel

Each Linear map comes with two spaces. The first is the *image*, which is the subset of the target space that can be reached by the map. The second is the *kernel*, which are all vectors from the source space that are mapped to zero. We denote this formally by the definition

Definition 1.2.2. Let $f : V_1 \to V_2$ be a linear map. The image of f are all vectors that can be reached by the map,

$$\mathrm{im}(f) = \{y \in V_2 \mid \exists x \in V_1 \;\; f(x) = y\}$$

The kernel are all vectors that are mapped to zero:

$$\ker(f) = \{x \in V_1 \mid f(x) = 0\}$$

Sometimes, the image is the complete target space, as in our examples with the rotating disk or the doubled prices. Here, any number (for example, \$4) is the target of another number ($f(\$2) = \$4$). However, the image can be a true subset; imagine the map that corresponds to the statement 'all drinks are free tonight'. It maps all prices to \$0, or in math notation $f(x) = 0 \cdot x$. Although the target space is still the realm of all possible prices, the image of this map is only \$0. If we think of a plane like the driving directions we gave our friend who

[2] Or at least are frequent as approximations to reality.

was driving to our house, consider a projection to the latitude (the west-east line at '0 steps north'). In math notation,

$$f\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ 0 \end{pmatrix}$$

This map is linear (something you may want to check as an exercise). The image of the map is a one-dimensional space, all vectors with a zero in the lower entry.

The kernel is a subspace of the source space. It consists of all vectors that are mapped to zero. In the price-doubling example, \$0 is the only element of the kernel, since every other element is mapped to a non-zero price. For the 'free beer' example, the kernel is the complete source space, as every price is mapped to zero. In the projection example, every element is in the kernel that has a 0 as the first entry, that is, every element on the south-north line for '0 steps west'. We can easily check that, because all these elements are mapped to zero:

$$f\begin{pmatrix} 0 \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Observe that because we are talking of linear maps, if x and y are two vectors from the kernel, then $x+y$ and λx for numbers λ are also in the kernel. That means that the kernel itself is again a vector space (a sub-vectorspace of the source space). The same holds for the image. The dimensions of these two spaces always add up to the dimension of the source space:

Theorem 1.2.3. *for a function $f : V_1 \to V_2$,*

$$\dim(V_1) = \dim(\mathrm{im}(f)) + \dim(\ker(f))$$

1.2.4 Describing Linear Maps by Their Result on Basis Vectors

If we know that a map is linear, its description becomes very easy: We only have to say what happens to the basis vectors to describe the complete map. In our example with double grocery prices, it would have been a sufficient description to say that 'what used to cost a dollar now costs two dollar'. Even though we don't say that something which used to cost \$5 in 1980 now costs \$10, this is implicitly included if we assume the map is linear, because we can change the order in which we do the multiplication and the linear map:

$$f(\$5) = f(5 \cdot \$1) = 5 \cdot f(\$1) = 5 \cdot \$2 = \$10$$

Observe that f maps infinitely many vectors (all possible prices) to a space of again infinitely many vectors, but we can describe the map by a single number (the description of the target vector for the basis vector 'one dollar'). Nice, isn't it? That is one reason why linear maps are so comfortable.

The same works with multiple dimensions. Assume the basis of the source vector space is $u_1, ..., u_n$ and of the target vector space is $v_1, ..., v_m$, and we know for a linear map f that

$$f(u_i) = \sum_{j=1}^{m} f_{j,i} v_j$$

with some numbers $f_{i,j} \in \mathbb{R}$. We already know that each vector u is uniquely described by $u = \sum_{i=1}^{n} \lambda_i u_i$. In fact, by only knowing the numbers $f_{j,i}$, we already know $f(u)$ in the basis $v_1, ..., v_m$ of the target vector space. To proof this, we again only use that we can change the order of f and the operations:

$$\begin{aligned} f(u) &= f\left(\sum_{i=1}^{n} \lambda_i u_i\right) \\ &= \sum_{i=1}^{n} \lambda_i f(u_i) \\ &= \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{m} f_{j,i} v_j \\ &= \sum_{j=1}^{m} \left(\sum_{i=1}^{n} \lambda_i f_{j,i}\right) v_j \end{aligned}$$

which, also it looks bulky, is again the standard representation of a vector in the target space we are already familiar with.

In short: If you know where the basis vectors are mapped to, you know where every vector is mapped to. The map f is hence completely described by $n \cdot m$ numbers, the $f_{j,i}$ in the above equation. From realizing this, it is a short step to represent f by a 'matrix', that is, by writing the numbers $f_{i,j}$ in an array. This is what we will do in the next section, but let's briefly talk about translating the real world into linear maps first.

1.2.5 Transforming Real World Change Processes into Linear Maps

Solving science problems has three steps: Translating a real situation in math lingo (fun), doing the math computation (boring), and translating back again (even more fun). The middle step can be done by good computer programs, so the scientist is only called in on step 1 and 3 (good for us, as these are the fun steps). The mechanics to create a 'linear map' are always the same: (1) Fix the basis of the source space, (2) fix the basis in the target space, and (3) for each basis vector in the source space, describe its image under the map in terms of the target space basis.

Remember that fixing the basis is an act of power; you can choose the first basis vector completely freely, and for every other basis vector you just have to pay attention that it is not already described by the previous. However, we have to include our world knowledge, or common sense, with our formal abilities, which sometimes resembles moving two equally poled magnets onto each other.

Let's go through our examples, let's start with the statement 'All drinks are free tonight'. The source space apparently are drinks. How many dimensions? Probably that depends on the bar we are in, let's make an example of a bar that sells beer and liquor. I choose the first basis vector to be 'one ounce of beer' and 'one ounce of liquor'. The target space are prices. Observe that I

could choose cent and dollar here if I like, but that is not a good choice as in the real world, the two are not independent (it is, however, not 'wrong' in any sense). So I choose 'dollar' as the basis of a one-dimensional vector space on the target side. The linear map maps 'one ounce beer' to 'zero dollar' and 'one ounce liquor' to 'zero dollar',

$$f\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 0 \end{pmatrix} \qquad f\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 0 \end{pmatrix}$$

Observe that the last step is the only 'work' in the sense that I'm not free what to do, but if I make good choices before, it is an easy step.

Next example, the child on the rotating disk. I observe that target and source space are identical. For a basis, I choose 'up' and 'right'. Observe I could add 'elevation' as a third dimension, but I don't because my common sense tells me that nothing interesting will happen if I do that. However, it would still be a 'correct' choice. Now to step 3, describing the result of both basis vectors under a 90° clockwise rotation. Obviously, 'up' becomes 'right'. 'right', in turn, becomes 'down', which is negative 'up'. So in numbers, the map looks

$$f\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \qquad f\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

What about this one: 'Without wind, ships don't move'. Okay, really? We want to go through all the math to describe something that trivial? The answer is yes, sometimes we want to, because we don't have to treat this situation separately later on (it is just another, albeit trivial, case of linear maps). Source and target space are the same and apparently describe the ocean plane, so we choose 'north' and 'east' as basis vectors, and any point as zero. A ship that sits one mile north of the origin and zero miles east (the first basis vector) after the map still is there (what surprise), and the same with 'east', so

$$f\left(\begin{pmatrix} 1 \\ 0 \end{pmatrix}\right) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad f\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Again, we observe that the lines describing where the basis vectors are mapped to look identical in all three examples. The simple idea to write the results vectors in a single array is what creates 'matrices', our next section.

1.3 Matrices

1.3.1 Goals

Terms: Matrix, identity matrix, matrix inverse, determinant, trace, eigenvector, eigenvalue
Skills: Matrix-Vector multiplication, matrix addition, matrix multiplication, Translation of real world problems to matrices
Understanding: Simplicity of matrices to describe linear maps, the additional simplification of using eigenvectors

1.3.2 Notation of Linear Maps as Matrices

We have seen that every linear map f can be written by giving the result of the basis vectors, for example,

$$f(u_1) = \begin{pmatrix} 5 \\ 3 \\ 1 \end{pmatrix} \qquad f(u_2) = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}$$

A *matrix* is a notation of a linear map in this form, but combining all target vectors in a single array:

$$\begin{pmatrix} 5 & 1 \\ 3 & -1 \\ 1 & 0 \end{pmatrix}$$

You see that matrices are just a shorthand notation. Because matrices are used so excessively, it is sometimes forgotten that the matrix is nothing else but the results of the basis vectors written next to each other.

There are some standard notations with matrices which may or may not be respected in the literature. Usually, upper case letters (A, B, Θ, ...) are used to denote matrices. The elements of the matrix are often the same letter in lowercase with the row i and column j as index, that is, $a_{2,1}$ indicates the element in the 2nd row in column 1 ('3' in the example above). The columns of a matrix are called *column vectors* (remember the column vectors are the result of the basis vectors in the mapping described by the matrix), the rows are called *row vectors*. The number of columns is called the *column dimension* or *source space dimension* (one column for every basis vector in the source space), while the number of rows is the *row dimension* or *target space dimension*. The set of all matrices with row dimensions n and column dimensions m is typically written as $\mathbb{R}^{n \times m}$. Matrices with $n = m$ are called *square* (mathematician are not very inventive if it comes to naming). The *diagonal* of a square matrix usually refers to the diagonal from top left to the bottom right, that is, the elements $a_{i,i}$.

Sometimes it is convenient to have a notation to turn a matrix around, that is, to write the row vectors in the columns (which automatically writes the column vectors in the rows). This is called a *Transposition*, and typically denoted by a superscript 'T' (that is, A^T is the matrix A with its rows written in the columns). In the social science literature, you will sometimes find A' instead.

1.3.3 Matrix-Vector Multiplication

A matrix is a shorthand for a linear map. With the short notation comes a technical trick how the result of a vector in this map can be computed easily. This computation is called a *matrix-vector multiplication*, and is typically denoted by Ax for a matrix A and a vector x just like standard multiplication.

Assume we apply the map (let's call it A) represented by the matrix above (let's call that also A, since it's really the same thing) with a vector $x = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$. Remember this notation for x means $x = 2u_1 + 1u_2$. The result of the map on

x is 2 times the result of u_1 (the first column of A) plus 1 times the result of u_2 (the second column of A). This is

$$A(x) = 2A(u_1) + 1A(u_2) = \begin{pmatrix} 5 \cdot 2 + 1 \cdot 1 \\ 3 \cdot 2 + (-1) \cdot 1 \\ 1 \cdot 2 + 0 \cdot 1 \end{pmatrix}$$

Observe that the result looks like A, with the two elements of v interwoven. The first entry of $A(x)$ is the first row of A multiplied element-wise with v, and adding up all the results. The second entry of $A(x)$ is the second row times v, and so on.

So the mechanics to multiply a matrix to a vector is to move through one row of the matrix and the column vector, multiply matching entries, and add up the results, which then goes in the corresponding row of the result vector. If we write A left to the vector v, the above computation looks

$$\begin{pmatrix} 5 & 1 \\ 3 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \cdot 2 + 1 \cdot 1 \\ 3 \cdot 2 + (-1) \cdot 1 \\ 1 \cdot 2 + 0 \cdot 1 \end{pmatrix} = \begin{pmatrix} 11 \\ 5 \\ 2 \end{pmatrix}$$

To remember the mechanic to move through a row on the left and a column on the right, Steven Boker suggests an easy muscle memory trick, the 'Kung Fu of Matrix Algebra'. To learn this, stand up. Cry 'Ah' (you may skip this if you are in a public place) and move your hands in a Kung Fu fighting position, the left one in a defensive horizontal and the right one ready to strike vertically. If you now look at your forearms, you see how they form the horizontal row of the matrix and, right of it, the vertical column of the vector. This is the order in which you have to go through the matrix.

1.3.4 Matrix Addition and Multiplication

If we already have two maps A and B with equal dimensions, we might define a new map C by saying

$$C(x) = A(x) + B(x)$$

so the result of a vector x is the sum of the two results of x under A and B. You can quickly check that C is also a linear map, so it must have a representation by a matrix. In fact, the matrix C is the element-wise sum of the matrices A and B, so $c_{i,j} = a_{i,j} + b_{i,j}$ for all positions (i, j) of the matrix. In particular, C has the same dimension as A and B. Again, you may want to put the book aside for three minutes to check this.

This operation is called *matrix addition*. It is a concept worth stopping for a moment that we can *add* two *maps*. We treat maps, represented by matrices, just in the same way as we treat numbers. Again, we benefit from the fact that we don't have to learn everything new: As we are used to, $A + B = B + A$ (so matrix addition is commutative), and $A + (B + C) = (A + B) + C$ (associative law). Both are easy to see considering that matrix addition is just element-wise addition of real numbers. There is also a 'distributive law' between matrix addition and vector addition in the result, although that is really the formal definition of matrix addition:

Definition 1.3.1. Let $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{n\times m}$. Then, the map $A + B$ is defined as

$$(A + B)x = Ax + Bx$$

Matrices also have a multiplication. Here the definition is not that we multiply the two results (which wouldn't make sense since multiplication in vector spaces is with numbers, not with other vectors), but we chain up both mappings. So assume we have a map $A : V_1 \to V_2$ (which reads 'a map A from V_1 to V_2) and a map $B : V_2 \to V_3$. A vector $x \in V_1$ can be mapped to V_2 by A, and from there to V_3 by B. The combined map from V_1 to V_3 is called the *matrix product* of A and B, written $B \cdot A$ (where the '$\cdot$', as usual, can be omitted). Observe that the first map, the A, is on the right, as we always write a matrix-vector multiplication with the matrix on the left and the vector on the right; therefore, the vector first meets the A and then the B.

In real world, a matrix product describe two change process that occur one after the other. Consider for example the following two statements:
A: A dollar in 1980 represents the same value that two dollar have today.
B: For a dollar today, you can buy 3 lollipops.
Both A and B are 1×1 matrices; A maps from 1980-dollars to today-dollar, B maps from today-dollar to lollipops. The matrix product of both, BA, maps directly from 1980-dollars to lollipops today, that is, it tells us how much lollipops you can buy today for the value that a dollar had in 1980. As $A = (2)$ and $B = (3)$, we have obviously $C = (6)$. So matrix multiplication looks like scalar multiplication for 1×1 matrices.

However, it looks different for larger matrices. To find the product BA, we have to map each column vector of A by the map B to find the new column of BA. Assume the column vectors of A are $a_{.,1}, ..., a_{.,n}$, that is, A is

$$A = (a_{.,1} \;\; a_{.,2} \;\; ...a_{.,n})$$

Then,

$$BA = (Ba_{.,1} \;\; Ba_{.,2} \;\; ...Ba_{.,n})$$

where each $Ba_{.,i}$ is the matrix-vector multiplication of B with the ith column of A.

If we speed up this mechanic, we again have the Kung Fu of matrix algebra: We go in parallel through a row of B (left arm in horizontal defensive position) and a column of A (right arm in vertical attack position), multiply the matching entries and add up all these. For the ith row of B and the jth column of A, the sum goes into position (i, j) of the product matrix.

Observe that this process only works if the row vectors of B have the same length as the column vectors of A (in other words, the column dimension of B must match the row dimension of A). Going back to the definition, this is expected since the target space of A is the source space of B. The dimensions of the product are the row dimension of B and the column dimension of A. So if $B \in \mathbb{R}^{n\times m}$ and $A \in \mathbb{R}^{k\times l}$, then BA only exists if $m = k$, and its dimensions will be $BA \in \mathbb{R}^{n\times l}$.

In one aspect, our intuition about usual multiplication breaks with matrix multiplication: $A \cdot B$ is in general not equal to $B \cdot A$, so matrix multiplication is

not commutative. Unless A and B are square matrices of the same dimensions (that is, both are from $\mathbb{R}^{n\times n}$), one of the orders will even not exist. But even if both are square matrices with the same dimension, $AB \neq BA$. Everyone who seriously works with matrices will mess this up at least once, so probably it's good to make a small reading break again now and create a counter example, that is, create a matrix A and a matrix B and compute both AB and BA to check that they don't match.

The other laws, the associative law $A(BC) = (AB)C$ and the distributive law with matrix addition $A(B + C) = AB + AC$ both are true as we would expect intuitively, so the commutative law is the only real trap. However, it entails some others, as we will see later.

1.3.5 Identity Matrix and Inverse Matrix

The matrix

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

describes a rotation of 90° clockwise. The matrix

$$B = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

describes a rotation of 270° clockwise (or 90° counterclockwise). When we think about A and B as linear maps, then we quickly observe that transforming a vector first by A and then by B will always result in the same vector that we started with. This obviously holds for all vectors; no matter with what we start, if we turn it 90° in one direction and then 90° in the other direction, we end up at the original vector. In this situation, B is called the *inverse* of A (and A the inverse of B).

If we multiply both matrices, we get the matrix that describes the linear map 'no change' - a rather boring map. In our example, this matrix looks

$$AB = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

instead of computing the product mechanically, we could have translated the natural language statement 'no change' into a matrix by writing in each column the result of each basis vector, which is again the same basis vector because we describe 'no change'. This matrix is called the identity matrix.

Definition 1.3.2. The square matrix

$$I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$$

is called the *identity matrix*. If for a square matrix A there exists a matrix A^{-1} such that

$$AA^{-1} = I$$

then A^{-1} is called the *inverse* of A, and A is called *invertible*.

Observe that in this case, A^{-1} is also invertible using A as the inverse; so $\left(A^{-1}\right)^{-1} = A$. As an exercise, confirm that for two invertible matrices A and B, $(AB)^{-1}$ is $(B^{-1}A^{-1})$.

Not all matrices are invertible; the following matrix, for example, is not invertible:

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

We can easily see this if we think of A as a linear map. To be invertible, there must be another linear map that inverts the action of A. The following vectors are all mapped to the same result:

$$A\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad A\begin{pmatrix} 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad A\begin{pmatrix} 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} ...$$

so the poor inverse matrix would have no idea where to map $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$, since it has multiple different source points.

All these vectors are mapped to zero, so they are the kernel of A. We hence generalize that a matrix that has a kernel with more than one element (remember that the zero vector itself is always in the kernel) is never invertible. In fact, the opposite implication is true, too:

Theorem 1.3.3. *A matrix is invertible exactly if* $\ker(A) = \{0\}$.

The reason is that if the kernel has only one element, then every vector of the source space is mapped to a different vector of the target space; otherwise, their difference would be a second element in the kernel (again, it is a good exercise to check this). So we can construct an inverse map; this map is also linear (again fairly easy to show), so it corresponds to an inverse matrix.

1.3.6 Determinant and Trace

For vectors, we have discussed norms which intuitively put all numbers of the vector into a single number, telling us how 'big' the vector is. We are interested in something similar for matrices, a single number that roughly tells us how large the entries in the matrix might be. There are multiple such short descriptions for matrices. Two of these, both for square matrices, deserve our special interest. The first one is the *trace* and is fairly easy to compute:

Definition 1.3.4. The *trace* of a matrix is the sum of its diagonal elements

$$\mathrm{tr}\left(\begin{pmatrix} a_{1,1} & \dots & a_{1,n} \\ \vdots & \ddots & \vdots \\ a_{n,1} & \dots & a_{n,n} \end{pmatrix}\right) = \sum_{i=1}^{n} a_{i,i}$$

The other one-number descriptor of matrices is the determinant, which is typically harder to compute but more useful. The determinant describes the volume[3] that is spanned by its column vectors:

[3]Where 'volume' always means the 'size' of an object with maximal dimensions; in two dimensions, the volume refers to what is typically called the area, and the volume means 'length' in one dimension.

Definition 1.3.5. The *determinant* of a matrix, $\det A$ or $|A|$, is the volume spanned by its column vectors.

