Elementary statistics and causality

Thomas Colignatus, April 23 2007, preliminary first chapter

http://www.dataweb.nl/~cool

Start

Introducing causality and correlation

Causality versus correlation

Consider the game of billiards - "a game played with small, hard, heavy balls and long tapering sticks (called *cues*) on an oblong, cloth-covered table", Hornby (1985:81). When one ball hits another one and transfers momentum then the first is the cause of the subsequent movement of the other. What happens in reality with all those atoms and energy states differs from the model in our mind, differs from the words we use to describe both, differs from the (mathematical) model that we might make. The model can reflect the causality and capture it but the model does not embody it. Causality is what these balls do with each other, one being the cause and one being the effect.

The notion of **cause** may be taken as a primitive concept. This means that it is not further explained so that when you don't get it then you are lost to the discussion. However, since you have learned to read, you likely know about cause and effect. It suffices that we talk about cause and effect, so that you remember what it is. The example of the billiard balls probably helps too.

We say that rain causes the streets to become wet. Thus let c = "It rains" and e = "The streets are wet". The situation becomes more complicated when there are more causes and effects. For example c_1 = "It rains" would be our main cause of interest but we may

allow for some c_2 = "The city street cleaning car passed by" or something else that might cause the streets to get wet. Combining {cause, effect} with {single, multiple} with {And, Or} we get, focussing on c or c_1 with some possibly unknown c_2:

```
TypesOfCausality["Arrow"]
```

1	Single	$c \to e$
2	Only sufficient	$(c_1 \vee c_2) \to e$
3	Only necessary	$(c_1 \wedge c_2) \to e$
4	Multiple effects	$c \to (e_1 \wedge e_2)$
5	Uncertain effects	$c \to (e_1 \vee e_2)$
0	Not relevant	$\{c \to e', c' \to e\}$

Comments on the types are: (1) If rain would be a unique cause then there would be no other way of the streets to get wet, then there are no other causes to consider, and then rain would be both necessary and sufficient for the streets to become wet. A simple direct cause, thus has no intervening breaks in the causal chain, the cause always has the effect and the effect arises only from the cause. (2) If rain isn't a unique cause but still effective, then there are other causes, but because of the effectiveness rain is still a sufficient condition for wet streets. (3) If rain isn't a unique cause but also requires co-causes, then rain is necessary but not sufficient. In the latter case we can say that the effect is sufficient proof that all its (combined) causes have occurred. (4) A cause can also have multiple effects. If some cannot be observed directly then the observable effects can be used to predict the unobservables. (5) A tricky possibility is that a cause can have various effects, uncertain which it would be at any time. Like flipping a coin and see one of two possible outcomes, heads or tails. For a particular outcome e_i one would hold that it could be explained not only by the flip but also by additional causes, such that (5) reduces to (3).

Note that we can always replace $x = f[x_1, ..., x_n]$ for example with connectives from propositional logic, and for now it seems that cases 1 to 5 are sufficient for our analysis (with (5) already reducible). Causality nevertheless quickly gets complicated, even though we keep the discussion elementary. The complexity multiplies by combinations. (a) For example a relation like $c_1 \to c_2 \to e$ is a combination of (1) $c_2 \to e$ and (3) $(c_1 \wedge c_2) \to e$, so that we need not specify this as a 6th form, though it is useful to be aware of it and the possibility of this reduction. (b) Another example are multiple causes with multiple effects, e.g. the combination of (2) with (4). (c) A case discussed in the literature is the Mackie INUS condition: "an event *C* is perceived to be the cause of event *E* if *C* is "an *insufficient* but *necessary* part of a condition which is itself *unnecessary* but *sufficient* for the result"", with the understanding that *C* precedes in time to other sufficient causes and is the first sufficient cause to be completed. In the

legal literature this condition has the easier label NESS “necessary element of a sufficient set” Pearl (2000:314).

Simple causality can be taken as {(1), (0)} such that only those two alternatives are considered, so that two variables have a simple direct causal relation or not. Deterministic causality can be taken as {(0) to (4)}, excluding (5).

The notion of **correlation** means that phenomena occur at the same time but need not have a direct causal relation between them. For example, many people in Paris wake up between 7 and 8 o’clock in the morning and many people in Amsterdam wake up between 7 and 8 o’clock in the morning. Do the people in Paris wake up *because* the people in Amsterdam do so, or the other way around - with city inspectors or spies keeping track and co-ordinating alarm bells ? No, the true cause is sunrise at that latitude and longitude, common to both. When there is a national holiday in Amsterdam that allows people to turn off their alarm clocks then we see a different pattern between Amsterdam and Paris, and the other way around, which means that they don’t watch each other but the sun.

The distinction between causality and correlation holds *by definition* (and by these definitions). Books in statistics emphasize that correlation does not mean causation. For example, when research finds a strong correlation between smoking and cancer then this does not prove that smoking causes cancer. Yet, the notion that correlation and causation are different does not derive from deep statistical theory. It derives from mere definition. Books in statistics emphasize the difference though because of a human habit to anyhow confuse the two (even when those humans never heard about statistics before).

A property of causality is that a cause precedes the effect in time. This is **time’s arrow**. This is only a necessary property and not sufficient. If the people in Amsterdam wake a bit earlier than Parisians then they still are no cause. A necessary and sufficient condition to separate cause and effect is that you have a true **model** of them. The model explains what is the cause, what is the effect, and how these can be found true in empirical reality. Because causality is primitive and because having a true model is necessary and sufficient to handle causility, henceforth the notion of having a true model is a primitive concept too. Thus, the notions of a model and its relation to reality already presume the notion of causality, so they cannot be used to explain causality. When we discuss such issues the words around these notions only enrich them but do not further explain them.

We need to mention the notion of having a **reason**. Causes are in Nature while reasons are in the Mind. Logic and inference can be seen as the mirror of causality in nature. The mind mirrors nature to take better advantage of knowledge and experience about nature. Given natural selection, language, experiences and patterns of reasoning have been wielded out that seem to enhance survival. If we want to understand causality, we already get far by understanding reasoning, logic and inference.

Confounding

Note the triad *causality, correlation and confounding*. This book is called "Elementary Statistics and Causality" (ESAC) because a scientific theory relies on that triad. If the book were called "Elements of Causality" (EOC) then we could have evaluated TypesOfCausality[] and stopped there. It is in the interaction with data and the statistical handling that the complexity arises. Also, Nature first has causality and only secondly we extract data. From Nature's point of view this book should be titled "Elementary Causality and Statistics". For us however, starting as babies, we first collect data and then infer causality. We better express this processs in the title as well. Finding a good title for a book is a serious affair. But it would be confounding if we had "confounding" in the title as well, since we are not discussing "elementary confounding".

Confounding is the conceptual difference between causality and correlation. A confounder can hide a cause or create a seeming causal relationship. In both such cases there is a deviation from the true effect.

Pearl (2000:196) gives the example "that the effect of treatment is unconfounded if the treated and untreated groups resemble each other in all relevant dimensions". Kleinbaum et al. (2003:26) mention three types of research bias: study design, data selection, and "failure to adjust for variables other than the exposure [causal] variable (confounding)". It is also possible to over-adjust. PM. In epidemiology they also speak about exposure and response, equivalent to cause and effect.

If the cause is unique then we would always see the effect and have a correlation of 1 so that the difference between cause and correlation would be 0. (Except when the study is confounded.)

If the cause isn't unique then the correlation depends upon how often all causes occur, and there would be confounding if some of the causes are not measured or non-causes mistaken for causes.

As confounding gives the conceptual difference between causality and correlation, we might also consider taking the numerical difference. As we express the strength of a relationship by correlation, we might measure causality by the true correlation and confounding by the difference with observed correlation. It makes sense to say so, yet, the true difference resides in the causal **model**, which is something else than just a number measure. It seems better to say that the difference is *reflected* in such a number.

Let ρ be true correlation and R observed correlation. Note that both depend upon the study design. It would be awkward to speak about 'the' correlation between rain and wet streets in all space and time. It would neither be relevant since the correlation in our segment of space and time might differ from that overall theoretical value. It only makes sense to speak about, say, the streets in Amsterdam in 2060 and we might use that as a sample for the rest of Holland 2060 (and thus with the possibility of flooding due to global warming).

(1) With simple causality then there is a study design $d[\rho]$ such that $\rho = 1$ or 0. When $\rho - R \neq 0$ then we can say that there is confounding, due to the fact that we apparently have not set up the study such that the true causality directly shows itself.

(2) With simple causality then there can be a class of study designs $d[\rho] = \{d_1, ..., d_n\}$ that are more useful to us but such that $0 \neq \rho \neq 1$. When $\rho - R \neq 0$ then we can say that there is confounding, due to the fact that we apparently have not set up the study such that the true correlation is recovered. In this case $-1 \leq \rho, R \leq 1$.

