0.1

Scales of measurement

As the previous section indicates, the outcome of a psychological measurement is called a variable. But not all variables are of the same qualitative type and so it's useful to understand what types there are. A very useful concept for distinguishing between different types of variables is what's known as **scales of measurement**.

0.1.1 Nominal scale

A **nominal scale** variable (also referred to as a **categorical** variable) is one in which there is no particular relationship between the different possibilities. For these kinds of variables it doesn't make any sense to say that one of them is "bigger' or "better" than any other one, and it absolutely doesn't make any sense to average them. The classic example for this is "eye colour". Eyes can be blue, green or brown, amongst other possibilities, but none of them is any "bigger" than any other one. As a result, it would feel really weird to talk about an "average eye colour". Similarly, gender is nominal too: male isn't better or worse than female. Neither does it make sense to try to talk about an "average gender". In short, nominal scale variables are those for which the only thing you can say about the different possibilities is that they are different. That's it.

Let's take a slightly closer look at this. Suppose I was doing research on how people commute to and from work. One variable I would have to measure would be what kind of transportation people use to get to work. This "transport type" variable could have quite a few possible values, including: "train", "bus", "car", "bicycle". For now, let's suppose that these four are the only possibilities. Then imagine that I ask 100 people how they got to work today, with this result:

Transportation	Number of people
(1) Train	12
(2) Bus	30
(3) Car	48
(4) Bicycle	10

So, what's the average transportation type? Obviously, the answer here is that there isn't one. It's a silly question to ask. You can say that travel by car is the most popular method, and travel by train is the least popular method, but that's about all. Similarly, notice that the order in which I

list the options isn't very interesting. I could have chosen to display the data like this...

Transportation	Number of people
(3) Car	48
(1) Train	12
(4) Bicycle	10
(2) Bus	30

...and nothing really changes.

0.1.2 Ordinal scale

Ordinal scale variables have a bit more structure than nominal scale variables, but not by a lot. An ordinal scale variable is one in which there is a natural, meaningful way to order the different possibilities, but you can't do anything else. The usual example given of an ordinal variable is "finishing position in a race". You *can* say that the person who finished first was faster than the person who finished second, but you *don't* know how much faster. As a consequence we know that 1st $>$ 2nd, and we know that 2nd $>$ 3rd, but the difference between 1st and 2nd might be much larger than the difference between 2nd and 3rd.

Here's a more psychologically interesting example. Suppose I'm interested in people's attitudes to climate change. I then go and ask some people to pick the statement (from four listed statements) that most closely matches their beliefs:

(1) Temperatures are rising because of human activity
(2) Temperatures are rising but we don't know why
(3) Temperatures are rising but not because of humans
(4) Temperatures are not rising

Notice that these four statements actually do have a natural ordering, in terms of "the extent to which they agree with the current science". Statement 1 is a close match, statement 2 is a reasonable match, statement 3 isn't a very good match, and statement 4 is in strong opposition to current science. So, in terms of the thing I'm interested in (the extent to which people endorse the science), I can order the items as $1 > 2 > 3 > 4$. Since this ordering exists, it would be very weird to list the options like this...

(3) Temperatures are rising but not because of humans
(1) Temperatures are rising because of human activity

(4) Temperatures are not rising

(2) Temperatures are rising but we don't know why

...because it seems to violate the natural "structure" to the question.

So, let's suppose I asked 100 people these questions, and got the following answers:

Response	Number
(1) Temperatures are rising because of human activity	51
(2) Temperatures are rising but we don't know why	20
(3) Temperatures are rising but not because of humans	10
(4) Temperatures are not rising	19

When analysing these data it seems quite reasonable to try to group (1), (2) and (3) together, and say that 81 out of 100 people were willing to *at least partially* endorse the science. And it's *also* quite reasonable to group (2), (3) and (4) together and say that 49 out of 100 people registered *at least some disagreement* with the dominant scientific view. However, it would be entirely bizarre to try to group (1), (2) and (4) together and say that 90 out of 100 people said...what? There's nothing sensible that allows you to group those responses together at all.

That said, notice that while we *can* use the natural ordering of these items to construct sensible groupings, what we *can't* do is average them. For instance, in my simple example here, the "average" response to the question is 1.97. If you can tell me what that means I'd love to know, because it seems like gibberish to me!