In two dimensions, the determinant is the area in the parallelogram spanned by the two column vectors. You may want to check that this area can be computed by

$$\left|\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right| = ad - bc$$

There is another helpful observation: The determinant, that is, the volume spanned by the column vectors, is zero only if the space spanned by the column vectors has less dimensions than the target space. Recall that the column vectors are the images of the basis vectors, that is, the space spanned by the column vectors is the image. So the determinant is zero exactly if the image has less dimensions than the target space. Another conclusion is that in this case, the kernel has more dimensions than zero because the matrix is square, and the dimension of the target space is the sum of the dimension of the kernel and the image. So we see:

Theorem 1.3.6. *A matrix is invertible exactly if its determinant is non zero.*

With this, we can ask any computer program easily if our matrix is invertible by having the program compute its determinant.

1.3.7 Eigenvectors and Eigenvalues

Consider the linear map create by a mirror (for simplicity, let's say in 2D). 'A mirror' can be seen as a very short natural description of a map, so let's follow our procedure to create a matrix from this description: Target and Source space are the same. As a basis, let's choose the first basis vector as a vector orthogonal on the mirror, and the second basis vector parallel to the mirror, with the zero being any point on the mirror. The first basis vector is then mapped to its own negative, while the second basis vector remains unchanged:

$$A\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \qquad A\begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{i.e.,} \quad A = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$

Note that although A is not the identity, all vectors on the mirror line are not changed by A. Also, the other basis vector (and all its multiples) are mapped to its (or their) own negative, even though the map is not a multiplication with -1. So A is a map that acts as a multiplication with one in one direction and a multiplication with negative one in another direction. These numbers are called the *Eigenvalues* of A. Eigenvalues always come together with their corresponding *Eigenvectors*, in our case the two basis vectors.

Definition 1.3.7. Let A be a square matrix. A vector x with

$$Ax = \lambda x$$

for a number λ is called an *Eigenvector* of A, and λ is called the corresponding *Eigenvalue*.

The eigenvectors do not necessarily point in the direction of the basis vectors. Let's make another toy example: We have measured happiness and success on two consecutive days. We say that happiness on day 2 is twice as large as on day 1, plus the success score. Success is also twice as large as one the day before, plus one times the happiness scores. We translate that into a matrix: The target and source vector space are 'happiness' and 'success' on day 1 and 2, respectively. The first basis vector (one unit of happiness) translates to two units of happiness and one unit of success, and analogously for the second basis vector:

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

The basis vectors are no eigenvectors here, but see what happens if we map a person with equal score on both happiness and success:

$$A \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$$

So $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ is an eigenvector with corresponding eigenvalue 3. Note that every multiple of this vector is obviously also an Eigenvector with the same Eigenvalue since if $Ax = \lambda x$, then

$$A(\mu x) = \mu A x = \mu \lambda x = \lambda(\mu x)$$

A has a second space of eigenvectors (typically, people say simply 'eigenvector' when they talk about the space of all the multiples of this eigenvector) to a different eigenvalue:

$$A \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

The eigenvalue to this eigenvector is -1, which means that a person who is as happy as unsuccessful will at the next day be as unhappy as successful, and will alternate between the two for the rest of his or her life.

Note that with these two statements ('people equally happy and successful triple both properties every day', and 'people equally unhappy as successful alternate between this and the mirrored position') also describes the linear map completely. For this description, we have used the eigenvectors as a basis, which made our lives easier since describing the result of the eigenvectors is just one number (the eigenvalue) instead of two. The same trick works in higher dimensions, but only if the eigenvectors form a basis of the space. They do this under two conditions:

Theorem 1.3.8. *If A is a $n \times n$ square matrix that is symmetrical, that is $a_{i,j} = a_{j,i}$ for all i and j, and A has n different eigenvalues, then the corresponding eigenvectors form a basis.*

A always has at most n Eigenvalues, but if A is not symmetrical, then some of these eigenvalues can be complex numbers. If A is symmetrical, however, the second condition (that we have n Eigenvalues) is almost always satisfied in

practical cases. It remains the question how we can change representations in one basis system into another basis system, which we will discuss in the next section.

1.3.8 Some Further Terminology

We have concluded discussing all basics of matrices. Since matrices are so often used, many special aspects or groups of matrices got their own names over the years. These terms bear no importance, but for reading other literature and also for making our lives easier, we will introduce them in one big table (or repeat the term if we have already used it above). The first definition is that a matrix is called *square* if its row and column dimension are identical. All other terms in the following table require that the matrix is square.

Term	Meaning
A Diagonal	A is zero except for the diagonal elements
A identity (or unit)	A is a diagonal with ones on the diagonal
A symmetrical	$A = A^T$
A invertible (or regular)	A^{-1} exists s.t. $AA^{-1} = I$
A singular	A is not regular
A orthogonal	$AA^T = I$
A (semi) definite	A symmetrical, all eigenvalues have the same sign
A positive (semi) definite	A (semi) definite, all eigenvalues of A are positive

1.4 Basis Changes

1.4.1 Goals

Terms: Basis transformation
Skills: Creating basis transformations, applying basis transformations to vectors and matrices
Understanding: Changing a basis is a linear map.

1.4.2 Bases Changes

Assume a galleon of milk costs you \$3 (so 3 times the basis vector \$1). If you want to, you can express the same price using cent as basis vector; in that basis, the price for a galleon of milk is 300 cents. Observe that the price did not change, but our representation. We could also choose Mexican pesos; if one dollar is 13.28 pesos, then the same galleon of milk is 39.83 pesos. In linear algebra, these re-representations are called *basis change*.

In multiple dimensions, the bases take over the role of the units we are used to in one dimension. For example, we could describe the position of another person as '12 feet north and 2 west of my current position' just as well as saying '12 feet north and negative 2 east', or 'half the way to the TV'. The actual translation of one basis system to another in 1D is a multiplication with

a number; in general, it is done by multiplication with a matrix. This works because a basis change is also a linear transformation. Instead of changing an object, it just changes the objects description.

Changing the basis can be crucial for understanding data in social sciences. For example, if we have collected information about participants 'before' and 'after' a treatment, it may be useful to represent this as 'before' and 'change' instead, where 'change' is the difference from the 'before' data point to the 'after' data point. If in another scenario we have measured the same variable twice by different experimenters (so we have a value for 'rater 1' and another for 'rater 2'), a representation of the data as 'sum' and 'rater 1' might be more useful if we think the sum score represents the real score of the participant better. Of course, in each of these simple examples we could explain the changes we made verbally, but being aware that these are really basis changes (and thus satisfy everything we know about basis changes in general) may help us with insights that otherwise would be lost for us, and at the same time will probably make the workload to do the actual translation easier.

1.4.3 Finding a Basis Transformation

Finding a Basis Transformation is actually simple, but confusing because we have to represent basis vectors in one system by the other. Assume we have one basis (the 'old' basis) $u_1, ..., u_n$ of a vector space, and another (the 'new') $v_1, ..., v_n$ of the same vector space. Both could come as natural language descriptions (e.g., u_1 = 'north' and u_2 = 'west', and v_1 = 'north-east' and v_2 = 'south'), or one in natural language description, and the second given in the first (e.g., u_1 = 'north' and u_2 = 'west', and $v_1 = u_1 - u_2$ and $v_2 = -u_1$).

We are on the lookout for a transformation matrix Q such that a vector x in terms of $u_1, ..., u_n$ is transformed into a representation which is the same vector, but expressed by $v_1, ..., v_n$. For example, if x is 12 feet north and 3 feet west, then it should become -3 steps[4] north-east and -15 steps south. To find Q, we express the first basis vector u_1 in terms of $v_1, ..., v_n$; this will be the first column of Q, since we want Q to transform u_1 into this representation. We then repeat the same process for all us and fill the columns of Q.

In our example, u_1, which is one step north, can be represented as one negative step south, so $u_1 = -v_2$. The first column of Q will hence be $\begin{pmatrix} 0 \\ -1 \end{pmatrix}$. The second basis vector u_2 of the old basis, one step west, can be represented as negative one north-east plus negative one south, which is $u_2 = -v_1 - v_2$. Taking both together, Q becomes

$$Q = \begin{pmatrix} 0 & -1 \\ -1 & -1 \end{pmatrix}$$

The fact that our short notation for vectors usually does not include the basis, but let's do that now for a short moment so that we can check our result

[4] 'steps' instead of 'feet' because a diagonal step that takes as one foot north and one foot south is a little more than a feet, which we don't want to confuse us with here

without completely confusing ourselves. So our above example of a vector x which represents 12 feet north and 3 foot west is

$$x = \begin{pmatrix} 12 \\ 3 \end{pmatrix}_{\{u_1,...,u_n\}}$$

We can transform this vector into the basis v by multiplying it with Q:

$$x = \begin{pmatrix} 12 \\ 3 \end{pmatrix}_{\{u_1,...,u_n\}} = \left(\begin{pmatrix} 0 & -1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} 12 \\ 3 \end{pmatrix} \right)_{\{v_1,...,v_n\}} = \begin{pmatrix} -3 \\ -15 \end{pmatrix}_{\{v_1,...,v_n\}}$$

which is what we expected it to be.

Once we have found the transformation Q from $u_1, ..., u_n$ to $v_1, ..., v_n$, the transformation in the other direction is simple. It is inverting the effect of the linear transformation, so not surprisingly, the back-transformation from $v_1, ..., v_n$ to $u_1, ..., u_n$ is Q^{-1}, the inverse of Q.

1.4.4 Basis Change on Matrices

Once we have found Q that translates a vector from one basis into another, we can multiply Q to any vector without having to start thinking again. In fact, Q does one more thing for us: It not only translates vectors, but also matrices. So we can use Q to translate a transformation described in one basis system into another. For example, assume a researcher measured cognitive skill 'before' and 'after' a cognitive training treatment. He also surveyed the satisfaction of the participants with the training and found that satisfaction is twice the skill before minus the skill after the training. This corresponds to a 1×2 matrix A,

$$A = \begin{pmatrix} -2 & 3 \end{pmatrix}$$

If the researcher is more interested in writing papers than understanding humans, he may stop here and report something of the style 'participants who are more clever before the training are less happy with the program, while after the program more clever people are more happy'. Although true, he may have missed something important here.

Another researcher wants to re-analyze the data, but his previous data sets are set up as 'intercept' and 'change', so he asks a research assistant to translate the data. The student is clever and knows that this translation is a matrix multiplication with

$$Q = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$$

You may want to check that Q is the correct transformation data. Now, after doing all the work, the student is curious what happens to A in the new basis. To do that, he translates a vector from the new basis system to the old (which is, multiplies the vector with Q^{-1}), and then multiplies the result with A. The linear map A represented in the new basis system on a vector x is

$$Bx = A(Q^{-1})x$$

which we can write as

$$Bx = (AQ^{-1})x$$

So the transformation in the new basis system is AQ^{-1}. Doing the multiplication, we see

$$B = AQ^{-1} = \begin{pmatrix} -2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 3 \end{pmatrix}$$

Observer that B tells a new story: Participant who increase a lot like the program. In addition to that, participants who start up high like the program *even more*, not less as has been reported before. In fact, the negative first entry above is an artefact from the fact that people who start high increase less if we fix the second basis to be simply the second measurement. This is a good moment to put the script aside for a moment and think about what this means.

Mathematically, we see that we can change the basis of the source space of a linear map by multiplying Q^{-1} from the right. Observe that analogously, we can change the basis of the target space by multiplying the basis change Q_{target} for that space from the left. If the target space and source space are identical, both Qs are the same matrix. We can put this together in a final theorem:

Theorem 1.4.1. *If A is a linear map from a vector space into itself in terms of a basis $u_1, ..., u_n$ and Q a basis change matrix from $u_1, ..., u_n$ to $v_1, ..., v_n$, then*

$$QAQ^{-1}$$

is the same linear map in terms of the basis $v_1, ..., v_n$.

Chapter 2

Probability Theory

Numbers are essential to describe humans, but almost no number is ever correct without specifying the uncertainty about this number. Also, science is useless without statements, e.g., that a certain treatment helps. However, these statements hardly ever are true under all circumstances, no matter whether these circumstances are known (e.g., whether the person is question is male), unknown (e.g., whether the person in question has an undetected brain tumor), or themselves uncertain, e.g., whether the person is a human (most likely, but who knows) or a demon in disguise (less likely, but scientists never dismiss a possibility).

Such uncertain values - be it numbers or true/false statements - are called *random variables*. Those come with a space of possible values and a *distribution* which tells us how likely different possible ranges of values (called *events*) are. We will discuss how we can compute distributions with or without knowledge of some of the circumstances. As before, we will try to communicate as much about a distribution as we can in as few numbers as we can, which are called *moments* of the distribution. Finally, we will get some first touch of data using *statistics* of distribution, which are random variables that, if frequently enough measured, approach the moments and so disclose us the distribution from the data.

The math is easier than in the previous chapter, but our intuition is constantly tricked in probability theory. Boldness again does the trick: Boldly making mistakes (which is a necessity when making anything at all), and bravely accept that it is wrong once the error is detected.

2.1 Random Variables

Terms: Basis transformation
Skills: Creating basis transformations, applying basis transformations to vectors and matrices
Understanding: Changing a basis is a linear map.

2.1.1 Random Variables

Let's say we have a die that shows the same number on all six sides, which is x. Then x is a variable; even if we don't know which number it is that is shown, we know it is exactly one. The outcome of a real die with the numbers from one to six, on the other hand, is a *random variable* X. It has six possible values (called the *universe*); instead of being a value, it is a function from an interval of numbers into the universe:

Definition 2.1.1. A random variable X is a function from the interval $[0, 1]$ into a set Ω. [1]

The best way to think of the $[0, 1]$ interval is your *believe space*, and the random variable is your way of distributing your believe over the universe of possible outcomes. Note that Ω can be any set, not only integer numbers as for the die, but also continuous numbers (e.g., the intelligence of a person), or truth values (that is, $\Omega = \{true, false\}$ is a set with two elements), or - to link to the previous chapter - multidimensional vector spaces. In this case - or more broadly, whenever Ω allows an operation - the space of random variables on this Ω allows the same operation. For example, $5X$ is the random variable you get when multiplying the result of X by 5, and $X + Y$ is the random variable that you get when adding the random outcomes of X and Y.

2.1.2 Elementary Events and Events

Every element of Ω, so a possible outcome of the random variable, is called an *elementary event*. The subsets of Ω, that is, all conceivable combinations of elementary events, are called *events*.

Definition 2.1.2. Let X be a random variable on Ω. The elements of Ω are called *elementary events*, the subsets are called *events*.

For a die, the outcome '1' or '2' would be elementary events. Note that these are of course also events in particular, but the term event is stronger: For example, all even outcomes ($\{2, 4, 6\}$) is also an event, or all numbers below 3 (which is, $\{1, 2\}$). For another example, assume you are supposed to guess the height of a person behind a closed door. In theory, 'a person' could refer to a baby, but I guess the event of all values below 2 ft of length don't cover much space in your believe space. Also, the event of values above 10 ft will not cover much believe space (even though of course no one explicitly excluded the 'person' is a cartoon giant from a fairy tale). The event between 5 and 6 ft, on the other hand, will cover an overly high proportion of your believe space. An

[1] This, again, is a simplification. The random variable also needs a σ algebra of Ω (which roughly speaking are some of the subsets of Ω that act like open sets), and the origin of every element of this σ algebra under X must be a member of the Borel set of $[0, 1]$, where the Borel set is the σ algebra of all open subsets of $[0, 1]$. We ignore this issue here since all random variables in social sciences satisfy these conditions anyway. So when we say 'subset' in the following, we assume it to be a member of a suitable σ algebra, and when we say 'random variable', we mean the random variable with this σ algebra.

elementary event in this example would be that the person is 5ft 6in, precisely to the nanometer.

Since events are sets, we use typical set symbols to describe combinations of events: $A \cap B$ is the intersection of both sets, so the event that both A and B happened. $A \cup B$ is the event that A, B, or both happened. $\overline{A}$ (sometimes in the literature also denoted as $\neg A$) is the event that A did not happen, so every element of Ω that is not in A.

2.1.3 Probability

The *probability* of an event is the range of numbers in our believe space $[0, 1]$ that are mapped to this event. Let's specify a random variable X that represents a fair die, which is a die for which every elementary event has the same probability. X could for example be the function:

$$X(x) = \begin{cases} 1 & 0 \leq x < \frac{1}{6} \\ 2 & \frac{1}{6} \leq x < \frac{2}{6} \\ 3 & \frac{2}{6} \leq x < \frac{3}{6} \\ 4 & \frac{3}{6} \leq x < \frac{4}{6} \\ 5 & \frac{4}{6} \leq x < \frac{5}{6} \\ 6 & \frac{5}{6} \leq x \leq \frac{6}{6} \end{cases}$$

To compute the probability of an event, we count the range of elements in the believe space $[0, 1]$ that point to this event. Take for example the event to roll a '1' or '2', which is the subset $\{1, 2\}$ of Ω. All values from 0 to $\frac{2}{6}$ point to this event, so its probability is $\frac{1}{3}$. The range in the believe space does not need to be connected; for example, the probability of the event $\{2, 5\}$ is also $\frac{1}{6}$.

Definition 2.1.3. For a random variable $X : [0, 1] \to \Omega$, the probability of an event A, denoted by $P(X \in A)$ or short $P(A)$ if the random variable doesn't need mentioning, is the range of all elements in [0,1] that map to this event.

The probability of an event is often given in parts of hundreds, of Latin 'per cent', denoted by %. The sign means nothing else than it's literal translation, i.e., $\frac{1}{100}$. So '80%' means $\frac{80}{100}$ or 0.8; all three are equal.

2.1.4 Probability Zero and Impossible is Not the Same

The construction of probabilities of events over the function X seems a little unnecessarily complex at first glance, but it is worth the thought effort; many things that are true in probability theory are mind-twisting when trying to grasp them directly, but fairly simple in this setting. For an example, remember the elementary event that a person behind a closed door is 5ft 6in. Since height is a continuous value, the random variable X that describes the height maps the numbers of $[0, 1]$ to the possible values of the height 1:1, so the probability of the elementary event (which is the *range* of numbers that point to 5ft 6in) is zero. In fact, we don't expect any given value to be precisely right to the nanometer if we measure well enough. So absurdly, every elementary event of

this random variable has zero probability, but still the random variables and the definition of probability (for non-elementary events) make perfect sense.

There are two special events: All of Ω is called the *certain event*, and the empty set is called the *impossible event*. Note that of course, the certain event has probability one, and the impossible element has probability zero. We have already seen, though, that there could be elements that are not the impossible event and still have zero probability - for example the height of 5ft 6in for a person. That is confusing at first; think of it like this: The impossible event even after the measurement can not have happened. An event with zero probability was inconceivable before the measurement (e.g., that we can name a persons height by nanometer, or higher, precision), but after the measurement, one of the zero probability events has actually happened.

2.1.5 Conditional Probability

Assume secretly rolls a die covered by his hands. Your believe space will assign every number the same range, so the probability for the event 'even' (which is the set $A = \{2, 4, 6\}$) is $P(A) = \frac{1}{2}$. But assume the person who rolled the die peeks at it and reveals to you that the number is between 1 and 3, in other words, that the event $B = \{1, 2, 3\}$ has happened. We denote the probability for the event A under this knowledge by $P(X \in A|X \in B)$, or again in short notation $P(A|B)$ (which reads 'the probability of A given B'). This probability is obviously the ratio of the part in our believe space covered by A and B divided by the part covered by B:

Definition 2.1.4. The probability of an event A given an event B is

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

Conditional probability is the only way that we can accumulate knowledge in social sciences. For us, the event A typically is a psychological hypothesis (e.g., my treatment works), while B is the outcome of an experiment (e.g., the treatment improved cognitive scores for 90 % of my sample). After knowing the outcome of the experiment, which is a undeniable hard fact, we can now update the probability that the treatment works 'in general' from our initial guess $P(A)$ to $P(A|B)$.