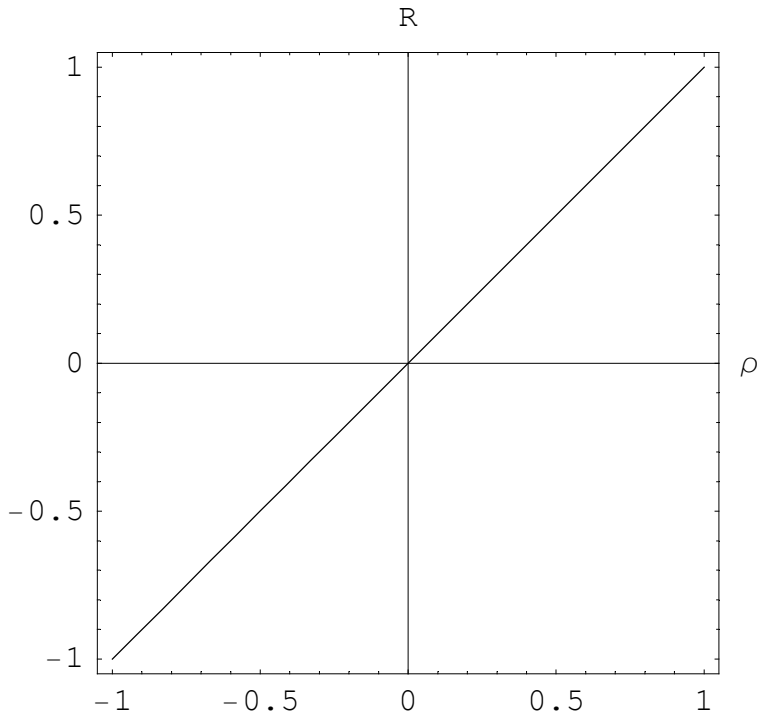
This summarizes the situation.

Cause["Correlation"]

	Cause	Correlation	Confounding
	ρ	R	$\rho - R$
True model (squared)	$\rho^2 \in \{0, 1\}$	$R^2 \in \{0, 1\}$	$\rho^2 - R^2 \in \{-1, 0, 1\}$
True model (not squared)	$\rho \in \{0, 1\}$	$-1 \leq R \leq 1$	$-1 \leq \rho - R \leq 2$
Statistical model	$-1 \leq \rho \leq 1$	$-1 \leq R \leq 1$	$-2 \leq \rho - R \leq 2$

We can also make a plot such that (i) the statistical model uses the whole area, (ii) the true model (squared) uses only the four corners of the upper right hand subsquare, (iii) the the true model (not squared) uses only the vertical lines $\rho = 0$ and $\rho = 1$.

```
ParametricPlot[{r, r}, {r, −1, 1}, AspectRatio → 1,
    Frame → True, AxesLabel → {"ρ", "R"}, TextStyle → {FontSize → 12}];
```



Time, space, matter and energy

One of the key notions of causality is that the cause precedes the effect in time.

Space or expandedness exists so that everything doesn't happen at the same time. Time or sequentiality exists so that everything doesn't happen at the same moment. We presume these. There are some physicists who suggest that time doesn't exist. Perhaps those physicists might follow the discussion by hypothesizing that time does exist.

Space and time are continuous, but there is no 'causality' there. Causality as a notion arises from matter or bouncing particles - while the distinction between particles and waves is mainly a conceptual one. How can an amount of energy of one particle / wave 'cause' another amount of energy of another particle / wave ? By hitting it, we presume, but hitting in itself does not explain what hitting does. In the end of such "philosophical" considerations, the example of the billiard balls may appear to be used as an analogy for causality and analogy only. Note that economists use the notion of *propensity* as for example the propensity to consume out of income. The notion of cause versus effect then

is replaced by *influence*. This is useful but shifts the problem to what ‘influence’ is. Note that physicists in the 1900s decided to drop these philosophical issues on ‘what things are’ or their ‘essence’ and set down to measure quantities. Quantative description was key to the surge in progress.

We may try to explain Time’s Arrow by the calculation order when the universe runs its equations. Perhaps the equations are such that they can be solved only in one direction. At least they are being solved in one direction.

For cause and effect we would like to see some local explanation in terms of matter or energy (or taking those the same as particle / waves) and not something like “distant causation”. When a stone drops from a mountain on the moon then according to Newton’ s model its different position immediately affects us, at infinite speed. This might be caused by a Big Bang that exploded in the dimensions of time and space but remained at zero in some dimension so that all things are still connected. Or we presume a prime cause at the time of some Big Bang such that all events here on Earth and on the Moon are just occurring simultaneously without directly affecting each other - so that Newton’ s model merely expressed a regularity and no causal connection. Or we switch to gravity waves that travel at the speed of light or their own superlightspeed. Or whatever else. Yet for a causal model we would like to see something like billiard balls hitting each other. That namely is the definition of a causal model. If you are not interested in that kind of model then this is not the book for you.

Causality and time in equations

An equation $y = \beta\, x + u$ may also be presented as $x = (y - u) / \beta = \alpha\, y + v$, with u and v error terms. Thus the first equation cannot express causality since the other equation would do so too. See Pearl (2000:159) for a discussion of stochastic considerations. Theil (1971:461) however mentions the criticism by Wold on systems of simultaneous equations that true models would be "recursive". When $y[t] = \beta\, x[t\text{-}1] + u[t]$ is the true model per day t of the week but observations are $Y[T] = \gamma\, X[T] + U[T]$ per week T, then it makes sense to relate Y and X of the same week, so that the basic causal delay does not show itself. The point is merely mentioned here and the discussions by Pearl and Theil on the stochastics are important. The issue thus is that the shift in research to statistical correlation can derive from the study design on the observations. It may well be that causality is no longer the issue and that the structure of the model has already been established by piecemeal tests. Given overall costs and benefits, it might be cheaper to work with aggregate data and only determine overall parameters and correlations. Modelling depends upon the intended use.

In such equations, a variable will have no effect when its coefficient is zero. This is numerically the same as dropping the variable from the scope of the explained variable though operationally the two approaches are different.

The meaning of Not

When we say “If it rains then the streets are wet” then it follows logically that “If not(the streets are wet) then not(it rains)”. This is logically true but it may be doubted whether these two sentences are observationally equivalent. The observation of not-rain might be more complicated than the observation of rain. There is a whole universe that is ‘not-rain’. An orange certainly isn’t rain and neither a wet street. But the observation of an orange does not allow us to say anything about the weather and the national infrastructure. In itself, this kind of question is a bit of a non-issue. When observing the weather, we catalogue events in “rain” or “no rain”. When observing the national infrastructure, we catalogue the streets as “wet” or “not wet”. It would be a fallacy to hold that when asked about observing ‘not rain’ we would leave the realm of weather observation, such that an orange might count as an observation. The ‘not’ as used in “If it rains then the streets are wet” is conditional to the dimension on which the variable is measured and not unconditional to the whole universe. Thus the contraposition of the implication has the same observational content when we keep in mind in what dimensions the variables are measured.

Nevertheless, it is a useful observation that oranges have nothing to do with the relation “If it rains then the streets are wet”. Thus, suppose that we test a theory “If oranges are green then the streets are wet”. We observe oranges in their own dimension with categories “green” and “ripe”, and we may observe some correlation that oranges might turn ripe after the rain season. The implication however will not hold since we can observe a green orange and a dry street. Perforce, that causation does not hold. Oranges are in no way a causal factor for wet streets. Unless of course when we squeeze them, collect the juice, purify the water, and spray our streets. But we then rather would say that our effort to do so is the cause and not the oranges. Unless, of course, all this would be relevant, and we would really have to model this. Rather, though, we would hold that it is useless to apply the logical ‘not’ to the various categories in the dimension for oranges. If we do that, we remain in that dimension and we remain stuck with oranges. We can also apply the ‘not’ to the variable itself, expressing that oranges are not a cause whether green or ripe. We might use a separate operator like NotSuch to express that Not then is applied to the variable, yet it might suffice that we are aware of the distinction. The notion ‘not’ is rather universal and we should be hesitant about introducing all kinds of “not” operators.

Thinking over these issues, it also transpires that cause and effect are a bit different in their contraposition and are not as straightforward as the implication. Thus when C

causes E then we don't say that $\neg E$ causes $\neg C$. And when E is the effect of C then we don't say that $\neg C$ is the effect of $\neg E$. One reason that prohibits us from saying so is the dimension of time which requires that all causes precede the effect. When the cause C is at moment t and E is at moment t+1, it becomes a question: where in time are $\neg C$ and $\neg E$ located ? We can distinguish two possibilities: (1) When $\neg C$ is at the same time as C and $\neg E$ is at the same time as E, then $\neg E$ cannot cause $\neg C$ since something of t+1 cannot cause something at t. (Some physicists call this backward causation, which is another group than those who deny the existence of time at all. This is also reminiscent of teleology, were a goal for the future (Greek: $\tau\epsilon\lambda$os) "causes" people to do all kind of things now.) (2) When $\neg C$ are all causes but just not at t, and similarly for $\neg E$, then these have overlapping time periods, and also drop out from causal consideration. Thus in both cases (1) and (2) when C causes E then we cannot simply apply "not" as in logic. The useful contraposition is that when we don't observe the effect then we know that the cause has not occurred, which is reasoning backwards, but this is not backward causation. See also the discussion below on the application of "not" when we have properly taken care of the time dimension.

Notation

We might use the plain arrow to indicate causality $c \to e$. However, the arrow is rather easily abused with multiple confusing notions at the same time. Also, when we work in the environment of *Mathematica* then we better use the arrow for replacement rules, and when we start replacing then we should not unintentionally replace a causal relationship. The following special types of arrows are more expressive. Note the following: (i) $c \rightleftharpoons e$ remains asymmetric because of the left and right hand sides, (ii) for multiple causes and effects we use a bar on the side of the multiple, (iii) we express $\vee$ and $\wedge$ with arrow heads that suggest the same, (iv) there is a difference between (asymmetric) *relevance* as a cause and (symmetric) *relatedness* that allows causality both ways. PM. The table also includes the projection into logic. That projection destroys some subtleties so that these logical expressions may become confusing.