0.1.3 Interval scale

In contrast to nominal and ordinal scale variables, **interval scale** and ratio scale variables are variables for which the numerical value is genuinely meaningful. In the case of interval scale variables the *differences* between the numbers are interpretable, but the variable doesn't have a "natural" zero value. A good example of an interval scale variable is measuring temperature in degrees celsius. For instance, if it was 15° yesterday and 18° today, then the 3° difference between the two is genuinely meaningful. Moreover, that 3° difference is *exactly the same* as the 3° difference between 7° and 10°. In short, addition and subtraction are meaningful for interval scale variables.[*1]

[*1] Actually, I've been informed by readers with greater physics knowledge than I that temperature isn't strictly an interval scale, in the sense that the amount of energy required to heat something up by 3° depends on it's current temperature. So in the sense that physicists care about, temperature isn't actually an interval scale. But it still makes a cute example so I'm going to ignore this little inconvenient truth.

However, notice that the 0° does not mean "no temperature at all". It actually means "the temperature at which water freezes", which is pretty arbitrary. As a consequence it becomes pointless to try to multiply and divide temperatures. It is wrong to say that 20° is *twice as hot* as 10°, just as it is weird and meaningless to try to claim that 20° is negative two times as hot as -10°.

Again, lets look at a more psychological example. Suppose I'm interested in looking at how the attitudes of first-year university students have changed over time. Obviously, I'm going to want to record the year in which each student started. This is an interval scale variable. A student who started in 2003 did arrive 5 years before a student who started in 2008. However, it would be completely daft for me to divide 2008 by 2003 and say that the second student started "1.0024 times later" than the first one. That doesn't make any sense at all.

0.1.4 Ratio scale

The fourth and final type of variable to consider is a **ratio scale** variable, in which zero really means zero, and it's okay to multiply and divide. A good psychological example of a ratio scale variable is response time (RT). In a lot of tasks it's very common to record the amount of time somebody takes to solve a problem or answer a question, because it's an indicator of how difficult the task is. Suppose that Alan takes 2.3 seconds to respond to a question, whereas Ben takes 3.1 seconds. As with an interval scale variable, addition and subtraction are both meaningful here. Ben really did take $3.1 - 2.3 = 0.8$ seconds longer than Alan did. However, notice that multiplication and division also make sense here too: Ben took $3.1/2.3 = 1.35$ times as long as Alan did to answer the question. And the reason why you can do this is that for a ratio scale variable such as RT "zero seconds" really does mean "no time at all".

0.1.5 Continuous versus discrete variables

There's a second kind of distinction that you need to be aware of, regarding what types of variables you can run into. This is the distinction between continuous variables and discrete variables. The difference between these is as follows:

- A **continuous variable** is one in which, for any two values that you can think of, it's always logically possible to have another value in between.
- A **discrete variable** is, in effect, a variable that isn't continuous. For a discrete variable it's sometimes the case that there's nothing in the middle.

These definitions probably seem a bit abstract, but they're pretty simple once you see some examples. For instance, response time is continuous. If Alan takes 3.1 seconds and Ben takes 2.3 seconds to respond to a question, then Cameron's response time will lie in between if he took

Table1 The relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible.

	continuous	discrete
nominal		✓
ordinal		✓
interval	✓	✓
ratio	✓	✓

3.0 seconds. And of course it would also be possible for David to take 3.031 seconds to respond, meaning that his RT would lie in between Cameron's and Alan's. And while in practice it might be impossible to measure RT that precisely, it's certainly possible in principle. Because we can always find a new value for RT in between any two other ones we regard RT as a continuous measure.

Discrete variables occur when this rule is violated. For example, nominal scale variables are always discrete. There isn't a type of transportation that falls "in between" trains and bicycles, not in the strict mathematical way that 2.3 falls in between 2 and 3. So transportation type is discrete. Similarly, ordinal scale variables are always discrete. Although "2nd place" does fall between "1st place" and "3rd place", there's nothing that can logically fall in between "1st place" and "2nd place". Interval scale and ratio scale variables can go either way. As we saw above, response time (a ratio scale variable) is continuous. Temperature in degrees celsius (an interval scale variable) is also continuous. However, the year you went to school (an interval scale variable) is discrete. There's no year in between 2002 and 2003. The number of questions you get right on a true-or-false test (a ratio scale variable) is also discrete. Since a true-or-false question doesn't allow you to be "partially correct", there's nothing in between 5/10 and 6/10. Table **??** summarises the relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible. I'm trying to hammer this point home, because (a) some textbooks get this wrong, and (b) people very often say things like "discrete variable" when they mean "nominal scale variable". It's very unfortunate.