2.1.6 Independence and Computation Rules for Probability

If in the above example B (the experiment) would not change the probability of A, then we call A and B *independent*.

Definition 2.1.5. A and B are called independent exactly if $P(A) = P(A|B)$.

Independence means that the outcome of one event provides no information about another event. This makes the concept very intuitively; for example, if we roll a die twice, we probably can assume that events defined by the first die (e.g., the first die is '1', or the first die is even) are independent of events defined by the second die (e.g., the second die is higher than 4).

Note that in this case,

$$\begin{aligned} P(A) &= P(A|B) \\ &= \frac{P(A \cap B)}{P(B)} \\ \Leftrightarrow P(A \cap B) &= P(A)P(B) \end{aligned}$$

so equivalently, we could define independence as saying that the probability for the event A and B is the product of the single probabilities,

Theorem 2.1.6. *If A and B are independent, then*

$$P(A \cap B) = P(A)P(B)$$

Note that this is not necessarily true if A and B are dependent.

If two events A and B are *exclusive*, they share no common elementary events:

Definition 2.1.7. Two events A and B are called *exclusive* if they share no element, that is, $A \cap B = \emptyset$.

In particular, A and B share no space in our believe space $[0, 1]$. Therefore, the range in our believe space of $A \cup B$, the union of both, is the sum of both probabilities:

Theorem 2.1.8. *if A and B are exclusive, then*

$$P(A \cap B) = P(A) + P(B)$$

Also exclusiveness is easier to understand formally, it is in practical examples less easy to determine if two events are exclusive. The easiest way of doing that is to check whether B can still happen if we know that A already happened, and vice versa; if for both directions one event excludes the other from happening, then the events are exclusive in the definition above. You can check this by the above definition of conditional probability if you use the fact that $P(\emptyset) = 0$, that is, the impossible event has always probability zero.

2.2 Bayes Theorem

Terms: The Theorem of Total Probabilities, Bayes' Theorem, Bayes' Fallacy
Skills: Computing Probabilities from partition of events, computing conditional probabilities from the inverse probability
Understanding: Probabilities of hypotheses change if experiment results are given, hypotheses are not 'proved' or 'disproved'. The effect of a-priori knowledge is typically underestimated.

2.2.1 The Theorem of Total Probabilities

Assume the weather forecast predicts 40 % chance of rain and 10 % chance of hailstorm (and 50 % sunshine) for tomorrow, when your neighborhood picnic is scheduled. The picnic will be canceled in hailstorm, and with 75 % chance if there is rain. Let call the event 'picnic takes place' A, and the three weather events B_1 (sun), B_2 (rain), and B_3 (hailstorm). Observe that since one of the three weather conditions must occur, we have

$$A = (A \cup B_1) \cap (A \cup B_2) \cap (A \cup B_3)$$

in other words, the event that the picnic takes places is one of the events that picnic takes place in sun, rain, or storm. All three unions are exclusive (since the weather conditions are exclusive), so

$$P(A) = P(A \cup B_1) + P(A \cup B_2) + P(A \cup B_3)$$

again in natural description, the probability that the picnic takes place is the probability that the picnic takes place in sun, that it takes place in rain, or that it takes place in storm. In this situation, we don't know the joint probabilities, but we do know the conditional probability (which are, the probability that the picnic takes places given the three weather conditions); using this, we can transform the line to

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)$$

and finally compute the probability for our picnic:

$$P(A) = 100\% \cdot 50\% + 25\% \cdot 40\% + 0\% \cdot 10\% = 60\%$$

It is fairly frequent that, as in our example, we know a probability for an event under some conditions which, again as in our example, are *exclusive* and *complete* in the sense that they fill the full space and don't overlap. The above computation can then be performed to compute the total probability of the event. Because mathematician have no inventivness if it comes to names, this computation rule is called the *Theorem of Total Probabilities*:

Theorem 2.2.1. *If some events $B_1, ..., B_n$ are a partition of Ω (that is, they are exclusive and together are Ω), then for every event A,*

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

One of the most frequent choice of $B_1, ..., B_n$ is an event and its complement. For example, the events 'the hypothesis is true' (B) and 'the hypothesis is false' ($\overline{B}$) obviously satisfy the condition that they are exclusive (if the hypothesis is true, than it is not false), and also fill the whole space (either the hypothesis is true or not, nothing else is possible).

2.2.2 Bayes' Theorem

The definitions for events and probabilities are, for mathematical terms, new. Therefore, the definition of conditional probability has a reformulation which somehow stole the name *Bayes' Theorem*, although it is almost exactly the original definition:

Theorem 2.2.2. *For any events A and B,*

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}$$

Sometimes, Bayes' theorem is written by replacing the denominator with the Theorem of total Probability using A and $\overline{A}$ as the two pieces:

$$P(A|B) = P(B|A)\frac{P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})}$$

2.2.3 Example: Hypothesis Probability Update

Although Bayes' theorem is very close to the original definition, it is in fact most important in social sciences for the same reason already mentioned at the definition of conditional probability. In social sciences, A often is an hypothesis, e.g., 'mean brain volume increases from age 10 to age 20 by a factor of 1.1'. B is often a concrete, substantial outcome of a real data set, e.g., 'the average brain volume of my 100 participants (Joey, Michael, Fred, Anne, Kathy,...) increased at least by a factor of 1.25'. What we are interested now is the probability that A is true given that B is the outcome of our experiment, which is called the *a-posteriori* probability of the hypothesis.

We need a *model*, a section for later in this script, that tells us what the probability for this outcome is given that the hypothesis is true ($P(B|A)$) or false ($P(B|\overline{A})$). Let's take these two values for granted by now, say they are

$$\begin{aligned} P(B|A) &= 0.3 \\ P(B|\overline{A}) &= 0.1 \end{aligned}$$

In natural language, the probability to find an increase of 1.25 or higher is 30 % if the hypothesis is true, and 10 % if it is not true. We also need the probability $P(A)$ that the hypothesis is true under no other event, that is, if we have no information about the experiment outcome. This is the probability that a naive person would assume for A, and is called the *a-priori* probability for A. For this example, let's say that $P(A) = 0.2$. With this, we can compute

$$\begin{aligned} P(A|B) &= P(B|A)\frac{P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})} \\ &= 0.3\frac{P(A)}{0.3P(A) + 0.1(1 - P(A))} \\ &= 0.3\frac{0.2}{0.3 \cdot 0.2 + 0.1 \cdot 0.8} = 0.429 \end{aligned}$$

So the probability for the hypothesis has changed from 20 % to 42.9 % by our experiment.

Note that even though our experiment worked in favor of the hypothesis (since the outcome we found is three times more probable under the hypothesis than without it), the probability for the hypothesis after the experiment is still below 50%. This is due to the fairly low a-priori probability we assigned to the hypothesis. If the a-priori probability for the event would have been $P(A) = 50\%$, that is, both guesses are equally likely, you can quickly check that $P(A|B) = \frac{P(B|A)}{P(B|A)+P(B|\overline{A})} = \frac{3}{4}$.

2.2.4 Bayes Fallacy

Intuitively, we are tempted to forget the a-priori probability when we think about a case as above. This can in fact have horrible consequences. A classical example is the following: Suppose the witness of a car accident states that a black taxi caused the accident and afterwards flew from the scene. Suppose further that 85 % of all taxis in the town in question are yellow, 15 % are black. The attorney of the defense questions the ability of the witness to separate the two colors (maybe it was a dark night), so a tests is performed that shows that the witness recognizes the color of taxis in similar conditions 90 % of the cases correct (in both directions, so he correctly identifies 90 % of the yellow taxis as yellow and 90 % of the black taxis as black). Put the book aside for a moment to make a guess at the probability that the taxi involved in the accident was black.

Most people will answer 90 %, based on the fact that the witness has been tested to be 90 % correct and identified the taxi as black. This would only be true if yellow and black taxis were equally likely a-priori, which they are not since we know yellow taxis are more frequent. This neglect is called the *Base Rate Fallacy*. But even if we do consider the base rate, the quantity of the effect is typically still underestimated. Let's compute the correct answer; let's call 'isB' the event that the taxi really is black, and 'saysB' that the witness says the taxi is black. We know

$$\begin{aligned} P(isB) &= 0.15 \\ P(saysB|isB) &= 0.9 \\ P(saysB|\overline{isB}) &= 0.1 \end{aligned}$$

The first line is the a-priori probability that (without any witnesses) the taxi would be black. The second two lines are the result from the experiment: The probability for the witness to call the taxi black is 90 % if it is black, and 10 % if it is yellow. We are interested in $P(isB|saysB)$, which by Bayes theorem is

$$\begin{aligned} P(isB|saysB) &= P(saysB|isB)\frac{P(isB)}{P(saysB)} \\ &= P(saysB|isB)\frac{P(isB)}{P(saysB|isB)P(isB) + P(saysB|\overline{isB})P(\overline{isB})} \\ &= 0.9\frac{0.15}{0.85 \cdot 0.1 + 0.15 \cdot 0.9} = 0.61 \end{aligned}$$

Even though the witness is fairly good at detecting colors, and the a-priori rates are not extreme in the sense that you could say 'there hardly are any black taxis', the probability after the witness' statement is not far away from 50:50 guessing. The probability is by a long shot not close enough for a conviction.

2.3 Distribution and Density

Terms: Probability Distribution, Probability Density
Skills: Computing Distributions from Densities and the other way round
Understanding: Distributions are probabilities to be at a point or below the point, and densities are the probability to be close to a point.

2.3.1 The Cumulative Distribution Function and the Density Function

When X is the height of a person, we realized that every single point of Ω, every elementary event, has probability zero, roughly because no one what ever guess the height of a person correct by infinitely many digits. But 5ft 6in is still more 'likely' than 20ft, even though both have probability zero. Luckily, if Ω is a vector space, there is a convenient way to express this intuition that comes in two steps: First, we talk about the probability that the height is *smaller or equal* 5ft 6in, which is the event of all outcomes up to 5ft 6in. In multidimensional vector spaces (say if we are concerned about height and weight), we mean by *smaller or equal* (6ft 5in, 170lb) every point that is smaller on both dimensions. This is called the *cumulative* probability of a point:

Definition 2.3.1. Let X be a random variable over an $\mathbb{R}$-vector space. The function

$$F(x) = P(X \leq x)$$

is the *Cumulative Distribution Function* (sometimes called *cumulative probability* or just *probability distribution*) of X

The Cumulative Distribution Function always approaches zero for very low value, and one for very high values. Somewhere in between, it increases from zero to one. Observe that the function increases only very little at points that are intuitively 'unlikely'. In our example, the probability that a person is below 2ft is almost as close to zero as that the person is below 3ft. Also, that the person is below 20ft is almost as large as that the person is below 21ft - very close to one for usual humans. However, the probability that a person is smaller than 5 ft is much less likely than that the person is smaller than 6ft, because many humans lie in between these two values. So the increase of the Cumulative Distribution is much steeper at 5ft 6in than at 20ft, exactly what we were hoping for. Mathematicians call the steepness of the increase of a function the *derivative* of the function, which for Cumulative Distributions is called the *Density*:

Definition 2.3.2. The norm of the first derivative of F,

$$f(x) = |\partial F(x)|$$

is called the *Probability Density Function* (sometimes just called the *Density*, or the *likelihood*).

So the Cumulative Distribution is the probability to be below a point, and the density at that point is the increase of this probability. If Ω is multidimensional, then F has multiple arguments (one for every dimension). The term '$\partial F(x)$' then is a vector, where the first entry is the derivative of F with respect to the first basis vector, the second with respect to the second, and so on. Remember that the $|.|$ indicate the norm of the vector. Geometrically, F can be seen as a surface that looks much like a hill you could ski on (with the hilltop at the upper right, and the valley at the lower left). $|F(x)|$ is the steepness in the steepest direction at every point.

2.3.2 Density is the Probability to be Close to a Point

The density has another intuitive meaning: The probability to be close to a point is the density at that points multplied by the the range that you consider 'close'. If we think of the height X of a person again, and assume that the density of 5ft 6in is 0.8. This means that approximately, the probability to be in the range of one foot around 5ft 6in, i.e. between 5ft and 6ft, is $0.8 \cdot 1 = 80\%$. The probability to be at most half a foot away from 5ft 6in is approximately $0.8 \cdot 0.5 = 40\%$, the probability to be at most a quarter of a foot away is 20%, and so on. The smaller the range is chosen, the better is the approximation.

The reason for this is simple: Since the CDF (short for 'Cumulative Distribution Function') is the probability to be below a point, so the probability to be between two points (e.g., 6ft and 5ft) is $F(6ft) - F(5ft)$. This difference is the change of F from 5ft to 6ft, which for small ranges is given by the derivative of F. Put the book away for a minute to understand this reason.

2.3.3 Moving between Cumulative Distribution and Density

Cumulative Distribution Function and Density are two close siblings. If the Ω of a random variable has a reasonable $\leq$ order and allows a derivative, then either of the two can be used to describe the probability function of the random variable. We don't need both of them, but they come with different advantages for our intuition, so it's worth to look at how to move from one representation to the other.

By definition, the density is the derivative of the CDF. This implies that the CDF is the area under the density, which is also intuitive since we can think of the CDF as *cumulating* the probability of being below a point. In mathematical terms, the 'area' means the integral under the curve:

Theorem 2.3.3. *Let F be the CDF and f be the density of a random variable over a vector space. Then,*

$$f(x) = |\partial F(x)| \quad \text{and} \quad F(x) = \int_{-\infty}^{x} f(x)dx$$

2.3.4 Intuition for Cumulative Distribution and Density

Both CDF and Density have their benefit as representations of X. The CDF represents probabilities, so it is always between 0 and 1. It is easy to translate into a probability statement; if $F(x)$ is 0.2, that means we have a 20 % of being at or below x. The density is also always above 0 (since the CDF is monotonously increasing), but can exceed 1, since to translate the density into the probability to be close to a point, we have to multiply it with a small range that indicates what we mean by 'close'. However, the density is very useful if we want to compare to points: 5ft 6in in our example has higher density than 20 ft, which means it is more likely to be close to 5ft 6in than to be close to 20ft.

This means that while the CDF always is a probability, the density comes with a less intuitive unit (which loosely speaking is 'probability per area of Ω'). In particular, if we change the basis of the vector spaces (e.g., express the height of a person in inches instead of feet), the CDF is just stretched, but the density actually changes values; in our example, since 12 inches are one foot, the density would be reduced by a factor of 12. Check that this is consistent with both interpretations of the density: If the CDF is stretched, it is 12 times less steep at every point, so its derivative must be reduced. Also, the probability to be 1in close to a point is of course less (for small enough ranges, 12 times less) than the probability to be 1ft close to the same point.

So the density most often comes as ratios of the density of two points, and then its interpretation is again very simple. If the density at 20ft is 40 times smaller than the density at 5.5 ft, then this means that 20ft is 40 times less likely. Strictly, we should say that 'a small range around 20ft is 40 times less likely than the same range around 5.5ft', but since the ratio stays the same for any range, we omit it in natural language. Observe that although single points in the CDF are very helpful in interpretation, ratios of two different points don't give us any valuable information.

Over discrete subsets of vector spaces, for example the outcomes of a die $\{1, 2, 3, 4, 5, 6\}$, CDF and density are defined by assuming that Ω would be all real numbers (or the whole vector space in multiple dimensions), and the believe for all other numbers are zero. So we treat the die as if we could roll a 7, but do so with zero probability. If we do this, the CDF is constant zero at all points before 1 and jumps from 0 to $\frac{1}{6}$ as we pass the 1. The density is obviously zero for every number that is not $\{1, 2, 3, 4, 5, 6\}$, but mathematicians have a lot of fun defining the density at these points: Obviously, it is infinity (since the CDF raises in a step from 0 to $\frac{1}{6}$), but at the same time the integral over this function is $\frac{1}{6}$. You can think of the density as a flat line that has infinitely high spikes at $\{1, 2, 3, 4, 5, 6\}$ such that the integral over this spike is $\frac{1}{6}$. Although not mathematically correct, it become custom to denote these such a spike at position 1 with integral $\frac{1}{6}$ as $\frac{1}{6}\delta_1(x)$. If we allow this notation, we can use CDF and density equally on discrete and continuous random variable.

2.4 Moments

Terms: Expectation, mean, variance, standard deviation, moments, covariance matrix, correlation
Skills: Computing mean and variance from a distribution
Understanding: A distribution can be described efficiently by its moments, which are mean, variance, and some more. *Covariances* describe how two variables move together, and correlation are standardized covariances.

2.4.1 The Expected Value

Assume we want to describe the distribution of a fair die X by a single number. Since X is a whole function, we actually can't do that, but we can find a number which describes our expectation of the outcome best. Note that again, following the general rules of science, we aim at simplifying something to represent it more easily (this time, in a single number). What is typically done to get at this as-informative-as-possible value is to add up all possible outcomes of the die, weighted by the probability of that outcome:

$$\mathbb{E}(X) = \sum_{x=1}^{6} xP(X = x) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + ... + 6 \cdot \frac{1}{6} = \frac{1 + 2 + ... + 6}{6} = \frac{7}{2} = 3.5$$

This value is called the *Expected value* of the random variable X, our die. Note that the probability for 3.5 is actually zero, so even though 3.5 is the most efficient description of the die, it is a value that actually never occurs in a single roll. It is still our 'best guess' if we want to be as close as possible to the true result.

Assume a second die Y is not fair, but shows six with probability $P(Y = 6) = \frac{1}{2}$ and all other five values with probability $\frac{1}{10}$. Then our expected value should be higher; again, we add up the possible outcomes and weight them by their (now different) probability:

$$\mathbb{E}(X) = \sum_{x=1}^{6} xP(X = x) = 1 \cdot \frac{1}{10} + ... + 5 \cdot \frac{1}{10} 6 \cdot \frac{1}{2} = \frac{9}{2} = 4.5$$

For a general, potentially continuous, distribution, we have to sum over all infinitely many outcomes and use the infinitely small probability to hit this number. Luckily, we have already the concept of the density to our disposal; so in general, the expected value is defined as

Definition 2.4.1. Let f be the density of a random variable X over a vector space Ω. Then,

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} f(x)xdx$$

is called the *Expected Value* of X, also called the *mean* or *first moment* of X.

We can think of the integral as a big sum where the steps of the sum is infinitesimally small. In fact, for a discrete distribution like in our example above, the integral becomes a finite sum over all possible outcomes of Ω with the probability instead of the density,

$$\mathbb{E}(X) = \sum_{x \in \Omega} P(x) x dx$$

What if X is a random variable over a multi-dimensional vector space, for example, five cognitive variables we are interested in? Again, we profit from our general representation of vector spaces: The above equations work equally well if x is a number or if x is a vector. The integral is then a integration in all basis directions. Note that in this case, the expected value also is a vector with five dimensions, which is consistent with our intuition because the 'best bet' of the outcome should have as many dimensions as the outcomes themselves.