TypesOfCausality[]

N	Label	Function	Expression	Definition	Logic
1	Single	Cause	$(c \rightleftharpoons e)$		$(c \Leftrightarrow e)$
2	Only sufficient	OnlySufficient	$(c_1 \mapsto e)$	$((c_1 \vee c_2) \rightleftharpoons e)$	$(c_1 \Rightarrow e)$
3	Only necessary	OnlyNecessary	$(c_1 \mapsto e)$	$((c_1 \wedge c_2) \rightleftharpoons e)$	$(e \Rightarrow c_1)$
4	Multiple effects	CauseToo	$(c \rightarrowtail e_1)$	$(c \rightleftharpoons (e_1 \wedge e_2))$	$(c \Leftrightarrow e_1$
5	Uncertain effects	CauseChance	$(c \rightarrowtail e_1)$	$(c \rightleftharpoons (e_1 \vee e_2))$	$(e_1 \Rightarrow c)$
9	Any of 1 to 4 (5), not 0	CausallyRelevant	$(c \updownarrow e)$		
0	Not relevant	CausallyIrrelevant	$(c \downarrow e_1)$	$((c \rightleftharpoons e') \wedge (c' \rightleftharpoons e))$	
10	Related	CausallyRelated	$(c \updownarrow e_1)$	$((c \updownarrow e_1) \vee (e_1 \updownarrow c))$	
−1	Not related	CausallyUnrelated	$(c \downarrow e_1)$	$((c \downarrow e_1) \wedge (e_1 \downarrow c))$	

An example of the use of this notation is the INUS cq. NESS condition:

NESSCause["Example"]

$(((c_1 \wedge c_2) \vee c_3) \rightleftharpoons e)$

and we might say that $c_1 \in C \wedge C \rightarrowtail e$.

For a simple direct cause, thus without any intervening breaks in the causal chain, and such that the cause always has the effect and such that the effect arises only from the cause:

Implies[Cause[c, e], And[Time[c, e], Equivalent[c, e], "¬ ∃ y Path[c, y, e]"]]

$((c \rightleftharpoons e) \Rightarrow (\text{Time}(c, e) \wedge (c \Leftrightarrow e) \wedge \neg \exists\, y\ \text{Path}[c, y, e]))$

Simple cause & effect imply these consequences but are not equivalent to them. The expression Time[*c, e*] summarizes:

Implies[Cause[c, e], Exists[Δt, Δt > 0, Equivalent[c[t], e[t + Δt]]]]

$((c \rightleftharpoons e) \Rightarrow \exists_{\Delta t, \Delta t > 0}\ (c(t) \Leftrightarrow e(t + \Delta t)))$

Implies[Effect[e, c], Exists[Δt, Δt > 0, Equivalent[¬ e[t], ¬ c[t − Δt]]]]

$(\text{Effect}(e, c) \Rightarrow \exists_{\Delta t, \Delta t > 0}\ (\neg\, e(t) \Leftrightarrow \neg\, c(t - \Delta t)))$

Using the proper contraposition find that these expressions are logically the same. PM. We dropped the $\forall t$, and it may be noted that each t can have its own Δt.

Quantifiers are key

To understand causality we actually must use quantifiers, thus ∀ (for all) and ∃ (exists). For example, when we would express the "sufficient but not necessary" condition $p \mapsto q$, and use propositional logic only, then we would get the rather absurd result that this is only satisfied by a false p and a true q.

```
TruthTable[(p ⇒ q) && Not[q ⇒ p]]
```

p	q	$((p \Rightarrow q) \wedge \neg (q \Rightarrow p))$
True	True	False
True	False	False
False	True	True
False	False	False

What $p \mapsto q$ actually means is more complex:

```
ForAll[t, p[t] ⇒ q[t + Δt]] && Exists[t, ¬ p[t] && q[t + Δt]]
```

$$(\forall_t (p(t) \Rightarrow q(t + \Delta t)) \wedge \exists_t (\neg p(t) \wedge q(t + \Delta t)))$$

Thus, it is not sufficient just to consider all logical possibilities but we have to use more observations over time, causing us to transcend from logic to statistics.

This also helps us to understand what we mean with $C \downarrow E$. This is a symmetric relation:

```
CausallyUnrelated[Definition, p, q]
```

$$((p \downarrow q) \wedge (q \downarrow p))$$

And for a single case of asymmetric irrelevance we need to keep track of time:

```
CausallyIrrelevant[Implies, p, q, t]
```

$$(\exists_t (p(t) \wedge \neg q(t + \Delta t)) \wedge \exists_t (\neg p(t) \wedge q(t + \Delta t)))$$

Statics and dynamics

We already had the difference between equality (==), definition (≡) and assignment (=) (using the notation in *Mathematica*). Equality is a condition that can be tested and evaluates to True or False. A definition would be necessarily true unless it causes a contradiction. Assignment sets a variable to some value and further works with that value. In logic we had equivalence ⇔, implication ⇒ and inference ⊢. Equivalence and implication are relations that can be tested to be True or False. Inference produces a conclusion. Inference can be valid or invalid. ALOE interpreted the distinction between implication ⇒ and inference ⊢ as the distinction between statics and dynamics. We can extend that as follows. First note that there is also the relation Because[*c, e*] meaning that *e* is the case because of *c*.

```
Because[c, e]
```

$c \because e$

Collecting all these symbols we get:

```
TypesOfCausality["Time"]
```

	Static Symmetric	Static Asymmetric	Dynamic Asymmetric
Logic	⇔	⇒	⊢
Equations	==	≡	=
Causality	↥	⇌	∵

There might be a structural isomorphism between the various concepts. In logic, proof theory discusses various properties of axiomatic systems. It may be that such results could be easily transformed to the causal realm, and it may be that in fact such proof theory was originally created by thinking in a causal manner anyway. Yet it seems fair and perhaps even advisable to develop the issue of causality on its own merits. Thinking hard about causality might help us finding some key difference. And if we don't find a key difference and everything is just an isomorphism, then at least we have thought hard about the matter and then know for sure that there isn't a real difference.

Proving causality

When we write $c \rightleftharpoons e$ then we merely express or denote that *c* is a simple cause for *e.* Such as when we would state "rain causes wet streets". It is another matter how we are going to prove such a causality. Here the notion of "conditional independence" will be useful. Before discussing this notion, let us see how far we can get with understanding causality without this notion.

Structure of the next

We will start with one cause and one effect. Subsequently, we will introduce a third variable, that can either be irrelevant or a hidden cause or a confounder.

We can express causality and correlation in an absolute manner (yes / no) or numerically by some degree. For now we take causality as yes / no and correlation as a degree.

Before trying to express causality it is easiest to first express correlation.

This discussion presumes that you know logic, preferably from reading ALOE. We start with ALOE (2007:137), the example on empirical induction.

Note that logic uses truth tables and that the next step in statistical observation is the use of contingency tables or crosstables. These tables collect numbers of observations on categorical or nominal data. Hence we use a concept of "nominal correlation" to discuss causality, correlation and confounding.

The 2 × 2 case

The example from ALOE on empirical induction

You note that sometimes it rains and sometimes it doesn't rain. You also observe that sometimes the streets are wet and sometimes they aren't. You become interested in the combined occurrences. You collect some 100 cases. In all the 25 cases that it rains the streets are wet. There are 3 cases where it didn't rain but the streets are still wet (e.g. because the road cleaning car came by). The data can be collected in the following table. We assume that rain is the cause. Is the observation that the streets are wet a good predictor for what is the cause ?

```
mat = {{"Observation count", "It rains", "It doesn't rain", "Total"},
       {"The streets are wet", 25, 3, □},
       {"The streets are not wet", 0, □, □},
       {"Total", □, □, 100}};
```

```
res = Headed2DTableSolve[mat]
```

Observation count	It rains	It doesn't rain	Total
The streets are wet	25	3	28
The streets are not wet	0	72	72
Total	25	75	100

Instead of remembering all these 100 cases either individually or by frequency distribution, the memory processing unit might save on storage and retrieval costs by adopting a general rule that "If it rains then the streets are wet".

This can become a general rule for which we can use a truthtable. The truthtable tests the condition whether the frequency is zero or non-zero.

- The SquareTruthTable procedure sorts the texts in alphabetical order, unaware of our hypothesis on causality. Thus we transpose its result.

```
SquareTruthTable["It rains" ⇒ "The streets are wet"] // Transpose
```

(It rains ⇒ The streets are wet)

	It rains	¬ It rains
The streets are wet	True	True
¬ The streets are wet	False	True

Clearly, such general rules must be treated with care since you might run into cases where the rule is refuted. For example, a street running under a bridge or through a covered shopping mall might appear to be dry even while it is raining.

When we know only two variables then there are no alternatives such as “street under a bridge” or “street under a tunnel” and hence the issue becomes quite simple. To get to the core of things we maintain that simplicity for a while till we introduce the possibility of a third variable.

Variables and scales

In above example the causal variable is the weather, with categorical values “rain” and “no rain”. The effect variable is the status of the streets, with categorical values “wet” and “not wet”.

(1) The weather and the status of the streets are measured with a nominal scale, as distinct from measuring them with a ratio scale. A ratio scale might use humidity of the air and inches of rain for the weather, and resistance or degree of slippery-ness or inches of water for streets. For both variables it might make most sense to use inches of water, origin rain or other, since this makes for easy conversion.

(2) Every-day language is that we say that the rain is a cause that makes streets go wet while it is scientese to say that the weather has causal impact on the status of roads. The scientese expression collects various combinations of every-day expressions. The every-day expression that rain causes wet streets implicitly contains the notion that there may also be *no-rain* and *not-wet*, yet, it might also be possible that one lives in a country where it always rains, and hence the scientific expression better captures the various categories.