0.1.6 Some complexities

Okay, I know you're going to be shocked to hear this, but the real world is much messier than this little classification scheme suggests. Very few variables in real life actually fall into these nice neat categories, so you need to be kind of careful not to treat the scales of measurement as if they were

hard and fast rules. It doesn't work like that. They're guidelines, intended to help you think about the situations in which you should treat different variables differently. Nothing more.

So let's take a classic example, maybe *the* classic example, of a psychological measurement tool: the **Likert scale**. The humble Likert scale is the bread and butter tool of all survey design. You yourself have filled out hundreds, maybe thousands, of them and odds are you've even used one yourself. Suppose we have a survey question that looks like this:

> Which of the following best describes your opinion of the statement that "all pirates are freaking awesome"?

and then the options presented to the participant are these:

> (1) Strongly disagree
> (2) Disagree
> (3) Neither agree nor disagree
> (4) Agree
> (5) Strongly agree

This set of items is an example of a 5-point Likert scale, in which people are asked to choose among one of several (in this case 5) clearly ordered possibilities, generally with a verbal descriptor given in each case. However, it's not necessary that all items are explicitly described. This is a perfectly good example of a 5-point Likert scale too:

> (1) Strongly disagree
> (2)
> (3)
> (4)
> (5) Strongly agree

Likert scales are very handy, if somewhat limited, tools. The question is what kind of variable are they? They're obviously discrete, since you can't give a response of 2.5. They're obviously not nominal scale, since the items are ordered; and they're not ratio scale either, since there's no natural zero.

But are they ordinal scale or interval scale? One argument says that we can't really prove that the difference between "strongly agree" and "agree" is of the same size as the difference between "agree" and "neither agree nor disagree". In fact, in everyday life it's pretty obvious that they're not the same at all. So this suggests that we ought to treat Likert scales as ordinal variables. On

the other hand, in practice most participants do seem to take the whole "on a scale from 1 to 5" part fairly seriously, and they tend to act as if the differences between the five response options were fairly similar to one another. As a consequence, a lot of researchers treat Likert scale data as interval scale.[*2] It's not interval scale, but in practice it's close enough that we usually think of it as being **quasi-interval scale**.

[*2] Ah, psychology ...never an easy answer to anything!

1. 研究デザインについての短い導入

> *実験が終わった後に統計家にアドバイスを求められることは，ほとんどの場合，死後の検査を求めるようなものである。統計家は実験がなぜ死んだかをいうだけだろう。*
>
> – Sir Ronald Fisher[*1]

この章では，研究をデザインし，データを集め，そのデータが機能するかどうかチェックする，といったことについての基本的なアイデアについて考えることから始めよう。あなた自身が研究をデザインするのに十分な情報生えられないかもしれないが，ほかの人がやる研究を査定するのに必要な基本的ツールの多くは手に入れられると思う。しかし，この本の狙いはデータの収集よりもデータの分析にあるので，ごくあっさりと全体像を見るにとどめる。この章は，ふたつの意味で「特別」である。まず，ほかの章よりも心理学に特有の状況を扱っている。次に，研究の方法論についての科学的問題にかなりの重点を置いていて，データ分析の統計的問題にはそれほど重点をおいていない。とはいえ，このふたつの問題はお互いに関係しているから，この問題をちょっとばかり詳しく扱うのが，統計のテキストの伝統的なやり方になっている。この章は研究デザインについては **Campbell1963** に，測定の尺度に関する議論については **Stevens1946** に依っている。

1.1 心理学的測定への導入

データ収集を理解するにはまず，**測定**について考えなければならない。つまり，ここで我々がやろうとしていることは，人の行動や心についての何かを測るということだ。では「測る」とは何だろう？

[*1] 第 1 回インド統計会議会長講演, 1938. 出典: http://en.wikiquote.org/wiki/Ronald_Fisher

1.1.1 心理学的測定についてのいくつかの考え方

測定そのものは微妙な概念だが，基本的には「なにか」に数字やラベル，あるいはほかのよく定義された記述を割り当てる方法を見つけること，と言える。であるから，次に挙げるようなものも心理学的測定の中に含まれる。