We have previously said that we can add random variables over vector spaces, or multiply them with scalars. For example, if we have two random variables X and Y, the random variable $X + Y$ would be the outcome of X plus the outcome of Y. Because the integral can be treated like a little overly proud sum, we can exchange the order of the addition and the integral:

$$\int f(x)(x+y)d(x+y) = \int f(x)x + \int f(y)dy$$

And the same holds for multiplications with scalars. We can conclude that the expectation is a linear function:

Theorem 2.4.2. *Let $\lambda \in \mathbb{R}$ be a scalar and X and Y random variables on the same vector space. Then,*

$$\begin{aligned} \mathbb{E}(X+Y) &= \mathbb{E}(X) + \mathbb{E}(Y) \\ \mathbb{E}(\lambda X) &= \lambda \mathbb{E}(X) \end{aligned}$$

So the expectation doesn't care whether we take it before or after doing additions or scalar multiplications. This observation makes our lives a lot easier.

2.4.2 Moments

The expected value of a random variable is also called its *mean*, or its first moment. The last name suggest that there are more moments, and in fact there are for variables over Ω that allow an exponential; here, the 'moments' are defined as

Definition 2.4.3. Let X be a random variable over a space that allows exponential with density f. Then, the nth *moment* of X is defined as

$$M_n(X) = \mathbb{E}(X^n)$$

The nth *central moment* is

$$C_n(X) = \mathbb{E}((X - \mathbb{E}(X))^n)$$

The 2^{nd} central moment is also called the *variance*, denoted by $\mathbb{V}(X)$. The square root of the variance is called the *standard deviation*.

The moments are a consequent continuation of the principle 'more bang for our bug', because intuitively each moment carries maximal information about the distribution if the previous moments are already known. In other words, if someone already told me the mean of a random variable and I still need more information (in as few numbers as possible), I will ask for the 2^{nd} moment, and then for the 3^{rd}, and so on.

The central moments are a little more intuitive than the raw moments. For example, the variance is the 'expected squared distance from the mean', that is, a measure how wrong I will be if I bet my money on the mean. The central moments are simple transformations of the raw moments. To get practice with computations with expectations, we will compute the variance in the following. Note that $\mathbb{E}(X)$, the expectation, is not a random variable.

$$\begin{aligned} \mathbb{V}(X) &= \mathbb{E}((X - \mathbb{E}(X))^2) \\ [\text{expanding square}] &= \mathbb{E}(X^2 - 2X\mathbb{E}(X) + \mathbb{E}(X)^2) \\ [\mathbb{E} \text{ is linear}] &= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + \mathbb{E}(X)^2 \\ &= M_2(X) - \mathbb{E}(X)^2 \end{aligned}$$

Take a moment to follow this computation. From the first to second line, we expand the square by the binomial equations. In the second to third line, we pull the addition in the expectation apart (which we can since the expectation is linear), and pull the $\mathbb{E}(X)$ out of $\mathbb{E}(\mathbb{E}(X)X)$ in the middle summand, again possible because $\mathbb{E}$ is linear. In the last summand, the expectation of a constant number is obviously that number, so $\mathbb{E}(\mathbb{E}(X)) = \mathbb{E}(X)$. in the final step, we add the second two summands.

2.4.3 The Covariance Matrix

In social sciences, our Ω is often a multidimensional vector space. Vectors in fact have an exponential, although we only have defined the 'square', and that in two different ways (the inner and outer vector product). For the variance of vectors, we use the outer product, so that the variance of a random variable over a vector space is

$$\mathbb{V}(X) = \mathbb{E}(xx^T) - \mathbb{E}(x)\mathbb{E}(x)^T$$

Observe that while the mean of an n-dimensional vector-valued random variable is a vector, the variance is a $n \times n$ matrix. [2] This matrix is also called the *covariance matrix* of the random variable. It is typically denoted by a Greek upper-case Sigma,

$$\Sigma = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & ... & \sigma_{1,n} \\ \sigma 1,2 & \sigma_{2,2} & ... & \sigma_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,n} & \sigma_{2,n} & ... & \sigma_{n,n} \end{pmatrix}$$

[2] For the interested readers, yes, there is a exponential to the power of three for vectors. The result is something like a three-dimensional matrix, a third-level tensor, and also a fourth and all other exponentials. For social sciences, these objects are completely ignored up to this day.

The covariance matrix is symmetrical. As an exercise, it is worthwhile to check that $\sigma_{i,j}$ is

$$\sigma_{i,j} = \mathbb{E}((X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j)))$$

which means, it is the expected product of the ith and the $j$the entry of X treated as single-dimensional random variables, after correcting for their corresponding means. This term even receives an own name,

Definition 2.4.4. For two random variables X and Y, the *covariance* of X and Y is defined as

$$cov(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

In particular, the entry $\sigma_{i,i}$ is the variance of the ith entry of X.

The naming in this field of math is a little confusing: The *variance* of a multi-dimensional random variable is a matrix which is called the *covariance matrix*, and the entries of this matrix are in turn the *variances* and *covariances* of the single dimensions. The notation in the literature is also not always helpful: The diagonal elements of the covariance matrix are sometimes denoted with squares (so 'σ_1^2' instead of $\sigma_{1,1}$).

The covariances have a strong intuitive meaning already foreshadowed by their name: They describe to what extent two variables vary together. A high positive covariance means that whenever one of the two variables is high, the other one tends to be high, too, both relative to their means. A high negative covariance on the other side indicates that if one variable is below its mean, the other one will tend to be above the mean. If the covariance of two variables is zero, then no prediction from one of them can be made whether the other one will, in expectation, be above or below its mean. Observe that this is not the same as the concept of *independence* introduced earlier; if two random variables are independent, then the outcome of one carries no information of the other. In this case, the covariance will be zero. The covariance can, however, be zero even for dependent random variables. For an example, think of a situation in which we measured how fast a runner is, and how much he approaches a given target. Assume the runner doesn't know in which direction the target is, and has an equal chance of running away from it as running towards it. Then the covariance of speed and approaching the target is zero, since higher speed can equally likely mean that the runner moves faster away from the target, or towards it faster. In expectation, the mean has no influence on the change in the distance of the runner to the target, but it is still not independent as higher speed values of course indicate higher absolute values in this change.

Another useful exercise is to verify that the absolute covariance between two variables will never exceed the *geometric center* of both variance,

Theorem 2.4.5.

$$|cov(X, Y)| \leq \sqrt{\mathbb{V}(X)\mathbb{V}(Y)}$$

If the covariance is at the geometric center, they covary perfectly, which means that if you know one, you can compute the other without any remaining random effect. Because of that, the covariance as a descriptive of a random

variable is often normalized by the variance, which is then called the *correlation* of the two random variables:

Definition 2.4.6.

$$cor(X,Y) = \frac{cov(X,Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$$

Analogously to the covariance matrix, there exists a correlation matrix which instead of the covariances has the correlation of X_i and X_j at position (i,j). Obviously, the diagonal elements of a correlation matrix are always one, since each variable correlates with itself perfectly.

The correlation is bounded between -1 and 1, and independent of the actual variances of the random variables. It is also independent of re-scaling one or both of the variables, since the variance will change to the same extend. This in some sense gives the correlation an absolute character that the covariance is missing, in a sense that a correlation of 0.8 can always be considered 'high' covariation, while a covariance of 0.8 depends on the variances of both variables to allow a classification as 'high' or 'low'. However, whether that is really the case often depends on the actual situation; a researcher should always be aware of the covariance and the variances of two variables rather than the correlation.

2.5 Statistics

Terms: Average, Large Number Theorem, statistics
Skills: Computing descriptive statistics (average, estimated standard deviation, ...)
Understanding: Moments are latent properties of random variable which are not directly accessible. Statistics are numbers computed from data that, for sufficiently many data points, approximate the moment.

2.5.1 The Average and the Large Number Theorem

Assume we have a (potentially unfair) die and want to know the mean of it. We can roll the die multiple times and take the *average* of the results

Definition 2.5.1. Let $X_1, ..., X_N$ be N identical random variables over a vector space. Then,

$$Av(X_1, ..., X_N) = \frac{1}{N}\sum_{i=1}^{N} X_i$$

is called the *average* of the random variables. Analogously, for N instances $x_1, ..., x_N$, the average is $\frac{1}{N}\sum_{i=1}^{N} x_i$

Observe that the average *is not* the mean of the random variable. The average is a random variable for which one instance can be computed from a given data set. This instance is situated in the real world, it is a real object that is directly accessible to the researcher. The mean on the other side is a non-random abstract number that comes with the random variable; it can

sometimes be computed if the distribution is known, but is never accessible from a data set. So the average (a random variable that can be computed from a real data set) and the mean (an abstract single 'correct' number) should not be confused. Having said that, what is true is that the average approaches the mean as N growth. We can conclude this from an even more general statement called the *Large Number Theorem*:

Theorem 2.5.2. *Let $X_1, X_2, ...$ be series of random variables with equal mean $\mathbb{E}(X)$ and finite variance. Then the average*

$$\frac{1}{N}\sum_{i=1}^{N} X_i$$

converges stochastically towards $\mathbb{E}(X)$ as N approaches infinity.[3]

In social sciences, we typically assume that each person is an identical random variable. We want to learn something about this distribution, for example the mean, by collecting multiple instances of the random variable. The Large Numbers Theorem connects the latent, otherwise inaccessible world of abstract random variables to real data sets; in other words, we should be very grateful to this theorem, for without it, social sciences would be completely lost.

2.5.2 Descriptive Statistics

Statistics are values which are computed from an actual observed data set, that is, from a series of identical random variables. They typically serve one or both of two purposes: Either describing the data set in fewer numbers, again following the tradition to describe something complex approximatively with something simple, or they are statistics for a parameter of the abstract distribution. The average is an example for a statistic for an abstract parameter, the mean, which we never are able to access directly. The average at the same time also serves the first purpose as it describes a *central tendency* of the data set, or in simpler words, the middle of the values. These statistics are sometimes called *descriptive* statistics. There are a number of other descriptive statistics. The following table gives a number of descriptive statistics for a data set $X_1, X_2, ...$ with identical random variables.

[3] Don't worry about the 'stochastically', what this effectively means is that with higher N, the average will be closer and closer to the mean. Also don't worry about the 'final variance'; in social sciences, all our random variables have finite variance. In fact, the Large Number Theorem is even stronger and allows infinite variance as long as it doesn't get large too quickly, but why bother if we never encounter those?

Name	Definition
average ($\overline{X}$)	$\frac{1}{N}\sum_{i=1}^{N} X_i$
pth percentile	For $0 < p < 100$, the pth percentile is the point such that the p% of the data is lower than the point; for example, if the 8th percentile is 32, then 8% of the data is below 32, and 92% is above 32.
median	The 50th percentile
quartile	the 1^{st}, 2^{nd} and 3^{rd} quartile are the 25th, 50th, and 75th percentile.
mode	For discrete Ω, the mode is the outcome that occurs most often in the data set.
estimated variance ($\hat{\sigma^2}$)	$\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2$
estimated standard deviation ($\hat{\sigma}$	$\sqrt{\hat{\sigma^2}}$)

The percentiles, quartiles and the median are typically only defined for $\Omega \subset \mathbb{R}$, although in theory they are also useful over multidimensional vector spaces. If the pth percentile is between two data points (e.g., the median of an even number of distinct results), one typically chooses the arithmetical middle of these two data points as the percentile. The estimated variance is applicable to multidimensional data sets, the result is then an estimated covariance matrix. The standard deviation is also defined in that case, [4] but is typically not used.

Descriptive statistics and different graphical representations of data sets are very important to understand a data set, so important that there enumeration has an own name, which confusingly is again *descriptive statistics*. There is actually nothing to *know* about statistics other than there definitions, but it is more than worthwhile to acquire a good intuition for what these definitions do, or in other words, what we would expect from a data set when we know some of its statistics. Since this script is more about explanation than about practicing, we will cover descriptive statistics here, even though using and understanding descriptive statistics (both the field and the values) is very important.

2.5.3 Statistics for Distribution Parameters

Statistics of the other kind, that is, statistics for distribution parameters, are the bridge between the abstract world of random variables and real-world data. As with the mean, there are multiple other parameters of distributions that we can never access. Since all statistics are to some degree descriptive of the underlying data set, the separation is a little artificial, and many statistics are considered to belong to both classes. The second class at least comes with a formal definition:

Definition 2.5.3. Let $X_1, X_2, ...$ be a series of identical random variables with a parameter θ and $\hat{\theta}(X_1, ..., X_N)$ a statistic that depends on the first N entries.

[4] For the interested reader: There are different definitions of the square root of a matrix. The one that is typically chosen here is the *Cholesky decomposition*.

We say that s is a statistic for θ if

$$\lim_{N\to\infty} \hat{\theta}(X_1, ..., X_N) = \theta$$

So $\hat{\theta}$ is a statistic for θ if it converges[5] towards θ, or in simpler words, the larger the data set that we consider, the closer the statistic is to the true population parameter.

The average is a statistic for the mean, and the 'estimated variance' given above is a statistic for the variance of a distribution. As an exercise, you may want to confirm the second statement using the Large Number Theorem. Both statistics, in addition to approaching the true value for increasing N, have another advantage: Even for small N, the expected value of the statistics are the true population parameters. This property is called a *true* statistic.

Definition 2.5.4. Let $s(X_1, ..., X_N)$ be a statistic for a population parameter θ. Then s is called a *true* statistic for θ exactly if

$$\mathbb{E}(X_1, ..., X_N) = \theta$$

for all sample sizes N.

This concept is simple, but mind-twisting at first glance, especially when we think about the average: The average (a random variable) of X (another random variable) is a true statistic for the expectation of X (which means, $\mathbb{E}(avg(X)) = \mathbb{E}(X)$). So the expectation of our measure for the expectation is this expectation. fairly confusing, but if you take a minute, you can work this out.

Average and estimated variance are simple equations for statistics that work. For many other parameters, statistics exist. Even better, there exists statistics called *universal* statistic, that is, a value computable from the data set that is always a statistic for a parameter of a distribution of the random variable. One of these universal statistics is the maximum likelihood we will get to know later in an own chapter.

2.6 The Normal Distribution

Terms: Central Limit Theorem, (standard) normal distribution
Skills: Computing the density of a normally distributed random variable
Understanding: The sum of many random variables (almost no matter which) always has the same distribution, the normal distribution. The normal distribution is therefore a very important real-world distribution.

2.6.1 The Central Limit Theorem

There are multiple distributions that occur in real life. For all of those, the cumulative distribution function, the density, and usually also some parameters

[5] again, it should precisely say 'converges stochastically'

like the expectation value can be looked up on the internet. A very condensed list is given in the Appendix, together with some rules how the distributions are linked together. Here, we will concentrate on the single most important distribution, the *normal distribution*. The reason for its importance is simple: If we have a series of independent random variables and add them up, then the sum normalized by the mean and standard deviation is normally distributed.[6] Think a moment about how important this observation is: Mostly everything we observe is a result of very many, very tiny, causes. For example, the height of a person is a result of thousands of genetical and environmental influences. To some degree, independence of these is often a reasonable assumption. Whenever that is the case, the result will be one and the same distribution. That is truly a miraculous action of mathematics in the real world. [7] The formal theorem for this observation, called the *Central Limit Theorem*, comes in many forms; we will use a very light form here:

Theorem 2.6.1. *Let $X_1, X_2, ...$ be a series of independent random variables over a vector space with variances bounded from above*[8]*. Let*

$$Y_N = \sum_{i=1}^{N}$$

be the sum of N of these random variables, and

$$Y'_N = \frac{Y - \mathbb{E}(Y)}{\sqrt{\mathbb{V}(Y)}}$$

a normalization of Y_N. Then for large N, Y'_N approaches one fixed distribution.

Definition 2.6.2. This distribution is called the standard *Normal* or *Gaussian* distribution.

2.6.2 The Standard Normal Distribution

A second miracle is that the density function of the standard normal distribution is known. It is the only equation you should know by heart. For a K-dimensional Ω (which means, for a K-variate random variable), it is

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

Let's take a close look at the equation. $\pi = 3.14...$ is the well-known constant, the area of a standard circle. The complete fraction is only a constant number independent of x. The major part is the exponential $e^{-\frac{1}{2}x^T x}$, which you can also write as $e^{-0.5|x|^2}$, so the exponential of half of the negative squared norm of x. This term is maximal for $x = 0$, which gives $e^0 = 1$. Every larger vector

[6] There are some exceptions to this rule, but they are not important in social sciences.

[7] Or however you want to phrase this, dependent on your view of the world. But whichever terminology you prefer, the effect of our mouth dropping open on this fact stays the same.

[8] Which means, there is a maximal number which is higher than the variance of each X_i

results in a number below 1, first decreasing slowly, then rapidly close to 1, and then approaching zero for very large vectors x. The complete area under this curve, not surprisingly, is $\sqrt{2\pi K}$, so that the area under f totals to one, as is required for a density function.

For the cumulative distribution function, there exists no solution that can be written more clever than '$\int e^{-\frac{1}{2}x^T x}$'; mathematicians say there exists no *close form solution* for the integral. Bad for them, but we don't care too much: We can always let a computer approximate the area under the curve numerically for us. In practical issues, we can hence assume we know the CDF of the normal function.

2.6.3 The Normal Distribution

The 'standard' in the name 'standard normal distribution' comes from the fact that the mean of the variables are zero, and the covariance matrix is the identity. In the real world, most variables are normally distributed, but not necessarily standardized (e.g., the height of humans is well approximated by a normal distribution, but the mean is obviously not zero). There are two ways of computing the density of a non-standardized normal distribution: We may always standardize our variables before looking up the standard normal density. But instead of doing this each time, we may do it once in the actual distribution. To add a non-zero mean μ, we just replace x by $x - \mu$ in the normal density equation:

$$f_\mu(x) = \frac{1}{\sqrt{2\pi K}} e^{-\frac{1}{2}(x-\mu)^T(x-\mu)}$$

To add a covariance matrix Σ, we have to transform x with $\sqrt{\Sigma^{-1}}$, which in the exponential term just adds a Σ^{-1} between the two vectors. However, this changes the area under the curve; you can comprehend this when imagining that we re-label the x-axis of the normal distributions by doubling all labels, then obviously, the area of the curve doubles. Since the density needs to have an area of one again, we have to divide by a norm of the matrix; it can be shown that the determinant of Σ divided by K does the job here. So in total,

Theorem 2.6.3. *For a normally distributed random variable stretched and translated such that the variable has mean μ and covariance matrix Σ, the density function is given by*

$$f_{\mu,\Sigma}(x) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{2.6.1}$$

Definition 2.6.4. The probability distribution described by the above density is called the *Normal* or *Gaussian* Distribution.

The density looks more complex now, but keep in mind how it is composed to see that it is actually no rocket science: The exponential term still is mostly the product of x with itself (only correct by the mean and standard deviation), and the fraction in front of it is just a normalizing constant that does not depend on x.

Just a reminder about how the density works. Assume you have a data point x (say, a vector (5,3,2)), and you want to know how likely it is to get this point (or one very close to it) for a random variable X that is normally distributed with mean μ and covariance matrix Σ, for instance

$$\mu = (4, 3, 2) \quad \text{and} \quad \Sigma = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

You then evaluate Equation 2.6.1 using these values:

$$\begin{aligned} f_{\begin{pmatrix} 4 \\ 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}} &= \frac{1}{\sqrt{2 \cdot 3.14 \cdot 2}} e^{-\frac{1}{2}(1,0,0)\begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}} \\ &= \frac{1}{\sqrt{12.56}} e^{-0.25} \\ &\approx 0.005 \end{aligned}$$

So the density for x under this distribution is 0.05. Remember what this means: Roughly, the likelihood to be in a box of area 1 around $(5, 3, 2)$ is 0.5%. One step more precise, the probability to be in box of area 0.1 around $(5, 3, 2)$ is 0.05%, or 0.005% to be in an even smaller box, and so forth.