2 × 2 tables and their border sums

Contingency tables generally are presented with table-headings and border-sums. Calculations are normally done with the inner matrix. Thus we will use routines that insert and strip those fringes whenever needed. In epidemiology we can define a 2 × 2 table as an object with the name DiseaseTestMatrix[], standardly defined with the cause (disease) in the columns and the effect (test results) in the rows, such that for this object all kinds of statistics like sensitivity, specificity and odds ratio are available by default. When a 2 × 2 table is presented as a 3 × 3 table with bordersums, the degrees of freedom is 4, and it is handy to have freedom to insert 4 values anywhere and calculate the remainder.

In probability theory there is also the approach of conditional probability in which one takes the border sums (marginal probabilities) as given. In that case there is only 4 - 3 = 1 degree of freedom (since the total of each border sum is equal to the total of the other border sum).

The following is an example where men and women are dieting or not. Observed frequencies are *a* to *d*. We wonder whether the behaviour of the groups differs.

```
DiseaseTestMatrix[] = mat2 = Bordered2DPrSolve[ a  b  □
                                                c  d  □ ];
                                                □  □  □
```

```
DiseaseTestMatrix[Table, "men", "dieting"]
```

	men	¬ men	Tested
dieting	a	b	$a+b$
¬ dieting	c	d	$c+d$
Sum	$a+c$	$b+d$	$a+b+c+d$

For analysis, we rely on the inner matrix only.

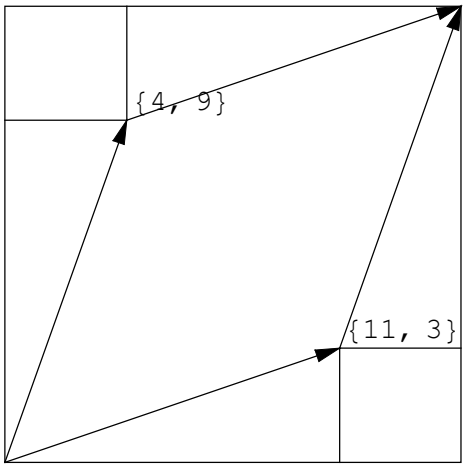
```
mat3 = StripMatrix[mat2]
```

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Diagram of the 2 × 2 case

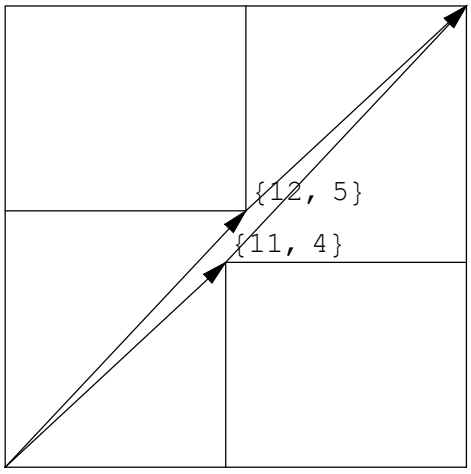
The 2 × 2 matrix {{*a, b*}, {*c, d*}} contains two row vectors {*a, b*} and {*c, d*} that together span a parallelepiped in two-dimensional space. When we draw a diagram of this, we find that the parallelepiped is contained in a rectangle with sides (*a* + *c*) and (*b* + *d*) which are the column sums of the matrix. The following gives a numerical example. Recall that we discuss numbers of observations so that we only consider nonnegative matrices.

```
ShowDet[{{11, 3}, {4, 9}}, AspectRatio → 1];
```



This is a plot of a case where the numbers are close together

```
ShowDet[{{11, 4}, {12, 5}}, AspectRatio → 1];
```



When the vectors are closer together the area of the parallelepiped is closer to zero, and when the vectors are further apart then the area of the parallelepiped is closer to covering the whole rectangle (which turns in a square when there is full equality). We can use the rate of coverage as a measure of association between the variables.

A measure for association or correlation

The total area of the rectangle is given by $(a + c)(b + d)$ while the area of the parallelepiped can be found by subtraction of the small rectangles and triangles, thus $(a + c)(b + d) - 2\,bc - 2 * (\frac{1}{2}ab) - 2 * (\frac{1}{2}cd) = ad - bc$. This latter value is the determinant of the matrix.

```
(a + c) (b + d) - 2 b c - 2 (1/2 a b) - 2 (1/2 c d) // Simplify
```

$$a\,d - b\,c$$

```
Det[{{a, b}, {c, d}}]
```

$$a\,d - b\,c$$

When we take the ratio of the areas $cr = (ad - bc) / ((a + c)(b + d))$ then we find a number between -1 and 1.

Note also that the determinant $ad - bc$ also holds for the dual (transposed) matrix, giving a ratio rr.

Since it is arbitrary which variable influences the other, a more robust measure is the geometric average $\sqrt{cr * rr}$. The numerator remains $ad - bc$ but the denominator becomes $\sqrt{}((a + c)(b + d)(a + b)(c + d))$. This gives us a "standardized surface ratio".

We shall take this ratio as a measure of association or correlation between the variables.

```
CorrelationPr2By2[{{a, b}, {c, d}}]
```

$$\frac{a\,d - b\,c}{\sqrt{(a + b)\,(a + c)\,(b + d)\,(c + d)}}$$

We can easily check that a diagonal matrix with $b = c = 0$ gives outcome +1 and with $a = d = 0$ gives outcome -1. Nominal data have no natural order, but one cannot avoid an order of presentation and the sign of the correlation in this case reflects that.

PM. CorrelationPr2By2 has been defined to take the inner 2 by 2 matrix even when input is larger, thus allowing for input with border sums.

Statistical independence means zero correlation

When the vectors in the contingency table are algebraically dependent, i.e. that a row (column) is a linear combination of the other rows (columns), then the determinant is zero, conforming to the idea that the association is zero as well. An extreme case is called statistical independence, when the matrix arises by multiplying the marginals (and with the total number of observations).

```
mat4 = PrTable[t, p] n
```

$$\begin{pmatrix} n\,p\,t & n\,(1-p)\,t \\ n\,p\,(1-t) & n\,(1-p)\,(1-t) \end{pmatrix}$$

```
CorrelationPr2By2[mat4]
```

0

Logical configurations

The 2 × 2 matrix {{*a, b*}, {*c, d*}} can have various equivalent configurations. A (logical) configuration is determined by the positions of zero values. The number of possible configurations is potentially large but can be reduced by equivalence, when one configuration can be turned into the other by permutation of rows or columns. We can achieve such permutation directly or by relabeling the variables. For example “not wet” can become “dry” and “not raining” can become “clear sky”.

Three zero’s

The minimal matrix that is non-zero has all observations in one spot only. This could be a country where it only rains.

```
res1 = Bordered2DPrSolve[ 100  0  □
                           0   0  □ ]
                           □   □  □
```

$$\begin{pmatrix} 100 & 0 & 100 \\ 0 & 0 & 0 \\ 100 & 0 & 100 \end{pmatrix}$$

```
SquareTruthTable[("It rains" ⇒ "The streets are wet") && "It rains"] // Transpose
```

((It rains ⇒ The streets are wet) ∧ It rains)

	It rains	¬ It rains
The streets are wet	True	False
¬ The streets are wet	False	False

Correlation is indeterminate.

```
CorrelationPr2By2[res1]
```

Indeterminate

And what about causality ? It is similarly indeterminate.

One column only

This is when it only rains but the streets can also be dry.

```
res2col = Bordered2DPrSolve[ 25  0  □
                             75  0  □ ]
                              □  □  □
```

25	0	25
75	0	75
100	0	100

Correlation is indeterminate.

```
CorrelationPr2By2[res2col]
```

Indeterminate

However, rain does not convince as a cause for wetness. That the streets would be dry contradicts the implication that rain is sufficient for wetness.

```
("It rains" ⇒ "The streets are wet") &&
    "It rains" && Not["The streets are wet"] // LogicalExpand
```

False

PM. Rain could still be a cause for wetness provided that there is a third variable that causes that the streets remain dry (e.g. a mall).

One row only

```
res2row = Bordered2DPrSolve[ 25  3  □
                             0   0  □
                             □   □  □ ]
```

$$\begin{pmatrix} 25 & 3 & 28 \\ 0 & 0 & 0 \\ 25 & 3 & 28 \end{pmatrix}$$

Correlation is indeterminate.

```
CorrelationPr2By2[res2row]
```

Indeterminate

```
SquareTruthTable[("It rains" ⇒ "The streets are wet") && "The streets are wet"] //
  Transpose
```

((It rains ⇒ The streets are wet) ∧ The streets are wet)

	It rains	¬ It rains
The streets are wet	True	True
¬ The streets are wet	False	False

The observations are consistent with a causal effect and don't refute it.

However, apparently we are observing streets in Atlantis, that are always wet (after the catastrophe). Our observational set-up allows for only two variables and does not allow for a third variable, such as being immersed under sea. Hence, without introducing the possibility of a third variable, we would not observe this.

Positive diagonal

The direction of the diagonal affects the sign of the correlation.

```
res2diag = Bordered2DPrSolve[ 25  0  □
                              0   72 □
                              □   □  □ ]
```

$$\begin{pmatrix} 25 & 0 & 25 \\ 0 & 72 & 72 \\ 25 & 72 & 97 \end{pmatrix}$$

```
CorrelationPr2By2[res2diag]
```

```
1
```

The causality is not refuted. When we do not allow for a third cause to wet the streets, then this is the result that we will observe. Rain and wet streets are observationally equivalent. Rain is the only cause and is not only sufficient but also necessary.