- 私の **年齢** は *33* 歳です。
- 私は**アンチョビが** *好きではない* 。
- 私の **染色体的な性別** は *男性である*。
- 私の **性自認** は *男性である*[*2]

上の短いリストには**太字になっているところ**が “測定されようとしている対象” であり，*イタリックになっているところ*が “測定されたもの” である。じつは，それぞれのケースで生じうる測定全体について考えることで，少し拡張することができる。

- 私の**年齢**は (年単位で数えるから)*0, 1, 2, 3 …*，となる。上限がどうなるかはちょっと曖昧だが，現実的には最大でも *150* ぐらいだと思っていれば良い。それ以上生きた人はいないのだから。
- *アンチョビが好き*かどうかは，答え方として*好き*あるいは*嫌い*，はたまた*どちらでもないとか*，*ときどきね*，と答えるだろう。
- 私の**染色体的な性別**は，ほぼ間違いなく*男性 (XY)* か，*女性 (XX)* だろうが，ごくわずかな例外もある。それは*クラインフェルター症候群 (XXY)* というやつで，ほとんど男性の染色体とおなじである。そのほかにもこうした可能性はあるかもしれない。
- 私の**性自認**は*男性*か*女性*のいずれかだが，私の染色体的な性別とは一致していない。私は自分を*どちらでもないとする*かもしれないし，はっきりと*トランスジェンダー*だ，と答えるかもしれない。

このように，ある事象 (年齢とか) は取りうる値が明らかのものもあれば，ちょっと特殊なものも

[*2] うむ… ちょっと引っかかるだろ？ このセクションはこの本の最も古いパートだから，ちょっと時代遅れで小っ恥ずかしいものになっている。これが書かれているのは 2010 年で，ここに書いてある事実は*確かに正しい*。これを今の 2018 年にも取って考えると，私はもう 33 歳じゃないし，驚くことじゃない。私の染色体が変わったとは思えないし，私の遺伝子は今も XY 型だとおもう。一方，性自認は… ううん。タイトルページには，今私のことを Daniel じゃなくて Danielle と書いているからわかってもらえるかと思うけど (訳註；Daniel は男性名的表記，Danielle は女性名的表記)，最近は性別に関するアンケートでは男性と答えないようにしていて，むしろ “she/her” で呼ばれる方が嬉しくなっている。話すと長くなるけど！ 実は，この本ではこのことをどう扱うか少し考えた。この本では作者の声が直接含まれているので，この本での表記を全部 Danielle にすると随分違ったものになるんじゃないかと思う。でもそれはかなりの作業量になるから，この本での私の名前は “Dan” で通そうと思う。とにかく Danielle のニックネームとして Dan はバッチリだと思わない？ ちっちゃいことは気にしない。私のことをどう呼んだら良いかわからない読者のために，少しでも気楽になってもらえればと思ってこれを書いてる。あ，アンチョビはまだ好きじゃないけどね (笑)。

ある。しかし，誰かの年齢と言ってもこれよりちょっと微妙なものであることは指摘しておきたい。たとえば，上の例では年齢を一年単位で測定することを想定していた。しかしもしあなたが発達心理学者であったら，これはちょっと雑すぎてて，年齢は*年と月 (月齢)* で表現するかもしれない (もし子供が 2 歳 11 ヶ月であれば，これを “2.11” と書いたりする)。新生児に興味があれば，年齢は生まれてから*何日めか*で表現するだろうし，もしかすると*生後何時間経ったか*で表現するかもしれない。言い換えると，あなたが測定値に何を許容するかというのが重要なのだ。

これをもう少し詳しく見てみると，“年齢” という概念は実際まったく正確ではないことに気づかれるかもしれない。一般的に言って，我々が “年齢” というときは，暗に “生まれてからの時間の長さ” を意味している。しかし，それが常に正しいわけではない。たとえば，新生児の目の動きをどのようにコントロールしているかに興味があったとしよう。もしそれぐらい若い子に興味があれば，あなたは “生まれたとき” から始めることが意味のある点だとは思わなくなるかもしれない。アリスが生後 3 週間で，ビアンカが 1 週間後に生まれてきたとして，彼女らに “生まれてから 2 時間後” に出会った時彼女らが “同じ年齢” だったといえるだろうか？ 社会的には，日常生活で年齢について語る時に，生まれた時を起点とするが，それは世界に生まれたって独立した存在として自らを扱ってきた人としての時間の総量，として定義しているからだ。科学的な観点からは，興味があるのはそこだけではない。生物としての人間を考えるなら，概念的には意識を持った生き物として成長してきた過程を考える方が便利だし，その観点から行くとアリスとビアンカは同じ年齢だとは言えない。だからあなたが “年齢” という概念を定義しようとするなら，ふたつのやり方があることになる。意識の時間的長さか，生まれてからの時間的長さか，である。大人を対象にしているときは，その差はほとんど意味をなさないが，新生児の場合はそうではないのだ。