If we repeat the same computation with $y = (4, 3, 2)$, that is, exactly at the mean, we get a density of 0.063, the highest density we can get.

Sometimes, the normal distribution is called *multivariate* or *univariate*, depending on whether the underlying vector space is multi- or one-dimensional. In the univariate normal distribution, the x and the x^T are usually combined in a single x^2, which is possible since the order of multiplication is not important in the real numbers. To allow a recognition effect if you read the equation in the literature, let's look at it using μ for the mean (this time a single number) and σ for the standard deviation:

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

It's not worth learning this equation if you know Equation 2.6.1, since it is the same equation in one dimension.

2.6.4 Logarithms for Densities

The density of many distributions, especially multivariate normal distributions, can sometimes be very small. This could involve numbers of the form 2.610^{-15}, which is difficult to read. Also, the normal distribution has a fraction, a multiplication, and an exponential, all three operations that we don't have a very good intuition for. For this reason, very often the logarithm of the density is reported instead of the density itself. Observe that if the density of x is larger than the density of y, the same holds for the logarithms,

in equations: $f(x) > f(y) \Rightarrow \log(f(x)) > \log(f(y))$. The logarithm is said to be *monotonous*. The logarithm in general converts products into sums, because $\log(a \cdot b) = \log(a) + \log(b)$, and exponentials into products, because $\log(a^b) = b \log(a)$. The density equation in particular becomes much simpler if we take its logarithm and multiply it by negative two:

$$f(x) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{2.6.2}$$

$$-2\log(f(x)) = \sqrt{2\pi|\Sigma|} + (x-\mu)^T \Sigma^{-1}(x-\mu) \tag{2.6.3}$$

You may want to take a minute to confirm that the logarithm has been taken correctly. Due to lack of imagination if it comes to names, mathematicians call this the *negative two log likelihood* of the normal distribution. However, it is difficult to deny that it reads better than the original density equation. As we will see later, there are even more advantages of using the log likelihood instead of the density itself.

Which basis the logarithm has is of less importance, because $\log_a$ and $\log_b$ only differ by the constant $\frac{a}{b}$, or in other words,

$$\log_a(x) = \frac{a}{b} \log_b(x)$$

To eliminate the exponential e^x in the normal equation, we used the logarithm to the basis e, also called the *natural logarithm*, above.

Chapter 3

Information Theory

Even though obviously at the center of science, *information* is typically used in a vague way. In this chapter, we will introduce information as a measurable entity. Very roughly, something has as much information as we need *bits* to store it on a computer in the shortest possible format. 2 minutes of a song, for example, can be packed to approximately 2 MB on the drive; a single, unchanged tone of 2 minutes length will be an MP3 much smaller (you can try this).

For a random variable, we are interested on how much information is lacking, that is, how many bits would it cost me to describe the output of the random variable. This "lack of information" is called the *entropy* of the random variable. In psychology, all variables of humans are random variable, and our job is to minimize the entropy. We can (and psychologists have for long) do this job without even knowing what entropy is, but it often comes in handy to know whether we are progressing, and how fast.

3.1 Information and Codes

Terms: Bits, Codes, Information
Skills: Inventing Codes and computing their bit costs
Understanding: The amount of space we need to store 'something' on a hard drive in our computer is its information.

3.1.1 Units of Information

The amount of information of a text that we can write down can be described by the length, that is, the number of alphabet signs it needs to write the text down. To avoid the problem of different alphabets, languages or writing styles, mathematicians agreed to define a smallest information unit called a *bit*, which can take one out of two values. The information contained in a text (where *text* refers very generally to everything that can be written down) is then the smallest number of bits needed to write down this text.

Definition 3.1.1. An information unit that can take on two values is called a

bit. The smallest possible amounts of bit needed to store a text is its *information*.

In this way, bits become the units of information. Note that the definition is based on the 'smallest' write-up of the text, but no one tells us how we actually can get at this length. In fact, information is a latent term (like the mean of a random variable) and can usually not be computed. However, there are programs that write texts in as few bits as possible (called *compression* programs) and do an excellent job at it; the length they achieve is a good approximation of the information. Also, observe that the text 'uacbzaec'[1] has more information that 'aaaaaaaa', even though both texts have 8 letters. Also, 'abcdefgh' has less information (hardly more in fact than 'aaaaaaaa'); we can figure this comparison out without actually having to compute the information.

3.1.2 Codes

An English text can be seen as a sequence of random letters. [2] You can think of each letter in an English text as a random variable. As letters come in different frequencies, we can create a code for 'English' that has more information density than the text you are reading at the moment. To see this, let's restrict ourselves, just to make the example easier, to a text that only contains the letters 'e', 't', 'j', and 'q'. In English texts, these letters make roughly 12.7, 9.1, 0.2 and 0.1 percent of the letters in an English text. We could use two bits to code each of these letters, for example like this:

letter	code
'e'	11
't'	10
'j'	01
'q'	00

In this code, the word 'tee'[3] would be '101111', so it would use 6 bits. In fact, every word would use two bits per letter. However, observe that we could also use the following codes

letter	code
'e'	1
't'	01
'j'	001
'q'	000

In this code, the word 'tee' is '0111', so only four bits.

It's one thought step more to compute how many bits are used per letter on average for this code. If we assume the four letters appear in the same ratio

[1] obtained by randomly typing on my keyboard

[2] Even though I hope this is not the impression you have from this text

[3] Which is the little plastic golfers use to shoot the ball from

as in English texts, their probabilities would be 0.575 for 'e', 0.410 for 't', 0.01 for 'j' and 0.005 for 'q'. Observe that the numbers of bits used per letter is a random variable which is '1' with probability 0.575 (the 'e'), '2' with probability 0.410 (the 't'), and '3' for 0.015 (the 'j' and 'q'). We compute the expectation of this random variable

$$\mathbb{E}\left(\text{bits per letter}\right) = 0.575 \cdot 1 + 0.410 \cdot 2 + 0.01 \cdot 3 + 0.005 \cdot 3 = 1.44$$

which shows that this code is more clever than the naive one, where we used 2 bits on average per letter.

Observe that no code is the beginning of another code. This property allows us to write 'tee' as '0111' instead of '01 1 1', which would use more information (because we would have to assign a code to the whitespaces). A code that satisfies this condition is called *prefix-free*. To put all this in one formal definition,

Definition 3.1.2. A *prefix-free* code for a set Ω is a function c from $\Omega \to 0,1^k$ such that for all different $\omega_1, \omega_2 \in \Omega$,

$$c(\omega_1) \not\preceq c(\omega_2)$$

The '$\preceq$' symbol means that the code for ω_1 is the beginning (or in math talk, the *prefix*) of the code for ω_2. For example, 001 can not be a code for an element if 0011 is a code for a different element. You know prefix-free codes from telephone numbers: No person can have the phone number 911-5478, because '911' itself is already a complete phone number. Also, no one will have the local phone number 434-7618, since '434' is the area code for Charlottesville, Virginia.

3.2 Entropy

Terms: The Coding Theorem, Entropy, Mutual Entropy
Skills: Computing the entropy of one random variable or the mutual entropy of two random variables.
Understanding: Entropy is the degree of the uncertainty we have about a random variable.

3.2.1 The Coding Theorem

Finding better and better codes can be for a random variable on a finite universe Ω can be fun, but the following theorem destroys a lot of the game:

Theorem 3.2.1. *In an optimal code, the number of bits used to code an elementary event of probability p is*

$$-\log(p)$$

Strictly speaking, you need an appropriate rounding since $-\log(p)$ is usually not an integer number, but that's a detail not very important to us. Another detail usually not important is what the actual code is. If we can compute how long the code will be, we are fine.

The expected length of the optimal code is the expected value of $-\log(p)$ for all possible outcomes. For finite Ω, this is given by

$$H(X) = -\sum_{\omega \in \Omega} P(X = \omega) log(P(X = \omega))$$

This amount tells us how many bits we need on average to code the outcome of the random variable. You can think of it as the amount of information we are still lacking about the random variable. This value is also called the *entropy*; more generally, it is also defined on continuous variables, where again the probability is replaced by the density:

Definition 3.2.2. For a random variable X on a vector space with density f, then value

$$H(X) = -\int_{\omega \in \Omega} f(\omega) log(f(\omega)) d\omega$$

is called the *entropy* or *Shannon entropy* of the random variable.

Observe that for a finite random variable, this is the sum as given above. For continuous variables, the entropy approximately describe the costs of the code if we would bin the outcomes in bins of area one. For example, if our random variable is a continuous number from -10 to 10, the entropy describes the expected number of bits we need if we round the outcome to the next integer.

If we halve the intervals we are looking at (so round to -10, -9.5, ..., 9.5 and 10), the probability for every slot roughly halves, too. Since the best code is the negative log of the probability for each event, this means that the best code increases roughly by one unit on average. The smaller we make our bins, the more precise this approximation becomes. We can phrase this formally as a theorem:

Theorem 3.2.3. *Let X be a random variable on a vector space with entropy $H(X)$. If we collect the events into bins of area ϵ, that is, round all results with precision ϵ, then the number of bits we need to represent the rounded outcomes of X is*

$$\text{costs} \approx H(X) + \log(\epsilon)$$

The approximation approaches equality as ϵ approaches zero.

So let's repeat this: On discrete variables, the entropy is the expected number of bits we need to represent the outcomes of a random variable. On continuous variables, it is the expected number of bits needed when rounding to a unit area.

Since the entropy is the number of bits we need to code the outcome, it describes our *uncertainty* about the random variable. In this way, it is comparable

to the variance of a random variable (remember: That's the expected squared distance to the mean). In fact, for a normal distribution, the two are proportional - the further we expect to be away from the mean, the more bits we need to describe the outcome to the same precision. However, assume we measure the height in a group that consists of equally many males and females. Height is normally distributed in both groups, but with a mean of 6ft for the man and 5ft for the women. Such a distribution is called a *mixture of Guassians*. Observe that for coding the outcome, we don't mind if the distance between is 1ft or 2ft (or any other distance); you may want to stop here for a second to confirm that. However, the variance of the height increases with the distance of the two modes.

3.2.2 Mutual Entropy

Like the entropy is related to the variance, the *mutual entropy* is related to the degree that two variables go together. Again, the mutual information describes the commonality of two random variables by bits: It is the expected number of bits that you save on coding the outcome of one variable if you know the outcome of the other. Let X and Y be these two variables; the number of bits that X is cheaper if we know that Y is a specific outcome y is given by

$$\begin{aligned} H(X) - H(X|Y=y) &= \int_x (-f(x|y)\log(f(x)) + f(x|y)log(f(x|y)))\,dx \\ &= -\int_x \frac{f(x,y)}{f(y)} \log\left(\frac{f(x,y)}{f(x)f(y)}\right) dy \end{aligned}$$

where $f(x,y)$ is short for $f(X = x \cap Y = y)$. If we want to know the average gain for all possible values of y, we have to compute the expected value of this, which is

$$\int_y f(y)\,(H(X) - H(X|Y=y))\,dy = \int_{x,y} f(x,y)\log\left(\frac{f(x,y)}{f(x)f(y)}\right) dxdy$$

This is the definition of the mutual entropy:

Definition 3.2.4. For two random variables X and Y over the same vector space Ω, the value

$$I(X;Y) = \int_{x,y} f(x,y)\log\left(\frac{f(x,y)}{f(x)f(y)}\right) dxdy$$

is the *mutual entropy* of X and Y.

We can see from the definition that mutual information is symmetrical regarding X and Y.

Recall that the mutual entropy is the expected numbers of bits that we can save on a code for X if we know Y (or vice versa). So if X and Y are independent, the value should be zero; in fact, if X and Y are independent, so

$f(x, y) = f(x)f(y)$, and the logarithm is zero. If both variables are the same, then $f(x, y) = f(x)$, so

$$I(X; X) = \int_x f(x) \log\left(\frac{1}{p(x)}\right) = -\int f(x) \log(f(x)) = H(X)$$

Again, this is similar to the covariance, for which we know that $cov(X; X) = var(X)$.

Chapter 4

Principal Component Analysis

This short chapter introduces Principal Component Analysis (PCA), which is at the core of exploring a data set which we know is normal, but don't know much else about. Even though a fairly simple idea, PCA has been the tool used for what probably constitute the most scientifically surprising findings in psychology, as for example the IQ and the 'Big Five' of the personality structure. PCA is the first time in this script that linear algebra and probability theory join forces to create something new. In a nutshell, PCA makes a basis change on the universe of a multivariate normal so that the entries in the new basis system are uncorrelated, and sorted by the variance they contribute to the total variance. For many applications, it is reasonable to assume that components with more variance are also substantially more important, while the others are less important and can be safely ignored.

4.1 PCA

Terms: PCA, Explorative Factor Analysis
Skills: Computing a PCA from a covariance matrix
Understanding: We can cut away a minor proportion of the data while strongly reducing the effort we have to write down this data.

4.1.1 The PCA idea

We talked about using an orthogonal matrix Q to change the basis of a vector space. This is useful, so we also want to apply this on the universe of a random variable X. If we know the moments of X, we can use our previous results to easily compute the first two moments of the transformed random variable QX:

$$\begin{aligned} \mathbb{E}(QX) &= Q\mathbb{E}(X) \\ \mathbb{V}(QX) &= Q\mathbb{E}(X)Q^T \end{aligned}$$

The first line is not surprising, the mean just gets changed along with the values. The second line also is not surprising if we remember that the variance is the expected *squared* distance to mean; both vectors in the 'squared' get transformed by Q, so Q appears on both sides.

Since we can freely pick any Q that suits us, there must be some potential here that we can exploit. Assume for example that X is two-dimensional normal with covariance matrix

$$\mathbb{V}(X) = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

Then a plot of the distribution looks like the usual bell-shaped curve of the Gaussian distribution, but stretched in the main diagonal of the two-dimensional plane. We can pick any rotation of this data set without changing the content (it's just a basis change, comparable to thinking in cents instead of dollars, or vice versa). Stop for a moment and look at the shape; which rotation might make our lives easier?

If you followed the PCA logic, you would have picked a rotation that rotates the longest axis of the normal distribution on one basis, so that the Gaussian curve now is stretched in horizontal direction. This basis has two advantage: Firstly, the scores are now uncorrelated[1]. Secondly, because the variance in the horizontal axis is much stronger than in the vertical axis, if we accidentally happen to forget the vertical axis, not too much harm is done. Why should we do this? Because of Ockam's razor, a shorter description is a better description, and here we could describe our data by one number instead of two.

The PCA is exactly this rotation, potentially on more than two dimensions:

Definition 4.1.1. Let X be a random variable on a vector space. Transforming the data set by an orthogonal matrix Q such that the covariance matrix of QX is diagonal with decreasing values on the diagonal is called a *Principal Component Analysis* (PCA). Q is called the PCA transformation.

Let's postpone the question how we compute Q for a moment. What do we gain from performing a PCA? An assumption that surprisingly often turns out correct is that multiple variables are linearly dependent on much fewer variables that have not been measured directly. These span a lower-dimensional space in the variables that has a lot of variance, while the remaining components show not much variance. For instance, assume three cognitive tests are measured which are (differently strong) dependent on cognitive skill and memory perfomance, which for this example we will assume to be independent. Imagine every participant as a point in 3D space describing his three scores; then, the points will all be close to an ellipse-shaped disc, where the long axis of the disc could be the variance caused by cognitive abilities and the smaller axis the variance caused by memory ability. If we do a PCA on this data set, the disc wil be rotated such that it lies on the ground surface (which means we can easily determine that the third direction is mostly unused), and the long axis (cognition) on one score and the short axis (memory) on the other score. In this idealized case, we would find out that our three scores are essentially 2-dimensional, and would have separated the two causes.

If you select a large number of cognitive tests and perform a PCA on these, you will find that the highest variance is much larger than would be expected

[1]Sometimes, the term *orthogonal* is used for uncorrelated here, because the expected outer product is zero

for random numbers. The reason is the high correlation of cognitive abilities. This observation coined the definition of intelligence as this major direction.

If you select a large number of personality trait question, you will find that the first five variances together are a good deal larger than would be expected for random numbers. These personality questions could be something like 'Do you feel comfortable around people', 'are you exacting in your work', or similar. Since this finding has been replicated multiple times [2] even though hundreds of different initial questions have been asked, it is assumed in personality psychology today that personality is a fairly low-dimensional space.

4.1.2 The PCA algorithm

The challenge that remains is to find the matrix Q. If $\mathbb{V}(X)$ is the covariance matrix of X, we want that the transformed covariance matrix $Q\mathbb{V}(X)Q^T$ is a diagonal matrix D:

$$Q\mathbb{V}(X)Q^T = D$$

So we want that for each basis vector, multiplication with the covariance matrix equals multiplication by a single number. In math terms, if e_i is one of these basis vectors, we want

$$\mathbb{V}(X)e_i = \lambda e_i$$

for a number λ. You may remember that such vectors are called *eigenvectors* of $\mathbb{V}(X)$; so the new basis we are searching is the basis of eigenvectors. Also remember that the inverse basis change matrix, Q^T, translates a vector that is one at one entry and zero at all others into a new basis vector represented by the old basis, which in our case means, one of the eigenvectors. In other words, Q^T has all the eigenvectors of $\mathbb{V}(X)$ in the columns, and therefore Q has all Eigenvectors in the rows. As a side remark, the basis transformation Q that transforms a matrix into a diagonal matrix is called a *diagonalization*.

To make sure that Q is orthogonal, all its rows are normalized to have a norm of one (as we saw earlier, stretching an Eigenvector doesn't change the fact that it is an eigenvector). Observe further that $\mathbb{V}(X)Q$ multiplies the covariance matrix to its Eigenvectors, so that all these vectors are strechted by their corresponding Eigenvalues, but otherwise not changed. Multiplying Q^T from the left than gives the diagonal matrix with the Eigenvalues on the diagonals, since the norm of the vectors themselves is one. In this way, we see that the variances of the scores in the new basis systems are the Eigenvalues of $\mathbb{V}(X)$.

The new basis system are sometimes called *factors*, and the entries of the transformed participants on these factors *loadings*. The PCA procedure, possibly with subsequent elimination of factors with very low Eigenvalues, is also called *explorative factor analyses*.

Let's look at our simple example above of two variables X with covariance matrix

$$\mathbb{V}(X) = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$$

[2] Although some discussion is still open whether the magic number is three or five

The two Eigenvectors of this matrix are $(\sqrt{0.5}, \sqrt{0.5})^T$ and $(\sqrt{0.5}, -\sqrt{0.5}$; you may want to check that these are in fact Eigenvectors, and that they are normed (that means, their norm is one). The two Eigenvalues are 1.8 for the first and 0.2 for the second Eigenvector. The corresponding Q matrix is hence

$$Q = \begin{pmatrix} \sqrt{0.5} & \sqrt{0.5} \\ \sqrt{0.5} & -\sqrt{0.5} \end{pmatrix}$$

You may want to check that the transformed covariance matrix is in fact

$$\mathbb{V}(QX) = Q\mathbb{V}(X)Q^T = \begin{pmatrix} \sqrt{0.5} & \sqrt{0.5} \\ \sqrt{0.5} & -\sqrt{0.5} \end{pmatrix} \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{0.5} & \sqrt{0.5} \\ \sqrt{0.5} & -\sqrt{0.5} \end{pmatrix} = \begin{pmatrix} 1.8 & 0 \\ 0 & 0.2 \end{pmatrix}$$

Observe that for a real data set, we of course don't have the covariance matrix, but we do have the estimated covariance matrix, which we may use instead.