Negative diagonal

The configuration with a negative diagonal is equivalent to the positive diagonal. We just relabel the categories.

```
res2neg = Bordered2DPrSolve[ 0  3  □
                             72 0  □
                             □  □  □ ]
```

$$\begin{pmatrix} 0 & 3 & 3 \\ 72 & 0 & 72 \\ 72 & 3 & 75 \end{pmatrix}$$

```
CorrelationPr2By2[res2neg]
```

```
−1
```

However, it is useful to observe that, if we stick to the original labels, this refutes that rain would be a cause. It rain is a cause, we would not observe this. Thus negative correlation refutes a causal effect (into that direction). If you would observe that smoking is associated with lower mortality then this refutes a positive causal effect, that is, when you can exclude a confounder.

One zero only

There are four possible positions to allocate one zero and positive numbers elsewhere. But all those possibilities can be reduced by permutation of rows and columns to the notion of *implication*. For example, consider the following table, not rain and wet streets but ret and wain streets.

```
mat5 =  "Observation count"      "It rets"  "It doesn't ret"  "Total"
        "The streets are wain"   25         3                 □
        "The streets are not wain" □        0                 □
        "Total"                  □          □                 100 ;
```

res3 = Headed2DTableSolve[mat5]

Observation count	It rets	It doesn't ret	Total
The streets are wain	25	3	28
The streets are not wain	72	0	72
Total	97	3	100

In this case we have to permutate the columns and then find the proper implication to summarize the logic of the situation. We might replace "not retting" with the more positive sounding "ter", so that we can say more positively that if it ters then the streets are wain. It is all just a matter of flipping labels so that all such seemingly different configurations all are just one, the *implication*.

SquareTruthTable[Not["It rets"] ⇒ "The streets are wain"] // Transpose

(¬ It rets ⇒ The streets are wain)

	It rets	¬ It rets
The streets are wain	True	True
¬ The streets are wain	True	False

Similarly for the effect. When a zero occurs in the top row, we relabel the effect.

The correlations indicate that rain might be cause for wet streets but that retting is not a cause for wain streets.

CorrelationPr2By2[Part[res, {2, 3}, {2, 3}]] // N

0.92582

CorrelationPr2By2[Part[res3, {2, 3}, {2, 3}]] // N

−0.282008

But of course, if we consider only two variables, then we would not observe any off-diagonal element. And, when we introduce a third variable then this might still be a confounder.

No zero's

A final configuration is that there are no zero's.

```
res4 = Bordered2DPrSolve[ 25  3   □
                          □   55  □
                          □   □   100 ]
```

$$\begin{pmatrix} 25 & 3 & 28 \\ 17 & 55 & 72 \\ 42 & 58 & 100 \end{pmatrix}$$

The correlation now indicates a strength of relationship but is vague on causality. When we are testing whether rain causes wet streets then the observation of dry streets refutes such causality.

```
CorrelationPr2By2[Part[res4, {2, 3}, {2, 3}]] // N
```

0.0670431

A negative correlation arises when off-diagonal elements dominate.

```
res4neg = Bordered2DPrSolve[ 25  30  □
                             □   1   □
                             □   □   100 ]
```

$$\begin{pmatrix} 25 & 30 & 55 \\ 44 & 1 & 45 \\ 69 & 31 & 100 \end{pmatrix}$$

```
CorrelationPr2By2[Part[res4neg, {2, 3}, {2, 3}]] // N
```

−0.244904

A special case of course is statistical independence

```
mat4
```

$$\begin{pmatrix} n\,p\,t & n\,(1-p)\,t \\ n\,p\,(1-t) & n\,(1-p)\,(1-t) \end{pmatrix}$$

Provided that there are no zero's this would be a sufficient condition to conclude that there is no causal relationship. Weather and the status of the roads would have their own independent pattern and would be unrelated. But this assumes the existence of at least a third variable with the risk that this could be a confounder.

Intermediate conclusion

Negative correlation rejects causality in that direction but does not prove it for the other direction. To determine causality, we would test whether at least one zero exists. But in all cases the data can at best show only correlation since causality follows from the model anyway. We haven't defined yet a function $C[x, y \mid \text{data}] = 1$ or 0 that determines whether x causes y or not.

The 2 × 2 × 2 case

Rain reconsidered

If a variable has only 1 possible outcome then we can neglect it while 2 possible outcomes is the minimum to make a distinction. For convenience we stick to binary variables. Let us assume a third cause. This means that the 2 × 2 table actually is a bit misleading since, when all 3 variables have 2 possible outcomes, then we have $2^3 = 8$ possible combinations. Let us assume that the streets can also become wet because of the city street cleaning. It makes sense to assume that they don't do that when it already rains or let us assume that it rained for a long while and did some cleaning anyway. The following table gives a two-dimensional layout of the 8 combinations.

mat24 =

"Observation count"	"Rains" "Cleaning"	"Rains" "No cleaning"	"No rain" "Cleaning"	"No rain" "No cleaning"
"The streets are wet"	1	24	3	0
"The streets are not wet"	0	0	0	□
"Total"	□	□	□	□

;

res24 = Headed2DTableSolve[mat24]

Observation count	Rains	Rains	No rain	No rain	Total
Null	Cleaning	No cleaning	Cleaning	No cleaning	Null
The streets are wet	1	24	3	0	28
The streets are not wet	0	0	0	72	72
Total	1	24	3	72	100

What is important of this matrix is that we find that there is a *diagonal*: rain or cleaning is sufficient and necessary for wet streets. This means that correlation now is 1. We now don't use the 2 × 2 correlation measure but the more general NominalCorrelation.

```
NominalCorrelation[Part[res24, {3, 4}, {2, 3, 4, 5}]] // N
```

1.

This decomposition can also be done for any matrix, also for those without any zero.

A disease test matrix

Let there be a disease that is difficult to observe, e.g. brain cancer. The hospital has a "golden standard test" to determine whether a patient has that cancer or not. This test may be an expensive post-mortem on 100 persons. There is also a alternative cheap diagnostic test available but that is not accurate. This test might be based on a urine sample and it has been taken from those 100 persons while they were still alive. If the cheap test has good predictive value then the hospital might consider treating the test-positives. Suppose that the results are the following - and the matrix is called a "disease-test-matrix".

```
res22a = Headed2DTableSolve[ {"Count"    "Disease"  "Healthy"  "Total"
                              "Positive"  25         13         □
                              "Negative"  12         □          □
                              "Total"     □          □          100 } ]
```

Count	Disease	Healthy	Total
Positive	25	13	38
Negative	12	50	62
Total	37	63	100

Clearly when 13 of 38 test-positives are still healthy then the test has dubious value, in particular if the treatment against cancer would have important negative side-effects. Nevertheless, the test is interesting due to the strength of the association.

```
CorrelationPr2By2[Part[res22a, {2, 3}, {2, 3}]] // N
```

0.46683

Thinking in terms of causality it appears a bit strange that the test gives such probabilistic result. When it rains, the streets are always wet, unless there is some exception like a tunnel or a shopping mall. The diagnostic test should also obey those laws of causality and always generate the same results, unless there are such exceptions.

A key insight is that the patients are not homogeneous but can be split up into subgroups, namely those who are sensitive to the test and those who are not. Sensitivity to the test may also differ per disease state. The true table thus is more like this.

```
res24a = Headed2DTableSolve[
  {"Count"     "DisSen"  "DisIns"  "HealSen"  "HealIns"  "Total"
   "Positive"  25        0         13         0          □
   "Negative"  0         12        0          □          □
   "Total"     □         □         □          □          100}]
```

Count	DisSen	DisIns	HealSen	HealIns	Total
Positive	25	0	13	0	38
Negative	0	12	0	50	62
Total	25	12	13	50	100

The sensitivity and specificity of the test will be discussed below at length in the chapter on epidemiology. For now it suffices to observe the following. A background "false sensitivity" for people in health of 13 / 63 is increased by the disease to an observed sensitivity 25 / 37 (in the proper definition of sensitivity), though that latter rate might also be seen as not a "true" sensitivity because of that 13 / 63 background base value. These issues seem confusing because in the back of our mind we think that outcomes should be determinate. But after splitting up the subgroups it appears that the issue is now fully determined.

```
NominalCorrelation[Part[res24a, {2, 3}, {2, 3, 4, 5}]] // N
```

1.

The value of decomposition

The value of looking at the 2 × 2 case in this manner is that (1) it restores the idea of deterministic causality, (2) it helps us focussing on the factors that cause the subgroups. What makes that a person falls in the group of diseased who are insensitive, or of the healthy who are sensitive ? Is there a relationship with actually getting the disease ? Of course, as good doctors or researchers we were already aware of these angles but the 2 × 2 × 2 transformation helps to focus.

We said for the 2 × 2 case: that a fully positive matrix didn't contain a zero, therefor refuted causality, and required a third variable, that might be a confounder. Looking at the 2 × 2 × 2 decomposition we find that a hidden variable could be found in some subgroup phenomenon. That subgroup phenomenon may be caused by bias in our study design or by Nature allocating people to such subgroups.

When we added a third variable above it generated some diagonal submatrix. However, adding a third variable does not have to do that necessarily. A third variable might also generate another matrix with fully positive numbers - so that we might need to introduce a fourth variable. Here we get a clear example of confounding.