The computation of the Eigenvectors is something which we safely can leave to good computer programs.

Chapter 5

Testing

The re-entrance point from the space of math to the atmosphere of real world are *tests*, which translate a number we got from a computation in a binary decision. The decision of the test is done by comparing a tested value against a critical value, with one decision resulting if it is higher and the other if it is lower than the critical value. For some scientific purposes, it may be questionable whether we want the result of a test, considering that it throws away all information about the decision value other than whether it is above the critical value. However, for real life, we often have no choice than to make a test in the end: We will never know whether a drug certainly helps a patient, but at some point we have to give him the drug or not give him the drug. In psychology, we usually have tests for an hypothesis, and our decision is whether we believe the hypothesis or not. Preferably, the test is based on the probability that the hypothesis is true, but that is not always possible. Then, we have to make do with what we have got.

5.1 Probability Based Hypothesis Test

Terms: Tests, Critical Value, Hypothesis Tests, Probability Based Hypothesis Tests, Type I and Type II error
Skills: Deciding where to put a critical value
Understanding: The fact that we have to make decision forces us to set critical values for tests. No decision is also a decision.

Assume a game where you bet one dollar and have a 50% chance of earning \$1.50 by picking one out of two nut shells. It is an easy call that this game is probably not worth playing, because in expectation we loose 25 cent at every game. However, assume the chances are 80% of winning; then we probably would play the game, because the expected outcome is

$$\mathbb{E}(gain) = 0.8 \cdot \$0.5 + 0.2 \cdot (-\$1) = \$0.2$$

In fact, we would choose to play the game if our chance for winning is above $\frac{2}{3}$. This decision based on comparing a value (the probability) to a critical value (the probability from which we expect to win) is called a *test*.

Definition 5.1.1. A function that maps a test value to a binary decision based on a critical value is called a *test*.

If a positive decision means to accept an event as true, this event is called an *hypothesis*, and the test is called an *hypothesis test*.

An hypothesis test based on the probability for the hypothesis is called a *probability based hypothesis test*.

Observe that a positive hypothesis test does not mean that the event took place, i.e., that the hypothesis is true. It just means we accept it as true for the purpose at hand. In our example, playing the shell game doesn't mean that we win, it just means that we treat reality as if we would win. The choice of the critical value reflects this unclean translation. In the shell game with one dollar, we choose the critical value so that the expectation is positive. However, if we would play for a million dollar (and gain 1.5 million dollar in case of a win), we would probably still not play because loosing a million dollar is much more severe for our lives than could be countered by a chance of gaining 1.5 million dollar, even if the game is 'fair'. The same holds for the choice of critical values in psychological settings; if we are talking about an hypothesis without much effect to anyones daily live, we can be less careful when setting a critical value. When we are talking about a medication with potentially severe side effects, our critical value needs to be much more conservative.

Since the test outcome for an hypothesis tested is conditional on the true outcome of the hypothesis, there are two errors the test can make: It can call for the hypothesis even though the hypothesis is not true, and it can call against the hypothesis even though the hypothesis is true. These two errors are called *Type I* and *Type II* errors:

Definition 5.1.2. Let T be the probability that event that a test accepts an hypothesis H. Then,

$$P(T|\overline{H})$$

is called the *Type I error* of the test, and

$$P(\overline{T}|H)$$

is called the *Type II error* of the test. The complement of the Type II error, $P(T|H) = 1 - P(\overline{T}|H)$, is called the *power* of the test. For a specific event in the hypothesis, $h \subseteq H$ the probability $P(T|h)$ is called the power for the event h.

If you start to confuse Type I and Type II, observe that you only have to replace the hypothesis by its complement to exchange the role of the two, so which is which is not really important. However, it is very important that both errors exist, since one can not be computed from the other. We sometimes are very concerned about type one errors, for example if we hand out a medication at a successful test that has strong side effects. Sometimes, we are more concerned about the Type II error, for example on an fire warning system were a false alarm is much less of a problem than a missed fire. Sometimes, we are more concerned about the relation of the two types of error rather than their

extend; in a juristical decision, for example, we prefer a correct decision over a wrong one, but more than that we want a fair decision that does not favor any of the two sides.

5.2 Significance Test and Bayesian Testing

Terms: Bayesian Tests, Significance Tests, p-Value, a-priori and a-posteriori probability, nil tests, null results
Skills: Computing the a-posteriori, conducting significance tests, avoiding common errors
Understanding: Significance test, despite widely used, a fundamentally flawed. They should only be used if bounds can be set for power and a-priori ratio, which mostly makes them Bayesian tests.

5.2.1 Bayesian Testing

If both errors of a test and the a-priori probabilities for the hypothesis is known, then Bayes' Theorem can be used to compute the probability of the hypothesis after the test. Let T be the event that a test was positive, H an hypothesis and $H_0 = \overline{H}$ its complement, sometimes called the *null hypothesis*. We are interested in

$$P(H_0|T)$$

which is the probability that the null hyposthesis is true even though the test was positive in favor of H. By Bayes rule,

$$P(H_0|T) = P(T|H_0)\frac{P(H_0)}{P(T)} = P(T|H_0)\frac{P(H_0)}{P(T|H_0)P(H_0) + P(T|H)P(H)} \tag{5.2.1}$$

In many situations, the critical value of the test can be chosen such that this a-posteriori probability has a specified value, so that the test becomes a probability based hypothesis test, which is the 'gold standard' we usually want to obtain. This method of testing is usually referred to as *Bayesian Testing*.

Equation 5.2.1 is sometimes re-written as

$$P(H_0|T) = \frac{1}{1 + \frac{P(T|H)P(H)}{P(T|H_0)P(H_0)}}$$

The inverse of the fraction in the denominator is called the *Bayes Coefficient*. In fact, the Bayes coefficient is also an upper bound for the a-posteriori probability, which we can see if we bound the expression $P(T|H_0)P(H_0)$ in Equation 5.2.1 to zero:

$$\begin{aligned} P(H_0|T) &= \frac{P(T|H_0)P(H_0)}{P(T|H_0)P(H_0) + P(T|H)P(H)} && (5.2.2)\\ &\leq \frac{P(T|H_0)P(H_0)}{P(T|H)P(H)} && (5.2.3)\\ &= P(T|H_0) \cdot \frac{1}{P(T|H)} \cdot \frac{P(H_0)}{P(H)} && (5.2.4) \end{aligned}$$

5.2.2 Definition of Significance Tests

In psychology, the most frequent hypothesis test used is, sadly, not a probability based hypothesis test, but a class of tests called *significance tests*:

Definition 5.2.1. An hypothesis test based on the probability of a empirical finding given the complementary event of an hypothesis is called a *significance test* for the hypothesis, or against the complementary event. The complementary event is sometimes called the *null hypothesis*.

The critical value is called the *significance level* of the test, or α level. An hypothesis decided for in a significance test is called *significantly* accepted, or the null hypothesis significantly rejected. The probability of the empirical finding given the null hypothesis (the test value) is sometimes called the p-value.

The logic behind a significance test is that if our finding is very unlikely under the null hypothesis, then the null hypothesis must be wrong. This logic is based on the Aristotelian way of thinking (which is completely correct for non-stochastic statements) that if A is impossible, then the opposite of A must be true. However, if a finding assuming A is unlikely, the opposite of A is not even necessarily likely, let alone true. Significance tests by definition control for Type I error (the α level is an upper bound for the Type I error), but they can have any Type II error. So in a nutshell, significance tests are not working.

5.2.3 Justified Applications of Significance Testing

To value the outcome of a significance test, it is worthwhile to go back to Equation 5.2.4, which said that if T is an event that a significance test was positive, H an hypothesis and $H_0 = \overline{H}$ its complement, the null hypothesis, then

$$P(H_0|T) \leq P(T|H_0) \cdot \frac{1}{P(T|H)} \cdot \frac{P(H_0)}{P(H)}$$

The first factor is the p value, the second is the inverse of the power, and the third factor is the a-priori ratio. If the product of all three is low, then this means that H_0 is unlikely; p alone makes no statement. This term is called the *Bayes' Factor*; in fact, we saw that $P(H_0|T) = \frac{1}{1+\frac{1}{b}}$ where b is the Bayes factor. The only objective of using the bound is to get an intuitive grasp of the effect a p value has on our believe in the null hypothesis.

It is in fact often possible to give reasonable bounds of the other two factors; in this case, a significance test is valid to learn something about the hypothesis. For example,

(1) Assume we know that roughly $P(H) = P(H_0)$, that is, a-priori we have no idea which of the two hypothesis is more likely. This could for example be the case if H is that one out of two groups is stronger, and H_0 that the other is stronger. Also, let's say we know that the power $P(T|H) \geq 50\%$, which is often a reasonable assumption. Then, $P(H_0|T) \leq 2p$, so twice the p-value. If p, for example, is just below 0.05, then this means that the null hypothesis has probability below 10%, which in many cases makes it acceptable to act as if H was true.

(2) Assume the same statement about power, but this time $P(H)$ is less likely than $P(H_0)$; for example, the null hypothesis assumes a typically found level of a neurotransmitter, while H assumes a much stronger level. Let's say we can at best agree that $\frac{P(H_0)}{P(H)} \leq 5$. In this case, the bound of the null hypothesis is $P(H_0|T) \leq 10p$. If p was 5% again, then this bound would only allow a 50:50 guess between H_0 and H (in fact, if we use $P(H_0|T) = \frac{1}{1+2}$, we see that $P(H_0|T) = \frac{1}{3}$).

(3) Assume the same statement, but with an a-priori ratio of $\frac{P(H_0}{P(H)} \leq 20$. Even a p value of 5% leaves us with no statement about the hypothesis by the bound (since this only tells us that the probability is less than 200), and the exact computation shows that $P(H_0|T) = \frac{2}{3}$, so *more* likely than the alternative hypothesis H, despite the fact that our result is significant at an α level of 5%.

5.2.4 Horrible Failures of Significance Tests

The third situation just given is not artificial. Cohen reduces the argument to absurdity by suggesting to test whether a person is American by asking whether he is in the US congress, where a person in the congress is 'tested' to be no American. In fact, under the null hypothesis that a person is American, his chances of being in congress is only $p = \frac{535}{312{,}000{,}000}$, so the test is a fully valid significance test. In other words, all congress-men are significantly non-American. The flaw is obviously that under the alternative hypothesis that a person is not an American, his chances to be in congress is zero, that is, the power of the test is zero. However, following usual practice in social sciences at the moment, editors of journals would have to accept Cohen's result as methodologically fully correct.

Sadly enough, significance testing is actively used in law. Assume that a blood sample has been found at a crime scene with blood group B^-, and a suspect has this blood group. Under the null hypothesis that the suspect is innocent, his chances of sharing the blood group with the sample is only roughly 2% (which is the frequency of B^- in the general population). Therefore, the suspect is significantly guilty. Not only that this is a true story reported by Schrage, the court made it worse by reasoning that the found evidence gives a 98% chance of the suspect being guilty.

Significance tests don't satisfy any requirements we typically have for our indicators. For example, significance tests are not anti-symmetrical, which means that if a result significantly indicates an hypothesis H_1, it can at the same time significantly indicate the complement H_2, even though H_1 and H_2 contradict each other. To see this, assume we are in front of a stadium where a red-jersey sports team plays a yellow-jersey team. If red wins, the chances that a fan with red clothes leaves the stadium first is low (because typically the supporters of the loosing team leave the stadium first to avoid the traffic). Let's for the sake of the argument assume the probability is less than 5%. The same analogously holds for the yellow team: So a non-red clothed person indicates a win for red, a non-yellow a win for yellow. If the person who leaves first happens to have neither of the fan colors (they, a police officer in a blue uniform), this is a

significant indication that red won and yellow won. No joke, this is truly how significance tests work.

5.2.5 Nil Tests

Another problem with significance test, in addition to the fact that they don't work without at least bounds on power and the a-priori ratio, is that they are often used wrong (remember, even if used correctly, significance tests are worthless without power and a-priori estimates). The problem is that a significance test is sometimes made against an hypothesis which a-priori already has a probability of zero. For example, say the average performance of a large population on a test is zero. For a new population (e.g., from another country), one might set the null hypothesis that this group has the same performance. Since the distribution of the new group is assumably a continuous distribution, the probability for every single elementary event is zero, so $P(H_0|T) = 0$. Even though every test of course can be interpreted as 'successful' then, there is no scientific statement done; in fact, the difference can be arbitrarily close to zero, and still the test will be significant with sufficient number of participants.

Observe, however, that the a-priori distribution is not necessarily continuous. It could for example be that the second group, with some non-zero probability, is really drawn from the same population as the first group (maybe by a possible mistake, or because the groups have been chosen on a variable that is independent from the cognitive test with some probability). In this case, the significance test makes sense as long as this a-priori probability can roughly be estimated or bounded.

5.2.6 Null Result Interpretation

A second real misstake for significance tests is that failure to find a significant result is interpreted as a substantiation of the null hypothesis. Even if the p value, $P(H_0|T)$, is very large, that does not mean that $P(H_0|\overline{T})$ is very low (which would be a significant indication for the H_0). Sometimes, it is reasoned that with large power, failing to find a significant result could be seen as indication for the null hypothesis. Again, this is a completely wrong conclusion since the third factor int Bayes coefficient, the a-priori ratio, could still be any value. Only a combination of low a-priori ratio (or at least, bounded a-priori ratio), a sufficiently high power and a sufficiently high p value can be interpreted in favor of the null hypothesis, but even then it is not sufficient to be larger than a typically used α value of 5%; relevant would be a sufficiently high value to get a high Bayes ratio.

This bad practice has been so strongly accepted that some even suggest to use significant tests to test assumptions in more complex tests. For example, assume the computation of $P(T|H_0)$, the p-value, requires that a random variable X used in the test (e.g., the test value) is normally distributed. Now, some suggest to first perform a second test against the null hypothesis that X is normally distributed (for which a number of significance tests exist), and to proceed if the test was not significant.

Such a test can be a first step to *exclude* that the assumptions are upheld (again, if reasonable power and a-priori values are available), but never to accept normality of X. In fact, there exist no tests for this, so the best proxy is to argue for normality based on descriptive outcomes, for example QQ-plots, and base all further investigations on the assumption that X is normality, being aware that all conclusions rests on this assumption.

5.2.7 How to Do Correct Significance Tests

Bayesian tests should always be used rather than significance tests if anyhow possible. However, in reality the a-priori ratio as well as the distribution of some test values under the alternative hypothesis are sometimes difficult to access, which means that unless we want to rely on argumentation of descriptive values alone[1], we have to go back to the next objective possibility. This is, so far, significance tests.

The first step for a correct significance test is to formulate both H and it's complement, the null hypothesis. Care needs to be taken that $P(H_0)$ is not zero by an unlucky choice of the H_0. For example, instead of testing against the null hypothesis that two groups are *identical* (which usually is a nil hypothesis), we should test against the null hypothesis that the difference is smaller than a certain value δ. If possible, the δ should be chosen such that the difference is meaningful for the application at hand, that is, that the information we gain from considering the groups different outweights the loss of model simplicity by differentiating the groups. The alternative hypothesis, H, in this case would be that the groups are further than δ away from each other. Sometimes it is also reasonable to phrase a *one-sided* hypothesis, for example that the first group is equal or stronger than the second group; in this case, H_0 would also probably have a non-zero probability as it includes all outcomes where the first group is weaker.

If something that seems to suggest itself as the null hypothesis is the outcome we want to confirm, we have to turn the test around to make sense. For example, if we want to show that a medication has no negative side-effect on reaction times, setting the null hypothesis to 'no differences' between the groups means to hunt for a null result. Instead, the null hypothesis should be 'there is an absolute difference greater than x between the two groups'. The choice of x must be made by substantial considerations, that is, x should be chosen such that a reaction time deficit of x by the medication has no substantial impact on the well-being of the patients. Of course, the null hypothesis can also be chosen to say that the medication *improves* reaction times; note, however, that this null hypothesis may be difficult to reject if in fact the medication has no effect on reaction time, or a slight but unimportant one.

The second step is to find the distribution of a test value under the null hypothesis, and to fix a critical value according to a suitable α level. By default,

[1] Which is by far not a negligible option. However, the danger is high that in this case, the louder speaker with more followers will be accepted rather than the more thorough researcher. Therefore, we sometimes have to compromise between the scientifically better approach in an ideal world, and an objective, even though less scientific, approach like significance tests.

this level should never be chosen to be 5% (just to avoid the danger of doing what everybody does). If the difference is a mere scientific observation, the α level can easily be set to 10% or even higher. If a Type I error, however, would have serious consequences (e.g., if a treatment is administered in case of a successful test that may have unwanted site effects, or the freedom of people is restricted, e.g., for a forced psychotherapy), the α level needs to be substantially lower.

The decision of the α level is of course also dependent on the other factors in the Bayes coefficient, which should be thought about carefully. The power can be difficult when the distribution of the test value under the alternative is not known, but in many cases reasonable estimates can be made. Simulations on some likely situations under the alternative hypothesis H can help to get an idea. Observe that a power of 50% and 90% doesn't make such a difference since the power occurs as it's inverse (between 50% power and 100% power, the Bayes coefficient only changes by a factor of two). 50% power is something that is usually easy to argue for. For the a-priori ratio, you can make an *agreement* with the reader, where you accept that your conclusions are only correct if the a-priori ratio is above a certain value which is acceptable for many. Often, the alternative hypothesis H can be reasoned for, while H_0 is something that would not be as expected from the theoretical knowledge. For example, maybe your H is that a treatment that is based on a lot of theoretical thinking will have an effect greater than zero. In such situations, an a-priori ratio of 1 is probably an acceptable assumption. If you are hypothesizing something which in fact is less expected, than you have to accept a larger a-priori ratio, but this is justified by scientific principles: Strong statements need strong substantiation. A p value of 5% for an hypothesis that surprises the world is just not enough.

Finally, compute the p value for your data set under the specified null hypothesis. If the null hypothesis is a range of numbers, it may be impossible to compute the p value exactly. However, it is usually easy to identify a most conservative point. In our previous example where the null hypothesis was that one group is equal or weaker than another, the point of no differences is the most conservative point. If you assume this difference and compute the likelihood of the test value if the group difference is zero, you are on the conservative side as the true p value will be bounded by this value.

Then, check the p value against your pre-set α level and compute an upper bound for the a-posteriori probability of the H_0, $P(H_0|T)$. Report this value. If you are testing multiple hypothesis, you may want to fix a most conservative choice of an a-priori ratio and power for all hypotheses so that you don't have to outline the text for the computation repeatedly. Also, if possible, it is helpful to report *effect sizes*. These can be a variety of statistics computed from your data; for example, for the two groups, you should report the descriptive estimated distance between the two groups. If possible, it is helpful to report a standardized effect size, where typically the effect size is standardized by an estimate of its own standard deviation. We will later introduce a number of values that could be used for effect size reports in different situations.

Chapter 6

Normal Modeling

Models are distributions of data, usually (although not necessarily) fairly complex distributions. We usually try to choose the models such that they satisfy two conditions: Firstly, they should be simple enough that we can work with them. Secondly, they should be as close to the distribution of the real world as possible.

Sadly, these two aims are usually contradicting each other, since the world is fairly complex. In the extreme case, we can say that every variable is zero; that is as simple as we can get for a model, but also as incorrect. In the other extreme, we can use the universe as a model of itself; that is definitely very accurate and even comes with an automatic computation of the outcome, but the computation always takes as long as the process itself - since both are identical. Luckily, we usually find ourselves in the situation that even strong simplifications don't tear us far away from a reasonably good descriptions. For example, assuming a normal distribution makes our models fairly simple (since, e.g., the normal distribution is already fully described by its first two moments, and sums of normal variables remain normal), while at the same time many variables in social science follow a good approximation of the normal distribution because of the central limit theorem (note, however: Almost none follow exactly a normal distribution). Models based on normal distribution are most generally referred to as *normal models*, which we will introduce in this chapter. Historically, these came under multiple different names like regression models, ANOVA-type models, general linear models, or Structural Equation Models. All of these are equivalent, and preferences are a matter of taste only. We will here use the Structural Equation Model (SEM) language.

6.1 Structural Equation Models

Terms: Structural Equation Models, latent variables, manifest variables, simple matrix notation, path diagrams, RAM notation
Skills: Creating and understanding models, translation of real-world situations into models, computing manifest distributions
Understanding: SEM (which is, all normal models) are a 'sweet spot' between simplicity and explanation power for many settings.

6.1.1 Definition

Models are distribution for some variables which are observable, called *manifest*. However, we sometimes want to give this distribution in terms of other variables which are not observed; these variables are called *latent* variables. While manifest variables must be measurable, latent variables can be constructs which are not accessible or even not existing in real world. For example, intelligence is a latent variable, while the score in a specific cognitive test is manifest. In Structural Equation Models, the manifest variables are defined by a normal distribution on some latent variables and a linear map from the latent variables to the manifest:

Definition 6.1.1. Let Y be a normally distributed vector. A model that predicts data X to be a linear map from Y,

$$X = \Lambda Y$$

is called a Structural Equation Model (SEM).

The matrix Λ is called the *Structure Matrix*. The entries of Y are called *latent* variables, and the entries of X *manifest* or *observed* variables. This notation of a SEM will be called simple matrix notation in the following.

We will later see alternative notations of SEMs.

As the manifest variables X are a linear transformation of Y, we can compute the mean and covariance matrix of X as

$$\begin{aligned} \mathbb{E}(X) &= \Lambda \mathbb{E}(Y) \\ \mathbb{V}(X) &= \Lambda \mathbb{V}(Y) \Lambda^T \end{aligned}$$

Often (but not necessarily), the expectation of the manifest variables are called $\mu = \mathbb{E}(X)$ and the covariance matrix $\Sigma = \mathbb{V}(X)$.

6.1.2 An Example SEM

We could make our lives easy by always setting $\Lambda = I$, the identity matrix, and directly set $\mathbb{V}(Y)$ to be the covariance matrix we want to model. However, SEM helps us to think structurally about our model in terms of underlying sources (typically few latent variables) that together generate the manifest variables.

For example, consider we assume two latent variables 'eye precision' and 'ski skills' of a group of participant. Both are latent, since we cannot measure these variables directly; however, we could measure participants' scores on Archery, their time in a Biathlon race, and their time in a downhill ski race. With some acceptable simplification, one could say that Archery is influenced by eye precision but not by ski skills, biathlon is influenced by both, and downhill by ski skills but not by eye precision (okay, that simplifies it quite a bit). The structure matrix of this SEM would be

$$\Lambda = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{pmatrix}$$

Assume for example that the covariance and mean of the two latent variables is

$$\mathbb{V}(Y) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \qquad \mathbb{E}(Y) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

which corresponds to saying that both latent abilities are uncorrelated, have a average squared value of 2, and a zero mean. In this case, we would predict the covariance matrix of the manifest variables as

$$\Sigma = \begin{pmatrix} 2 & 2 & 0 \\ 2 & 4 & 2 \\ 0 & 2 & 2 \end{pmatrix}$$

and the mean, of course, to be zero. If we would have modeled this Σ directly without using the latent distribution and the structure, we would have had a harder time, even though mathematically both are the same thing.

6.1.3 Path Diagrams

In order to allow even more conceptual thinking, we introduce a second, completely equivalent representation of SEM via path diagrams. In these diagrams, manifest variables are represented as squares, and latent variables as circles, in a graphical representation of the model. These objects are called *nodes* in diagram language. The nodes are connected by arrows which are called *edges*. Path diagram have two sets of edges: Double-headed arrows represent covariances between variables, and variances if they loop back to the same variable. Single-headed edges take on the role of the Λ in the simple matrix notation, they represent regressions from one variable to another.

In addition to being graphical, path diagrams are also more flexible. In the simple matrix notation, only the latent variables have covariances and variances. In path diagrams, double-headed edges are allowed between all variables, including the manifests. Also, the Λ in the simple matrix notation only describes a mapping from latent to manifest variables, but in path diagrams, single-headed edges can go from any variable to any other variable. This flexibility comes for a cost: It is by far not as easy to compute the manifest values X or their distribution as it is in the simple matrix notation.

Definition 6.1.2. A *Path Diagram* consists of a diagram with

Squares	Representing manifest variables,
Circles	Representing latent variables,
Triangles	Representing a constant variable,
Double-Headed Arrows	from one variable to another indicating a covariance, or a variance if the double-headed arrow points from a variable to itself,
Single-Headed arrows	from one variable to another indicating an additive contribution from the source to the target.

The objects representing variables are called *nodes*, the connecting arrows are called *edges* of the diagram. The single-headed arrows are also called *regression* edges. All edges have a numerical label that indicates the value of

the covariance or additive contribution, respectively. A series of single-headed arrows (of length zero to any length) is called a path.

An outcome X of the path diagram is modeled by first assigning a value to all nodes with respect to the covariance matrix generated by the double-headed edges and one for each constant node. Then, the mapping from all nodes to all nodes generated by the single-headed edges is applied to the current values and added to the initial values repeatedly. If this series converges, the values in the square nodes are then the model generated outcome X. Otherwise, the path diagram does not represent an SEM.

By convention, an edge that is *not* drawn in a path diagram is considered an edge with label zero. Also, the label 'one' is usually omitted from the edge, so an unlabeled edge is considered to have a label '1'. If there is no triangle, then of course all edges from the triangle are also not there, which means there is no constant contribution to the path diagram. In this case, all means will be zero.

The description of the model generation actually sounds more horrible than it is. To generate data, you do two steps:

First, you ignore all single headed arrows and collect all double-headed values in a big covariance matrix for all variables, which defines a normal distribution with zero mean; from this distribution, sample a set of values for all variables. Set the value for all triangles to be one.

Second, you ignore all double-headed arrows. Add the label times the value of the source to the target of each single-headed arrow. Usually, you are done; however, some models have pathes of multiple single-headed arrows stringed together. For those, multiply all edge labels on the path, and then treat the whole paths as if it was another single-headed arrow.

The distribution of the manifest variables is still difficult to read off from this description. This is the price we pay for an intuitive, graphical representation. There are two ways to do the computation. The first one are called *path rules* and work as follows:

Theorem 6.1.3. *To find the covariance of two variables A and B in a path diagram (or the variance if $A = B$), find all pathes from A to B which consist of a path of single-headed edges (of length zero to any length) against the direction the arrows a pointing, then exactly one double-headed arrow, and then a path (again of length zero to any length) of single-headed arrows in the direction the arrows are pointing. Multiply the edge labels for all edges in the path, and add up the product for all such pathes. The result is the covariance of A and B.*

To find the mean of a variable A, follow all pathes of single-headed edges from the triangle to A. Again, multiply the edge labels for each path, and add the results up for all pathes.

The game described in the theorem needs some practice and is fun afterwards. To play it, put your finger on A. Follow any number of single-headed edges backward, then follow a double-headed edge, and then again any number of single-headed edges forward until you reach B. Now you have found one path that creates covariance equal to the product of all edge labels you passed

on your way from A to B. Do this for all such pathes (remember that 'any number' of single-headed edges can include no single headed edges, that is, you may also start with a double-headed edge). If you add up all these covariance contributions, you are done. The game for the mean is even simpler, just start at the triangle and follow all pathes of single-headed arrows down to the target variable. Again multiply all labels you pass on a path, and add the results up if there are multiple such pathes.

The drawback of path rules is that we only get a single number in the covariance matrix or the mean vector. If we want all of them, there is an easier way by translating the path diagram into a matrix notation. Luckily, there is one that is very close to the graphical representation.

6.1.4 RAM notation

RAM stands for Reticular Action Model, but everyone refers to it just as RAM. In RAM, an SEM is represented by essentially two matrices which directly correspond to the double-headed edges and the single-headed edges in a path diagram. A filter matrix is needed to separate the manifest from the latent variables. Formally, an SEM in RAM is defined as follows:

Definition 6.1.4. For an SEM with n variables, let $S \in \mathbb{R}^{n\times n}$ be a positive definite symmetrical matrix and $A \in \mathbb{R}^{n\times n}$ a matrix for which all Eigenvalues are below one in absolute value. Let m be a vector. If n_M of the variables are manifest, let $F \in \mathbb{R}^{n_M\times n}$ be a matrix that filters out the manifest rows from a vector of all variables. Let Y be a n-dimensional normally distributed random variable with mean m and covariance matrix S, then the model predicted outcome is

$$X = F(I - A)^{-1}Y$$

Theorem 6.1.5. *The mean of the manifest variables in RAM notation can be computed as*

$$\mathbb{E}(X) = F(I - A)^{-1}m$$

and the covariance matrix as

$$\mathbb{V}(X) = F(I - A)^{-1}S(I - A)^{-T}F^T$$

Again, μ is usually used for $\mathbb{E}(X)$ and Σ for $\mathbb{V}(X)$.

RAM notation acts as something like a translator between the simple matrix notation and path diagrams. The following theorem explains how you can translate these two notations into RAM and vice versa:

Theorem 6.1.6. *For an SEM represented by a Path Diagram, let m be the labels of all edges from the triangle, A be the labels of all single-headed edges not coming from triangles, and S the labels of all double-headed edges. Let F be a matrix that has one row for each square in the diagram, and each row has only zeros with exception of the column that corresponds to the square. Then, the RAM notation using this m, A, S and F is the same SEM as given by the path diagram.*

For an SEM in RAM notation using m, A, S, and F, draw a path diagram with a square for each row in F and a circle for the remaining columns in F, and order the nodes with respect to F. Draw single-headed edges for each non-zero entry in A, labeled with that entry, and double-headed arrows for each non-zero entry in S analogously. Draw a triangle and an edge to each node for each non-zero entry in m, again with the entry in m as label. The resulting path diagram is the same SEM as given by the RAM notation.

For an SEM in RAM notation using m, A, S, and F, let $\Lambda = F(I - A)^{-1}$. The simple matrix notation SEM with Λ as structure matrix and m and S as distribution for the latent variables is the same SEM as given by the RAM notation.

For an SEM in simple matrix notation using Λ, m and C with n_M manifest and n_L latent variables, let A and S be the $n_L + n_M \times n_L + n_M$ matrices

$$A = \begin{pmatrix} 0 & 0 \\ \Lambda & 0 \end{pmatrix} \quad S = \begin{pmatrix} C & 0 \\ 0 & 0 \end{pmatrix}$$

and mean vector $m_{\text{RAM}} = (m, 0)$ a vector with $n_L + n_M$ entries. Let F be the $n_M \times n_L + n_M$ matrix given by

$$F = (0, Id_{n_M \times n_M})$$

where $Id_{n_M \times n_M}$ is the n_M-sized identity matrix. Then, the RAM notation SEM with matrices m_{RAM}, A, S, and F is the same SEM as given in the simple matrix notation.

Again, mastering these translations need some practice, but then allows to switch between the different notations. In fact, computer programs will do that for us, so we usually don't have to bother, but it's good to know that there is a 1:1 translation between these representations.

There are some more representations of SEM, partly tailored to specific instances of SEMs. We will not introduce them here since we have covered the ground from 'most intuitive' to 'mathematically most simple', but that doesn't discredit other representations in any way; it's just another way of describing the same underlying thing, an SEM, in other formats.

One source of miss-understanding in path diagrams and RAM is that the entries of Σ are the variances and covariances of the variables, not the entries of S (which are the labels on the double-headed edges in path diagrams). The labels on the double-headed edges are the *unique* variance of the variable; there is some sources of variance (and potentially covariance to other variables) that is added by the incoming single-headed arrows.

6.2 Parameter Estimation

Terms: Least Squares Estimation, Maximum Likelihood Estimation, Bayesian Point Estimation, FIML
Skills: Estimating parameters in SEM
Understanding: All three estimation methods are universal statistic for model

parameters. Bayesian estimation requires a-priori information, which is the only way to build knowledge, but includes necessarily subjectivity.

6.2.1 Bayesian Point Estimation vs. Maximum Likelihood

In all SEM notations, numerical values can be replaced by model parameters. For example, a label on a double-headed edge from a variable to itself (its variance) can be unknown. Such a parameterized model is strictly speaking a *family* of very many models (one for each possible parameter value), although in practice usually the word 'model' is still used for simplicity.

This is the point where empirical data enters the pictures: When we have a parameterized model, we can choose the parameters such that they fit our data best. For example, in our model predicting Archery, Biathlon and Downhill times from visual perception and skiing skills, we might be interested in the variances of these two latent variables based on a data set of some hundreds of athletes who have been tested in these three disciplines. Finding the 'right' member of a family of models, which means the 'right' parameter set, is called *parameter estimation*.

What 'right' means here, however, is unclear. If we assume that our family of models is *correctly specified*, which means that there is one member in the family that has the same distribution as the process that generated the data, the best answers seems to be to choose the parameter set that is most likely given the data. Let θ denote a vector of parameters; we would then like to choose θ such that the likelihood $\mathcal{L}(\theta|data)$ is maximal.[1] To compute the likelihood, we can use Bayes' Theorem. This estimation method is called *Bayesian Point Estimation*:

Definition 6.2.1. Estimating parameters by choosing θ to maximize

$$\mathcal{L}(\theta|data) = \mathcal{L}(data|\theta)\frac{\mathcal{L}(\theta)}{\mathcal{L}(data)} \tag{6.2.1}$$

is called *Bayesian Point Estimation*, and the resulting estimate is called the Bayesian Point Estimate.

Let us look at the ingredients of Equation 6.2.1 that we need to compute this: The a-posteriori likelihood for the data, $\mathcal{L}(data|\theta)$, is no problem, because the model together with values for the parameters θ will give us a distribution of the data using the previous section. The likelihoods without conditions in the fraction are the a-priori likelihood of θ and the *data*, and both are more difficult to access. The problem about $\mathcal{L}(data)$, however, is not our problem since the value is not dependent on θ, and we are just looking for a maximum of $\mathcal{L}(\theta|data)$. Therefore, $\mathcal{L}(data)$ is just a constant, and we can ignore it; with other words, the maximum of

$$\mathcal{L}(data|\theta)\mathcal{L}(\theta)$$

[1] strictly speaking, we should write $\mathcal{L}(\theta|data, \text{model family})$. However, for now we will assume that our choice of model is correct, and omit it for now

will be the same θ as if we maximize Equation 6.2.1. The $\mathcal{L}(\theta)$ can not be computed or simplified further, it needs to come from conceptual considerations. This is a confusing point for social scientists, because it means that our estimates will not only be results of the data set at hand, but be biased in the direction where the prior distribution $\mathcal{L}(\theta)$ is higher, even though this distribution comes more or less 'ad hoc'. If the a-priori likelihood is assumed to be constant, the $\mathcal{L}(\theta)$ has the same fate as the $\mathcal{L}(data)$: It can be omitted from the equation. What remains is the optimization of $\mathcal{L}(data|\theta)$, which comes by the name *Maximum Likelihood Estimation*:

Definition 6.2.2. Estimating parameters by choosing θ to maximize

$$\mathcal{L}(data|\theta)$$

is called *Maximum Likelihood Estimation.*

To reiterate, maximum likelihood estimation is a special case of Bayesian point estimation, under a constant prior.

The estimates from both methods are statistics for the true population parameters, and both are in fact true, that is, for infinitely many participants, both estimation methods will converge to the correct population values. For finite N, however, both methods have an estimation bias. Even though this bias is usually small, it is difficult to say up front whether the N is 'sufficiently large' for the bias to be tolerable.

There are ongoing discussions about how $\mathcal{L}(\theta)$ should be chosen. This discussion is completely non-mathematical, but all the more important, and you should stop after reading the next two paragraphs to consider the question and derive an own opinion, even if you should be prepared to be convinced otherwise later.

Roughly, there are two camps: the *Frequentists' Camp* say that $\mathcal{L}(\theta)$ necessarily includes a subjective choice because by design, it is the part that has nothing to do with the data. The consequence is to choose $\mathcal{L}(\theta)$ such that it carries no (or following the opinion of some in this camp, almost no) information. Such a prior is called a *flat* prior and is constant (or again, almost constant) in the range that the θ is expected, which leads to maximum likelihood estimation, a purely mathematical and objective process. This is the killer argument for the Frequentists' camp: Science needs to be objective to avoid that the scientists with most charisma (or the strongest military) decide what is true, and a-priori distributions always include this subjective component.

The *Bayesian Camp* claims that $\mathcal{L}(\theta)$ always should be included in the estimation. How to choose it depends on how much we already know about the parameters in question: If for example θ is the mean of intelligence, then values below 50 or above 150 should have very low a-priori probability, while values close to 100 (the constructed mean for the population) should be high. $\mathcal{L}(\theta)$ should be chosen more flat if less is known about the parameters, finally leading to a maximum likelihood estimation only as an extreme case. Even though the argument that $\mathcal{L}(\theta)$ always includes a subjective component is appreciated, the Bayesian Camp counters by the observation that using maximum likelihood

also implicitly claims a prior, which is a flat prior, and hence subjectivity is not eliminate, but just hidden, which is worse. The killer pro argument is that without priors, every new study will have to start from zero, which means there is no memory in science. In fact, without priors every estimation is useless to begin with, because after publishing the estimates, they can never again be used in later steps in the knowledge acquisition process of science.

There is no easy answer to this controversy (none that I know of), but it is already a remarkable step forward if psychologists will become aware of this problem's existence. Too many work on the presumption that exact values can be found, which is untrue: All we can do is get better and more information-rich distributions about all parameters. However, looking on the good side, observe that this is often completely sufficient to make good decisions: Assume we estimate the average depression level in a group of people that share a number of symptoms. Although we will not be able to find 'the' estimate for this level, if we get an a-posteriori distribution that only includes values below some critical stage, we can still make the decision that these people do not need medication (at least not based on the symptoms observed).