The Yule-Simpson paradox

The Yule-Simpson paradox is that something may hold for subgroups but not for the total group. A decision based upon subgroups would turn around if one would consider only the total. The reason for such a switch can lie in confounding. The following example is taken from Kleinbaum et al. (2003:277). The story is that there are two shops selling both blue and green hats while a customer visits both shops and tries all hats. In each separate shop the green hats fit relatively better, but for both shops combined the blue hats fit relatively better. The dispersion over the two shops is a confounder.

```
TableForm[mat = {{{5, 1}, {8, 2}}, {{2, 8}, {1, 5}}},
  TableHeadings → {{"Shop1", "Shop2"}, {"Green", "Blue"}, {"Fit", "No fit"}}]
```

		Green	Blue
Shop1	Fit	5	8
	No fit	1	2
Shop2	Fit	2	1
	No fit	8	5

When we consider the overall association or correlation between these data then we find a relatively high coefficient of correlation, 64.9%.

```
NominalCorrelation[mat] // N
```

```
0.648564
```

Above contingency table can be collapsed when we join the two shops:

```
matsum = Plus @@ mat
```

$$\begin{pmatrix} 7 & 9 \\ 9 & 7 \end{pmatrix}$$

We can use the odds to express whether Green hats or Blue hats fit better. For example in Shop1 the Green hats have a (Fit / No fit) odds of 5 to 1 and the Blue hats have a 8 to 2 odds (or 4 to 1). Hence the odds ratio is (5 / 1) / (4 / 1) = 5 / 4. We find two results: (i) for the two shops separately the odds ratio is larger than 1 (preferring Green) but (ii) for the total (7/9 versus 9/7) it is below 1 (preferring Blue).

```
OddsRatio /@ Append[mat, matsum]
```

$$\left\{\frac{5}{4}, \frac{5}{4}, \frac{49}{81}\right\}$$

When the tables only concern the problem of fitting hats then it makes sense to add the two tables, since this eliminates the confounder, i.e. the dispersion over the shops. Then we find a much smaller association.

```
NominalCorrelation[matsum ] // N
```

0.125

NB. If we had worked the other way around, started with the table on green and blue hats, discovered that it had only positive entries, and tried to decompose it into diagonal submatrices in order to determine the true causal relations, then our effort to decompose over shops showed that this didn't help. It isn't the case that one shop sells fitting hats and the other one non-fitting hats. So there still is some hidden causality here.

NB. These data tables might represent another kind of problem, alternatively, in which it is not sensible to merely add the tables over the shop dimension. For example in voting theory there can be different districts (Saari (2001)). Or if a bill must pass both Houses in Parliament (Congress and Senate) then we might accept that it is OK that it passes in both while it would not pass in a joint session. For example, we might do a meta-analysis on the findings of the separate shops, aggregating the problem in such a way that the overall direction reflects the individual ones.

NB. To base a decision on a coefficient of correlation, one might specify a confidence interval, as can be done for real data and ordinary least squares regression, see Colignatus (2006). Such an interval has not yet been developed for this measure for nominal data. However, commonly one will regard 0.125 as a low correlation and 0.649 as a higher correlation, and one may take this as indicative.

Interaction of variables

Hier ?

Moving from nominal to ratio scale

Measurement scales

It makes a difference "how wet" the grass is, just a tiny bit or a lot. Logic and the contingency table just distinguish between "wet" and "dry" but a baby may wet the grass noticably less than a good rain. With the model $\{rain, baby\} \rightarrow wet\ grass$, or $\{r, b\} \rightarrow w$, a person interested in quantitative and not just logical relations would write $w \approx x_{\text{sum}} = x_r + x_b + \varepsilon$ with x_c the amount of wetness from cause c, and ε an error term from left out variables. In this model the original focus was on the x's but there is also that ε to consider. It so happens that statisticians have created intricate models discussing such ε, its distribution, expectation and variance.

Causality versus statistics - dominance versus balance

In this manner, history shows that there developed a literature of "path analysis", "graphical models", and "structural equations models" (SEM). In this line of research the emphasis was on causality and the x's. In parallel fashion, there was the development of a "statistical approach" with a focus on the ε. Due to confusion amongst researchers and also due to the apparent rigour of statistical theory, that second line of research started to dominate research. Yet, there is a recent effort to better balance the two approaches.

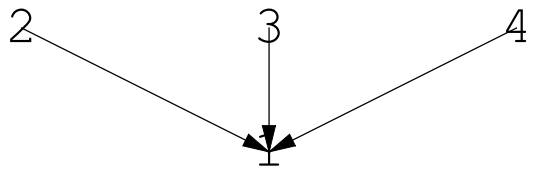
According to Pearl, the confounding as shown in the Simpson paradox is the core case where causality and the statistical approach differ. Pearl (2000:199): "As much as I admire the rigor introduced by Greenland and Robin's analysis through the framework of exchangeability, I am thoroughly convinced that the opacity and inflexibility of counterfactual contingency tables are largely responsible for the slow acceptance of the GR framework among epidemiologists and, as a byproduct, for the lingering confusion that surrounds confounding in the statistical literature at large. I am likewise convinced that formulating claims and assumptions in the language of structural models will make

the mathematical analysis of causation accessible to rank-and-file researchers and thus lead eventually to a total and natural disconfounding of confounding." One would also look at Greenland, Pearl & Robins (1999), "Causal diagrams for epidemiologic research" . A first conclusion is to be careful with the term "structural equations" as it is used in the literature.

Arrow diagrams and logic

In the SEM framework the researcher not only has a model but also an arrow diagram of the relationships. The following is a graph of the three causes $\{r, b, u\}$ that can wet the grass. For example, r = True iff $x_r \geq 0$. In the SEM framework the graph expresses the relationship $w \approx x_{\text{sum}} = x_r + x_b + \varepsilon$, with $\varepsilon = x_u$ from unknown causes u.

```
GraphPlot[{2 → 1, 3 → 1, 4 → 1},
    VertexCoordinates → {{−2, 1}, {0, 0}, {0, 1}, {2, 1}},
    VertexStyleFunction → Automatic, TextStyle → {FontSize → 20},
    EdgeStyleFunction → (Arrow[{#1, #2}] &)];
```



The use of these graphs has distracted the attention of researchers from mere logic. When rain or a baby may wet the grass, or an unknown error disturbance, then logically $w \Leftrightarrow (r \vee b \vee u)$ and we may even translate this into an algebraic relation (for values 1 or 0).

```
ToAlgebra[Equivalent[w, r || b || u]]
```

$$((1-b)(1-r)(1-u)+(1-(1-b)(1-r)(1-u))\,w)\,((1-(1-b)(1-r)(1-u))\,w-w+1)=1$$

These points then may be noted:

1. The latter expression looks rather forbidding and thus we can imagine that people rather work with graphs.
2. The key point: whenever we move from the nominal realm to the ratio scale realm, we never actually *leave* the logical realm, but we just *extend the scope* of the analysis. Apart from the quantitative variables in the model there still is the logical structure, that doesn't change.

Ergo. A model consists of (a) the logical or causal structure, expressed in logic or its algebraic translation or graphs, and (b) the manner of dependency, expressed in equations (e.g. linear or nonlinear).

Thus an empirical model consists of both types of equations in parallel, and the discussion of causality versus statistics is that causality focusses on the structure while statistics focusses on the manner of relationships.

A note on probability

A point to keep in mind is that causality itself would be deterministic and boolean while quantities might be probabilistic. ALOE p 177 contains the triangle of Determinism, Volition and Randomness (chance). These are different perspectives only and their differences are not as dramatic as may seem. For a die the list of possible outcomes {1, …, 6} is deterministic. Volition is when one takes an element by free will (or the illusion of that). If an element is taken by Nature then you might call it deterministic or chance, depending whether you assume some hidden volition/cause/process or not. The big words of Determinism, Volition and Chance thus are mainly perspectives that can be caught in simple operations. Which would transfer to the mathematics of modeling. But it would remain strange to try to describe causality in terms of probability, i.e. you can represent it as such but not explain it as such.

A note on convergence in science

Pearl (2000:29) gives a model in terms of logical relations where some of the variables are random disturbances. I had just completed ALOE (1981, 2007) when Richard Gill informed me about the existence of Pearl's great book. The title of ALOE is "A logic of exceptions" yet I managed not to mention a model in terms of logical relations where some of the variables are random disturbances. I simply expressed that exceptions can pop up but didn't specify them in a logical expression yet. Pearl's seemingly small but decisive step has been used now above, in the relation $w \Leftrightarrow (r \vee b \vee u)$, with u a random disturbance. Also, my background is in econometrics and structural equations modeling, but I also came to work with epidemiologists, and noticed a difference in approach that I could not put a finger on. I am very grateful to Pearl for providing the illumination. This is not to say that epidemiologists would not be interested in causality. Part of the problem may derive from the difference between nominal and ratio scale data. SEM resides in the paradise of real data while epidemiology has an important fraction of activity in nominal data. On the other hand epidemiologists are more used to randomized controlled trials while such a thing is less useful for a national economy. It seems that we now better grasp how these aspects are related.

A note on arrow diagrams

Arrow diagrams are not by themselves 'graphical models', since the latter are theory as discussed e.g. by Pearl (2000). It is also necessary that an arrow in an arrow diagram can stand for almost anything. One always must check what the researcher intends, just like looking at the axes of a plot to see what the plot means. For example, in the arrow diagram above the arrows coming into node 1 express the $\vee$ relation. In other cases such arrows, e.g. from a father and a mother to a child, can stand for an $\wedge$ relation, since it takes those two to transfer the genes.