6.2.2 Maximum Likelihood Estimation in SEM

We will now discuss how to compute Maximum Likelihood and Bayesian Point Estimates for SEM. We start with the Maximum Likelihood estimate, that is, we try to find the maximum of

$$\mathcal{L}(data|\theta)$$

where $\mathcal{L}(data|\theta)$ is given by an SEM with K variables. Let $\Sigma(\theta)$ (a $K \times K$ matrix) and $\mu(\theta)$ (a K dimensional vector) be the normal distribution predicted by the SEM. Out of lazyness, we will omit the (θ) and just write μ and Σ, but remember that both depend on the parameter values. We want to fit these two matrices to our data $x_1, ..., x_n$, each a K-dimensional vector with the empirical outcome of the K variables. Since the SEM predicts a normal distribution, the likelihood of the data for a single participant i is the normal density which we know is

$$\mathcal{L}(x_i|\theta) = \frac{1}{\sqrt{(2\pi)^K |\Sigma|}} e^{-\frac{1}{2}(\mu - x_i)^T \Sigma^{-1} (\mu - x_i)}$$

Remember that $|\Sigma|$ is the determinant of Σ, and Σ^{-1} its inverse. As we have seen previously, this equation becomes much easier if we take the natural logarithm. Since we are searching for the best θ, that is, the θ that maximizes this equation, the logarithm does not change our result since the highest value will also have the highest logarithm. The expression then becomes

$$\log \mathcal{L}(x_i|\theta) = -\frac{1}{2} \left(K \log(2\pi) + \log(|\Sigma|) \right) - \frac{1}{2} (\mu - x_i)^T \Sigma^{-1} (\mu - x_i)$$

The $-\frac{1}{2}$ is also not helpful; we multiply the equation by -2. The only thing we have to pay attention to is that the maximum θ after multiplication with

a negative number now becomes a minimum. The resulting term is called the *Minus Two Log Likelihood*:

$$-2\log\mathcal{L}(x_i|\theta) = K\log(2\pi) + \log(|\Sigma|) + (\mu - x_i)^T\Sigma^{-1}(\mu - x_i)$$

The likelihood of all participants $x_1, ..., x_N$ is the product of the likelihood for each participant if we assume that the participants are independent of each other. This product becomes a sum if we take the logarithm, which means that the minus two log likelihood of the complete data set is

$$\begin{aligned} -2\log\mathcal{L}(data|\theta) &= -2\log\prod_{i=1}^{N}\mathcal{L}(x_i|\theta) \\ &= \sum_{i=1}^{N} -2\log\mathcal{L}(x_i|\theta) \\ &= \sum_{i=1}^{N}\left[K\log(2\pi) + \log(|\Sigma|) + (\mu - x_i)^T\Sigma^{-1}(\mu - x_i)\right] \end{aligned}$$

We can apply the sum to each of the three terms separately; since the first two don't depend on i, they just get multiplied by N. The result is

$$-2\log\mathcal{L}(data|\theta) = NK\log(2\pi) + N\log(|\Sigma|) + \sum_{i=1}^{N}(\mu - x_i)^T\Sigma^{-1}(\mu - x_i) \quad (6.2.2)$$

Observe that the first term $NK\log(2\pi)$ is a constant number, so if we want to find the optimum of this expression, we can ignore it.

This representation of the -2 log likelihood in Equation 6.2.2 is sometimes called the *Full Information Maximum Likelihood* equation and is under some conditions directly used to find θ, especially if some participants do not have data on all variables (this situation is called *missingness*). But let's ignore this side-track for now and assume that all variables are available. In this situation, we can pull a little trick with the sum over all data points to simplify the computation of the equation. For this, we need the following theorem:

Theorem 6.2.3. *If*

$$S = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T$$

then

$$\sum_{i=1}^{N} x_i^T\Sigma^{-1}x_i = N\mathrm{tr}\left(\Sigma^{-1}S\right)$$

You may want to check this theorem for some simple examples, although knowing the proof is not critical here.

We factor out the term $(\mu - x_i)$ in Equation 6.2.2 and get the at first glance more complex expression

$$\sum_{i=1}^{N}(\mu - x_i)^T\Sigma^{-1}(\mu - x_i) = N\mu^T\Sigma^{-1}\mu - 2\mu^T\Sigma^{-1}\left(\sum_{i=1}^{N} x_i\right) + \sum_{i=1}^{N} x_i^T\Sigma^{-1}x_i$$

Even though longer, we observe now that instead of using the x_i's, we can now write the expression only in terms of the data covariance matrix and data mean (let's call those m and S):

$$\sum_{i=1}^{N}(\mu - x_i)^T\Sigma^{-1}(\mu - x_i) = N\left(\mu^T\Sigma^{-1}(\mu - 2m) + \text{tr}\left(\Sigma^{-1}S\right)\right)$$

To put all this in one result block, we have shown that

Theorem 6.2.4. *The minus two log likelihood of an SEM with K variables, mean μ and covariance matrix Σ, for a data set with N participants and estimated data mean m and covariance matrix S, is given by*

$$-2\log\mathcal{L}(data|\theta) = N\left(K\log(2\pi) + \log(|\Sigma|) + \text{tr}\left(\Sigma^{-1}S\right) + \mu^T\Sigma^{-1}(\mu - 2m)\right) \tag{6.2.3}$$

Remember we want to find parameters θ that minimize this expressions, and Σ and μ depend on θ. So we can omit the first constant and the N, and just optimize

$$\log(|\Sigma|) + \text{tr}\left(\Sigma^{-1}S\right) + \mu^T\Sigma{-}1(\mu - 2m)$$

To find the minimum of this function, you have to take the derivative with respect to each parameter and set it to zero (which gives you one equation per parameter). The solution of this system of equations is your maximum likelihood estimate. Showing the derivatives here is more than we want to cover for now, and anyway good computer programs exist that can do these steps for us.

One problem is that the minimum is not necessarily unique. We should more strictly say that the maximum likelihood solution is *among* the solutions of the system of equations. Usually, the numbers of solutions is finite, so we can solve the equation system and compute the value of Equation 6.2.3 for each solution. There are, however, situations where an infinite number of solutions exist, which are called under-identified models. We will later come back to this situation.

6.2.3 Bayesian Point Estimation in SEM

If you decided that you are not a Frequentist, but you want to add the a-priori distribution to the estimation, little actually changes from the math we just did. In this case, we want to maximize

$$\mathcal{L}(data|\theta)\mathcal{L}(\theta)$$

or, analogously to the above, minimize the minus two log likelihood. The log will transform the product of the two likelihoods into a sum, so we want to minimize

$$-2\log\mathcal{L}(data|\theta) + -2\log\mathcal{L}(\theta)$$

The first summand is just the same as before. Remember we divided by N, so we have to do this to both terms; other than that, we can just copy from the last subsection to get a new function we want to minimize,

$$\log(|\Sigma|) + \operatorname{tr}\left(\Sigma^{-1}S\right) + \mu^T\Sigma{-}1(\mu - 2m) + \frac{1}{N}\left(-2\log\mathcal{L}(\theta)\right)$$

The $\frac{1}{N}$ factor is important for the discussion between Frequentists and Bayesians, since we see that it weights the prior: The more data points we collect, the less important the prior information becomes. Eventually, with $N \to \infty$, the term becomes the same as the minus two log likelihood, and both methods coincide.

Observe that the last term is a known distribution in θ, so we are done already. If the a-priori distribution is also a normal distribution (say with mean m_{prior} and covariance matrix C_{prior}), we can write it as

$$\frac{1}{N}-2\log\mathcal{L}(\theta) = \frac{P}{N}\log(2\pi)+\frac{1}{N}\log\left(|C_{prior}|\right)+\frac{1}{N}(\theta-m_{prior})^T C_{prior}^{-1}(\theta-m_{prior})$$

We can omit the first two terms since they are not related to θ and end up with an objective function for normal priors given in the following theorem.

Theorem 6.2.5. *The Bayesian Point Estimate of an SEM with K variables, model mean μ and covariance matrix Σ for a data set of N samples with sample mean m and sample covariance matrix S, with a normal prior on P parameters with prior mean m_{prior} and prior covariance matrix C_{prior} is the global minimum of the function*

$$\log(|\Sigma|) + \mu^T\Sigma{-}1(\mu - 2m) + \operatorname{tr}\left(\Sigma^{-1}S\right) + \frac{1}{N}(\theta - m_{prior})^T C_{prior}^{-1}(\theta - m_{prior}) \tag{6.2.4}$$

Again, Equation 6.2.4 can have multiple minima, so the optimization must find the global minimum, and again, the global minimum can be an infinite set of solutions which then all have the same a-posteriori likelihood.

Note that the choice of a normal distribution for the prior is not necessarily a good one. While the normal distribution is a good choice for means and a reasonable approximation for variances well away from zero, it is less suited for variances close to zero if we aim to keep these variances positive (which we not always do, as we will see later), and usually also questionable for more complex parameters as for example regression weights. In the end, the choice of a good prior distribution can be as tricky as the choice of a good model.

6.2.4 Other Estimation Methods for SEM

Both estimation methods discussed previously have one disadvantage in common, they are fairly evolved to compute and unstable if the model is not correctly specified, especially under violations of the normality assumptions. A very simple alternative is *least square estimation*.

Definition 6.2.6. For an SEM with K variables, model mean $\mu(\theta)$ and model covariance matrix $\Sigma(\theta)$ and a data set with N participants, sample mean m and sample covariance matrix S, the least square fit index is defined as

$$LS(\Sigma, \mu, S, m) = N \left(\sum_{i=1}^{K} (\mu_i - m_i)^2 + \sum_{i,j=1}^{K} (\Sigma_{i,j} - S_{i,j})^2 \right)$$

The θ that minimizes this index is called the *least square estimate*.

Least squares takes the comparison of the model distribution to the data distribution literal by minimizing the Euclidean distance between the mean vectors and all entries of the covariance matrix. Because of the simple structure of this index, the derivative is usually easy to compute, which makes the Least Squares index so efficient. The least square index is also a true statistic for the population parameters if the model is correctly specified. However, we lack the nice interpretation that the least square index is the 'most likely choice' as in Bayesian or, with flat prior, maximum likelihood estimation. In fact, least square estimation for correctly specified models typically performs worse than the maximum likelihood estimation in terms of bias. It is though often more robust against miss-specification. A frequent application of the least squares estimate is to use it as a starting value in the optimization of Bayesian or Maximum Likelihood estimation processes.

Least Squares is the simple alternative to the two estimation procedures introduced before. On the other end of the complexity spectrum, Bayesian point estimate can be extended to *Bayesian distribution estimates* (or just 'Bayesian Estimates' sometimes). The idea here is that each parameter of the distribution is again described by its own distribution. There are multiple variants of this estimation technique: The distribution of each parameter is described by a sample, or the distribution is again parameterized, and a point estimated of these parameters (called *meta-parameter*) are found like in Bayesian point estimates. The latter mechanism can be escalated to finding meta-parameters of meta-parameters for some of the parameters. It can easily be shown that the information at each level is less.

For our purposes, we will concentrate on maximum likelihood in the following, because from a teaching perspective it has all components we need to know. However, in later application, it is probably worth checking whether an alternative estimation method could be more successful.

6.3 Likelihood Ratio Tests

Terms: Nested Models, Minus Two Log Likelihood Ratio, Likelihood Ratio Test
Skills: Performing significance tests against hypothesis given by linear constraints on parameters
Understanding: The LR test is a universal test for linear constraint hypothesis, but needs sufficient number of participants and may be frugal to model

miss-specification.

6.3.1 Nested Models

Maximum Likelihood provides a universal statistic for parameters θ in a model. Good for us, it also comes with a universal test for hypotheses that can be described by a linear constraints on the parameters. A linear constraint is simply a linear combination of the parameter that must be a constant,

$$w_1\theta + w_2\theta_2 + ... + w_K\theta_K = c$$

which you can also write as

$$w^T\theta = c$$

Most often, the linear constraint simply sets one parameter to a fixed value, e.g., $\theta_3 = 5$. Here, θ_3 could for example be the mean of one our latent variables A. So this constraint describes the hypothesis 'A has zero mean'. But constraints can also express other hypotheses, as for example 'A has the same mean as B', which would be $\theta_3 - \theta_4 = 0$ (or equivalently, $\theta_3 = \theta_4$), where θ_4 is the mean of B.

If θ is a parameter vector of a model $\mathcal{M}$, then $\mathcal{M}$ with the constraint is a new model $\mathcal{M}_0$; in fact, if $\mathcal{M}$ is a SEM, then $\mathcal{M}_0$ is still a SEM since the distribution it describes is still a parameterized normal distribution. We call $\mathcal{M}_0$ *nested* in $\mathcal{M}$ if it can be described as '$\mathcal{M}$ plus linear constraints':

Definition 6.3.1. If a model $\mathcal{M}_0$ can be described as another model $\mathcal{M}$ with some linear constraints on the parameters, then $\mathcal{M}_0$ is called *nested* in $\mathcal{M}$. Sometimes, $\mathcal{M}$ is called the *full* model in contrast to the nested model $\mathcal{M}_0$.

6.3.2 The Minus Two Log Likelihood Ratio

If we have an hypothesis H_0 that is described by linear constraints, then a surprisingly simple test statistic emerges, which is the difference of the minus two log likelihood for the data without the constraint and the minus two log likelihood with the constraint:

Definition 6.3.2. Let θ be the maximum likelihood estimate for a dataset *data* in a model, and θ_{null} the maximum likelihood for the same data under a nested model. Then the difference between the minus two log likelihood values for θ and θ_{null},

$$LR = -2\log\left(\frac{\mathcal{L}(data|\theta_{\text{null}})}{\mathcal{L}(data|\theta)}\right) = (-2\log\mathcal{L}(data|\theta)) - (-2\log\mathcal{L}(data|\theta_{\text{null}}))$$

is called the *Minus Two Log Likelihood Ratio*.

Observe that this name, again, is just a description of the operation we do. Also note that the LR is always positive, since θ will always be a better fit than θ_{null} which suffers from the constraint.

Let's compute the LR for an SEM. For simplicity, we will shift the variables such that the data mean is zero, which will for larger N also cause the predicted mean for the full model to be zero. Let's write Σ for the covariance matrix of the full model (that is, for the model prediction using θ) and Σ_0 and μ_0 for the estimated of the nested model (that is, for the model distribution using θ_{null}). We have to subtract Equation 6.2.3 for the nested model from Equation 6.2.3 for the full model,

$$\begin{aligned} LR &= (-2\log\mathcal{L}(data|\theta_{\text{null}})) - (-2\log\mathcal{L}(data|\theta)) \\ &= NK\log(2\pi) + N\log(|\Sigma_0|) + N\left(\mu_0^T\Sigma_0^{-1}(\mu_0 - 0) + \text{tr}\left(\Sigma_0^{-1}S\right)\right) \\ &\quad -NK\log(2\pi) - N\log(|\Sigma|) - N\left(0 + \text{tr}\left(\Sigma^{-1}S\right)\right) \\ &= N\left[\log(|\Sigma_0|) - \log(|\Sigma|) + \mu_0^T\Sigma_0^{-1}\mu_0 + \text{tr}\left(\left(\Sigma_0^{-1} - \Sigma^{-1}\right)S\right)\right] \end{aligned}$$

Assume that H_0 is not true, that is, in the population the constraint is not observed. Then θ and θ_{null} will converge to different values as N increases, and the term in the brackets converges to some fixed positive value. As N increases, the LR value increases with the same speed; nothing surprising here. Assume however that H_0 is true in the population. Then as N increases, the term in the bracket approaches zero (since Σ_0 approaches Σ and μ_0 approaches μ, which is zero). At the same time, the N in front of the brackets increases. Very surprising and extremely useful is that both happen at the same speed, so that the LR approaches some fixed value. This value depends on the data set we got from the population and follows a known distribution, called the χ^2 distribution. The χ^2 distribution comes with a parameter called the *degrees of freedom* (df), which is the number of constraints we applied in H_0. Let's summarize this in one theorem.

Theorem 6.3.3. *Let H_0 be an hypothesis described by df linear constraints which is true in the population. Then, the LR converges in distribution to a χ^2 distribution with df degrees of freedom as N approaches infinity.*

Since the LR increases towards infinity if H_0 is not true but stays in a reasonable range, as defined by the χ^2 distribution, if H_0 is true, we can use the LR as a test statistic for a significance test. If the LR value is high, we reject H_0. What constitutes 'high' we decide by the χ^2 distribution and an α level we fix for the significance test. This test is called the *Likelihood Ratio Test*:

Definition 6.3.4. Let H_0 be an hypothesis described by *df* linear constraints and $\alpha \in [0, 1]$. Let *crit* be the value such that the probability to get a value above *crit* is α in a *df* degrees of freedom χ^2 distribution, and LR the minus two log likelihood ratio. The *Likelihood Ratio Test* of H_0 for level α rejects H_0 if $LR \geq crit$.

Observe that the LR test is in fact a significance test with significance level α if N is sufficiently high.

6.3.3 Likelihood Ratio Test Procedure

We can summarize this section up to here by a stepwise procedure how to perform a likelihood ratio test

1. Set up a model $\mathcal{M}$ for the data.
2. Describe a null hypothesis H_0 by linear constraints. Let *df* be the number of constraints.
3. Fix an α level for the test.
4. Find the maximum likelihood estimate θ for $\mathcal{M}$
5. Find the maximum likelihood estimate θ_{null} for $\mathcal{M}$ with the constraints H_0.
6. Compute the difference of both minus two log likelihood values.
7. Reject H_0 if the probability to get this or a higher differences on a χ^2 distribution with *df* degrees of freedom is smaller than α.

For sufficiently large N, the LR includes all tests on normal data used in social sciences, and all these tests can be described as special cases of the LR test. That is good, because it means that for normal data, it is sufficient to understand this one test. It is hence worth the effort to stop reading at this point and recapitulate whether you understood how the LR test works, and why it works.

6.3.4 Interval Test with the LR Test

If the parameter constraints describing H_0 are linearity constraints, then the LR is a test against an event that usually anyway has an a-priori probability of zero, which doesn't need testing. A better H_0 is a range of parameters. To use the LR test for such an hypothesis, two steps are necessary: We first check whether the maximum likelihood estimate itself is in the H_0 range. If it is, there is no way we can reject H_0, so we stop. Otherwise, we search for the most conservative point in the H_0, usually the one closest to the estimate, and perform a LR test against this point. The p value we then get is an upper bound for the p value when we would consider the whole H_0.

For example, assume H_0 states that the mean of a cognitive skill is smaller or equal 100. Our maximum likelihood estimate of the mean skill is 105; as this is not in the H_0 (105 is not smaller or equal 100), we perform a LR test against a fixed value of 100 (that is, the linear constraint is $mean_{skill} = 100$). Let's say the difference of the two minus two log likelihood values with mean skill = 105 and mean skill = 100 is 6.1; in a χ^2 distribution with 1 df, that would correspond to $p = 1.4\%$. Tests against other points in the H_0 (say, mean = 95) will always give larger minus two LR values than 6.1, and hence smaller p values; so 1.4 is an upper bound, and we can claim a significant result against H_0 at an α level of, say, 5 %.

6.3.5 Limitations of the LR Test

All over this section, you find the expression 'for sufficiently large N'. That is because for low N, the distribution of the LR if H_0 is true is not precisely a χ^2 distribution. In effect, the Type I error rate may be larger than the α level we fixed. The sad part is that 'sufficient' can not be quantify in general because it relies on the information provided from each single case. With a high number of observations for each participant, $N = 1$ can be fully sufficiently large for the LR test. But even in the extreme case of a single observation per data and a test of the mean for this single parameter[2], an N of 20 is already sufficient to control the α level within one percent. Considering that the α level is ad hoc anyway, one percent control is more than is needed.

Note that the LR test only works because the distribution of the full model approaches the true data distribution, but not if the data distribution is not part of the distributions the full model provides. This is a fairly big 'But', because as we discussed earlier, the main objective of models is to simplify the world; so by design, models will always not have a perfect fit to the data. It is possible to correct the LR for this case, but that is more than we can cover here. Nevertheless, in many situations the model can be made reasonably accurate to justify the usage of the LR test.

Also, don't forget that the LR test is only a significance test. It comes with all weaknesses of this family of tests: We don't know anything about the probability of H_0 after the test. Even if we reject the H_0 by the LR test, that only means that the likelihood for our finding was 'small' under H_0, not that H_0 itself is of low probability. So whenever we use the LR test, we have to link it to the probability of H_0 by statements about the power, i.e. the probability of a correctly rejecting H_0 if it really is not true H_0, and about the a-priori likelihood of H_0.

[2]This situation is called a T test, and there are known corrections for the distribution in this case, which is called the T distribution