The Klein (1950) Model I of the USA economy 1920-1941

Lawrence Klein at the Cowles Commission

Lawrence Klein (1950) presented the "Klein Model I" , of the USA economy 1920 - 1941. We shall also use the discussion in Theil (1971:432). In his book Klein thanks Kenneth Arrow, Tjalling Koopmans, Trygve Haavelmo and Don Patinkin, amongst others, for their comments. Klein, Arrow, Koopmans and Haavelmo all got Nobel Prizes later on, in separate years. Patinkin should actually have got one too, for his analysis on Keynes's model. Theil too e.g. for the 2SLS method. See Bernstein (...) for a discussion of the Cowles Commission. So let us hope that the following is a good model of the US economy 1920-1941.

Model

We represent an endogenous variable as *y,* an exogenous variable as *x*[*t*] and a lagged variable as *z*[*t* - *i*]. We indicate what the coefficients are. We can also distinguish behavioural equations with an error term (to be estimated) and identities. Thus we can write routines to fully decompose a model in its structural elements.

KleinI["Legend"]

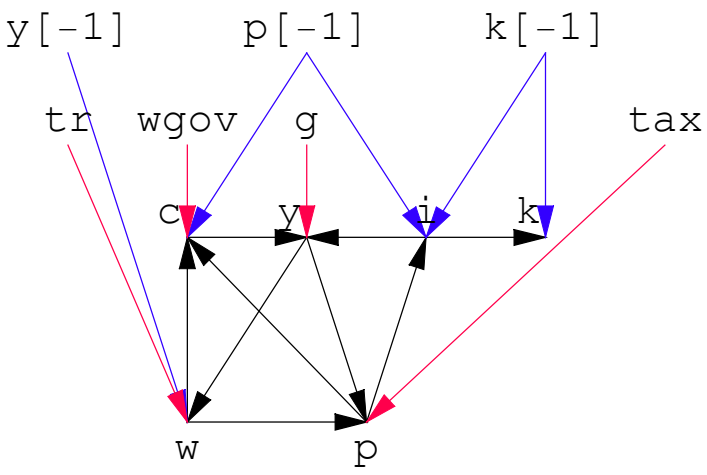
$$\begin{pmatrix} \text{c = consumption} \\ \text{i = investment} \\ \text{g = goverment} \\ \text{y = output = c + i + g} \\ \text{wgov = govt. wages} \\ \text{w = wages} \\ \text{tax = taxes} \\ \text{p = profit = y − w − tax} \\ \text{k = capital = k[−1] + i} \\ \text{tr = trend = annum − 1931} \end{pmatrix}$$

```
KleinI["Equations"] // MatrixForm
```

$$\begin{pmatrix} \text{Equation} \to c = \text{c0} + \text{ceps}(t) + \text{cp1}\ p(t-1) + \text{cp}\ p(t) + \text{cw}\,(w + \text{wgov}(t)) & \text{Coefficients} \to \{\text{c0, cp, c} \\ \text{Equation} \to i = \text{i0} + \text{ieps}(t) + \text{ik1}\ k(t-1) + \text{ip1}\ p(t-1) + \text{ip}\ p(t) & \text{Coefficients} \to \{\text{i0, ip, ip} \\ \text{Equation} \to w = \text{w0} + (t-1931)\,\text{wtr} + \text{weps}(t) + \text{wy1}\ y(t-1) + \text{wy}\ y(t) & \text{Coefficients} \to \{\text{w0, wy,} \\ \text{Equation} \to y = c + i + g(t) & \text{Coefficients} \to \{\} \\ \text{Equation} \to p = -w + y - \text{tax}(t) & \text{Coefficients} \to \{\} \\ \text{Equation} \to k = i + k(t-1) & \text{Coefficients} \to \{\} \end{pmatrix}$$

Endogenes are in black, exogenes in red and lagged variables in blue.

```
KleinI["ArrowDiagram"];
```



Estimating

We should use 2SLS but for ease allow for OLS that generates biased non-consistent parameter estimates.

```
toEstimate = KleinI["Equation"] /@ {c, i, w};
```

```
(eqs = (Equation /. toEstimate)) // MatrixForm
```

$$\begin{pmatrix} c = \text{c0} + \text{ceps}(t) + \text{cp1}\ p(t-1) + \text{cp}\ p(t) + \text{cw}\,(w + \text{wgov}(t)) \\ i = \text{i0} + \text{ieps}(t) + \text{ik1}\ k(t-1) + \text{ip1}\ p(t-1) + \text{ip}\ p(t) \\ w = \text{w0} + (t-1931)\,\text{wtr} + \text{weps}(t) + \text{wy1}\ y(t-1) + \text{wy}\ y(t) \end{pmatrix}$$

```
(neqs = eqs /. x_[t - 1] :> ToExpression[ToString[x] <> "lag"] /. y_[t] :> y /.
      Thread[(Error /. toEstimate) → 0]) // MatrixForm
```

$$\begin{pmatrix} c = \text{c0} + \text{cp}\ p + \text{cp1 plag} + \text{cw}\,(w + \text{wgov}) \\ i = \text{i0} + \text{ik1 klag} + \text{ip}\ p + \text{ip1 plag} \\ w = \text{w0} + (t-1931)\,\text{wtr} + \text{wy}\ y + \text{wy1 ylag} \end{pmatrix}$$

```
coefs = Coefficients /. toEstimate // Flatten
```

{c0, cp, cp1, cw, i0, ip, ip1, ik1, w0, wy, wy1, wtr}

```
vars = Variables[(neqs /. Equal → List /. Thread[coefs → 1]) // Flatten]
```

{*c*, *i*, klag, *p*, plag, *t*, *w*, wgov, *y*, ylag}

```
CopyData[KleinI, vars]
```

{plag, *t*, ylag}

```
Data[t] = KleinI[annum];
Data[plag] = Lag[Data[p]];
Data[ylag] = Lag[Data[y]];
```

```
dataUsed = ToDataRule[vars] /. (x_ → y_List) :→ (x → Rest[y]);
```

```
res = Estimate[neqs, dataUsed, coefs]
```

$$\Big\{\text{AdjustedRSquared} \to 0.998296,$$

$$\text{BestFitParameters} \to \{\text{c0} \to 16.2366,\ \text{cp} \to 0.192934,\ \text{cp1} \to 0.0898849,\ \text{cw} \to 0.796219,\ \text{i0} \to 10.1258,\ \text{ip} \to 0.479636,\ \text{ip1} \to 0.333039,\ \text{ik1} \to -0.111795,\ \text{w0} \to 1.49704,\ \text{wy} \to 0.439477,\ \text{wy1} \to 0.14609,\ \text{wtr} \to 0.130245\},$$

$$\text{Correlation} \to 0.999299,\ \text{CovarianceMatrix} \to \begin{pmatrix} 1.43027 & 0.000422575 & -0.0149235 & - \\ 0.000422575 & 0.00701158 & -0.00444197 & - \\ -0.0149235 & -0.00444197 & 0.00692541 & - \\ -0.0277429 & -0.00111157 & -0.000565591 & 0 \\ 0. & 0. & 0. & 0 \\ 0. & 0. & 0. & 0 \\ 0. & 0. & 0. & 0 \\ 0. & 0. & 0. & 0 \\ 0. & 0. & 0. & 0 \\ 0. & 0. & 0. & 0 \\ 0. & 0. & 0. & 0 \\ 0. & 0. & 0. & 0 \end{pmatrix}$$

$$\text{EstimatedVariance} \to 0.88641,\ \text{NumberOfEquations} \to 3,$$

$$\text{NumberOfObservations} \to 63,\ \text{ReducedFormQ} \to \text{False},\ \text{RSquared} \to 0.998598,$$

$$\text{StandardDeviation} \to \{1.19594, 0.0837352, 0.083219, 0.0366704, 5.09762, 0.0905771, 0.0940697, 0.0249283, 1.55867, 0.0397727, 0.0459281, 0.0391624\},$$

$$\text{TValues} \to \{13.5765, 2.3041, 1.0801, 21.7129, 1.98638, 5.29533, 3.54034, -4.48464, 0.960464, 11.0497, 3.18084, 3.32577\}\Big\}$$

```
coefsEst = (BestFitParameters /. res)
```

```
{c0 → 16.2366, cp → 0.192934, cp1 → 0.0898849, cw → 0.796219,
  i0 → 10.1258, ip → 0.479636, ip1 → 0.333039, ik1 → −0.111795,
  w0 → 1.49704, wy → 0.439477, wy1 → 0.14609, wtr → 0.130245}
```

Model run

```
SetOptions[Model, Coefficients → coefsEst];
```

```
errors = FitResiduals /. Results[Estimate];
KleinI[ceps] = Prepend[Error[1] /. errors, Indeterminate];
KleinI[ieps] = Prepend[Error[2] /. errors, Indeterminate];
KleinI[weps] = Prepend[Error[3] /. errors, Indeterminate];
```

```
(eqs = (Equation /. KleinI["Equations"])) // MatrixForm
```

$$\begin{pmatrix} c = \text{c0} + \text{ceps}(t) + \text{cp1}\ p(t-1) + \text{cp}\ p(t) + \text{cw}\ (w + \text{wgov}(t)) \\ i = \text{i0} + \text{ieps}(t) + \text{ik1}\ k(t-1) + \text{ip1}\ p(t-1) + \text{ip}\ p(t) \\ w = \text{w0} + (t - 1931)\ \text{wtr} + \text{weps}(t) + \text{wy1}\ y(t-1) + \text{wy}\ y(t) \\ y = c + i + g(t) \\ p = -w + y - \text{tax}(t) \\ k = i + k(t-1) \end{pmatrix}$$

```
SetOptions[Model, Equations → eqs];
```

```
SetOptions[Model, Begin → 1920];
```

```
Model[]
```

```
Results[Model]
```

```
Equations(t_) := {c = c0 + ceps(t) + cp1 p(t − 1) + cp p(t) + cw (w + wgov(t)),
    i = i0 + ieps(t) + ik1 k(t − 1) + ip1 p(t − 1) + ip p(t),
    w = w0 + (t − 1931) wtr + weps(t) + wy1 y(t − 1) + wy y(t), y = c + i + g(t),
    p = −w + y − tax(t), k = i + k(t − 1)} /. (Coefficients /. Options[Model])
```

```
{$Endogenes → {c, i, k, p, w, y}, $Exogenes → {ceps, g, ieps, tax, weps, wgov},
  Begin → 1920, TimeZone → {1921, Min(EndYear(ceps), EndYear(g),
      EndYear(ieps), EndYear(tax), EndYear(weps), EndYear(wgov))}}
```

```
vars = Endogenes ~Join~ Exogenes
```

```
{c, i, k, p, w, y, ceps, g, ieps, tax, weps, wgov}
```

```
CopyData[KleinI, vars]
```

```
{k}
```

```
Data[k] = Lead[Kleinl[klag]];
```

```
dataUsed = ToDataRule[vars] ;
```

```
Model[SetData → dataUsed]
```

```
First /@ eqs
```

$\{c, i, w, y, p, k\}$

```
nsp = NSolvePeriod[1921, 1941]
```

$$\begin{pmatrix}
41.9 & -0.2 & 182.6 & 12.4 & 25.5 & 45.6 \\
45. & 1.9 & 184.5 & 16.9 & 29.3 & 50.1 \\
49.2 & 5.2 & 189.7 & 18.4 & 34.1 & 57.2 \\
50.6 & 3. & 192.7 & 19.4 & 33.9 & 57.1 \\
52.6 & 5.1 & 197.8 & 20.1 & 35.4 & 61. \\
55.1 & 5.6 & 203.4 & 19.6 & 37.4 & 64. \\
56.2 & 4.2 & 207.6 & 19.8 & 37.9 & 64.4 \\
57.3 & 3. & 210.6 & 21.1 & 39.2 & 64.5 \\
57.8 & 5.1 & 215.7 & 21.7 & 41.3 & 67. \\
55. & 1. & 216.7 & 15.6 & 37.9 & 61.2 \\
50.9 & -3.4 & 213.3 & 11.4 & 34.5 & 53.4 \\
45.6 & -6.2 & 207.1 & 7. & 29. & 44.3 \\
46.5 & -5.1 & 202. & 11.2 & 28.5 & 45.1 \\
48.7 & -3. & 199. & 12.3 & 30.6 & 49.7 \\
51.3 & -1.3 & 197.7 & 14. & 33.2 & 54.4 \\
57.7 & 2.1 & 199.8 & 17.6 & 36.8 & 62.7 \\
58.7 & 2. & 201.8 & 17.3 & 41. & 65. \\
57.5 & -1.9 & 199.9 & 15.3 & 38.2 & 60.9 \\
61.6 & 1.3 & 201.2 & 19. & 41.6 & 69.5 \\
65. & 3.3 & 204.5 & 21.1 & 45. & 75.7 \\
69.7 & 4.9 & 209.4 & 23.5 & 53.3 & 88.4
\end{pmatrix}$$

```
Store[run1]
```

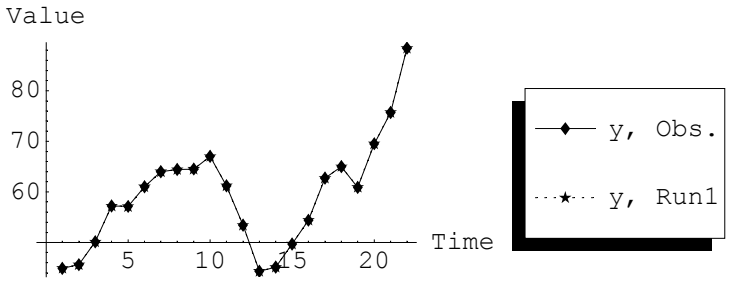
```
res = {Data[y], Data[y, run1]}
```

$$\begin{pmatrix}
44.9 & 45.6 & 50.1 & 57.2 & 57.1 & 61. & 64. & 64.4 & 64.5 & 67. & 61.2 & 53.4 & 44.3 & 45.1 & 49.7 & 54 \\
44.9 & 45.6 & 50.1 & 57.2 & 57.1 & 61. & 64. & 64.4 & 64.5 & 67. & 61.2 & 53.4 & 44.3 & 45.1 & 49.7 & 54
\end{pmatrix}$$

```
Data[y] == Data[y, run1]
```

True

```
MultipleListPlot[Sequence @@ DataFilter @@ res,
    PlotJoined → True, AxesLabel → {"Time", "Value"},
    PlotLegend → {"y, Obs.", "y, Run1"}, LegendShadow → {−0.05, −0.05}];
```




```
Equations[1941] // MatrixForm
```

$$\begin{pmatrix} c = 0.796219\,(w + 8.5) + 20.4937 \\ i = 4.9 \\ w = 53.3 \\ y = c + i + 13.8 \\ p = -w + y - 11.6 \\ k = i + 204.5 \end{pmatrix}$$

```
Data /@ {g, tax}
```

$$\begin{pmatrix} 2.4 & 3.9 & 3.2 & 2.8 & 3.5 & 3.3 & 3.3 & 4. & 4.2 & 4.1 & 5.2 & 5.9 & 4.9 & 3.7 & 4. & 4.4 & 2.9 & 4.3 \\ 3.4 & 7.7 & 3.9 & 4.7 & 3.8 & 5.5 & 7. & 6.7 & 4.2 & 4. & 7.7 & 7.5 & 8.3 & 5.4 & 6.8 & 7.2 & 8.3 & 6.7 \end{pmatrix}$$

```
impulse = Table[If[t ≤ 1930, 0, 5], {t, 1920, 1941}]
```

```
{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5}
```

```
SetData[{g → Data[g] + impulse, tax → Data[tax] + impulse}, 1920, run2]
```

nsp2 = NSolvePeriod[1921, 1941]

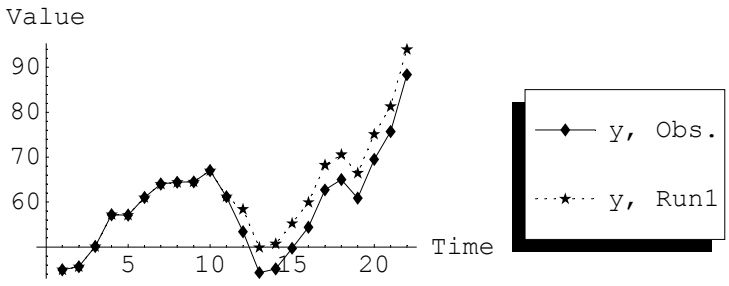
41.9	−0.2	182.6	12.4	25.5	45.6
45.	1.9	184.5	16.9	29.3	50.1
49.2	5.2	189.7	18.4	34.1	57.2
50.6	3.	192.7	19.4	33.9	57.1
52.6	5.1	197.8	20.1	35.4	61.
55.1	5.6	203.4	19.6	37.4	64.
56.2	4.2	207.6	19.8	37.9	64.4
57.3	3.	210.6	21.1	39.2	64.5
57.8	5.1	215.7	21.7	41.3	67.
55.	1.	216.7	15.6	37.9	61.2
50.9	−3.4	213.3	11.4	34.5	58.4
46.1816	−6.2	207.1	6.85115	29.7304	49.8816
47.1359	−5.14957	201.95	10.9709	29.3154	50.6863
49.3292	−3.07076	198.88	12.0423	31.4161	55.2584
51.9234	−1.37236	197.507	13.739	34.012	59.951
58.3222	2.03462	199.542	17.3459	37.611	68.2569
59.3235	1.94423	201.486	17.056	41.8118	70.5678
58.1257	−1.94619	199.54	15.0661	39.0134	66.4795
62.228	1.26236	200.802	18.7752	42.4151	75.0903
65.6301	3.2696	204.072	20.883	45.8167	81.2997
70.3318	4.87558	208.948	23.2894	54.1181	94.0074

Store[run2]

res = {Data[y, run1], Data[y, run2]}

44.9	45.6	50.1	57.2	57.1	61.	64.	64.4	64.5	67.	61.2	53.4	44.3	45.1	4
44.9	45.6	50.1	57.2	57.1	61.	64.	64.4	64.5	67.	61.2	58.4	49.8816	50.6863	5

MultipleListPlot[Sequence @@ DataFilter @@ res,
PlotJoined → True, AxesLabel → {"Time", "Value"},
PlotLegend → {"y, Obs.", "y, Run1"}, LegendShadow → {−0.05, −0.05}];



Relation to causality

1. The structural relations, our theory, specify how a variable is caused and explained by the other variables.
2. The specification explicitly determines zero coefficients for variables. These imply no arrows in the diagram.
3. If we hadn't those restrictions then all equations would look alike and we could not distinguish them.
4. In "path analysis" we would test those zero coefficients.
5. A lot of attention goes to estimating proper coefficients (e.g. OLS versus 2SLS). Yet a key issue is to keep a variable in the equation as well (testing on significant difference from zero).
6. The interpretation of such coefficients is also causal, which differs from the statistical approach (Pearl (2000:164)).
7. Simulating "what if" scenario's reflect "counterfactuals", which can only be properly understood in terms of causality.

Topics

Construction of contingency data from questionnaires (already in Guide)

Probability theory and the notation in *Mathematica* (notably conditionality) (already in Guide)

Epidemiology (already in Guide)

CrossTable and Chi2 (take from paper, and already in Guide)

Dump