この問題を超えた，方法論的な問いがある。どの “測定法” が，誰かの年齢を見出すのに用いられるべきだろうか。上で述べたように，そこにはいくつかの可能性がある。

- 単に “あなたは何歳ですか” と尋ねる。この事故報告式は早くて，やすくて，簡単だ。しかしその人が質問を十分理解している場合に限るし，人によっては歳を誤魔化すことがある。
- 権威者 (例えば親御さん) に “あなたのお子さんは何歳ですか？” と尋ねる。この方法は早くて，子供を扱うときには簡単な方法だ。親はいつも周りにいるのだから。でも “意識を持ち始めてから” を考えるときはうまくいかない。というのも，親はその子がいつ意識を持ち始めたかを知る由がないからだ。そのときは，別の権威者 (例えば産科医) に頼る必要があるだろう。
- 公的な記録を調べる。例えば出生証明書とか，死亡証明書である。これは時間はかかるし，手間もかかるが，それなりの利点がある (例えば故人を対象にする場合は。)

1.1.2 操作可能にする：あなたの測定を定義する

上の章で論じだれていたことはすべて，**操作化**という概念に関係している。もう少しこの概念を正確にいうと，操作化とは大きな意味があるが，何かしらぼんやりした概念を，正確な測定に落とし込む手続きのことを指す。この操作化のプロセスは，以下のいくつかの異なる要素を含んでいる。

- あなたが測定しようとしているものを明確にすること。例えば “年齢” も “生まれてからの時間” なのか，“意識が芽生えてからの時間” なのか，どちらがあなたの研究の文脈に沿うものなのか？
- それを測定するのにどの手法にするのか決定すること。事故報告式の測定年齢を使うのか，親に聞くのか，公的な記録を取るのか，どれだろう？ もし事故報告式を使うのなら，どういうフレーズでその質問をするだろう？
- 測定が取りうる値のセットを決定すること。この値は常に数量的なものとは限らないことに注意しよう。測定する年齢は数量的かもしれないが，どんな値が許されているのかをよく考えて見てほしい。年単位の年齢なのか，年月なのか，日，時間，どれだろう？ ほかの測定の種類によっては (例えばジェンダー)，値は数量的なものにならない。しかしすでに述べたように，私たちはどんな値が許されているかを考える必要があるのだ。もし事故報告式でジェンダーを尋ねたら，私たちは相手にどんな選択肢を許すだろうか。単に “男性” か “女性” だけで十分なのか？ “その他” のオプションは必要ないのか？ あるいは，特別な選択肢を考えて，回答者自身の言葉を使うことを許すのか？ そしてもしすべての言語報告を可能な値のセットに含めるのであれば，その回答をどのように解釈するのか？

操作化はトリッキーなビジネスで，“唯一の正解” があるものではない。“年齢” や “性別” のインフォーマルな概念を操作化する方法を選択し，それをフォーマルな測定にすることを選んでも，それはあなたが何を測定したいかに依存するのだ。大抵の場合，あなたの領域で活躍している人たちの科学的コミュニティーが，どうしたらいいかについて十分に確立されたアイデアをすでに持っていることと思う。いいかえると，操作化はケースバイケースで考える必要があると言うことだ。とはいえ，個々人の研究プロジェクトにはそれぞれ特別な問題が他にもいっぱいあるが，一般的な側面もあるものだ。

次に進む前に，用語を整理しておきたい。相互に関連する次の四つのことがある。

- **理論的構成概念**。これはあなたが測定しようとしているもの，例えば “年齢”，“性別”，“ジェンダー” や “意見” といったものだ。理論的な構成概念は直接観測されないし，実際にはちょっと曖昧なものであることが多い。
- **測定**。測定はあなたが客観的にしようとする際につかうツールや手法のこと。調査における質問，行動観察や脳のスキャンなど，すべて測定に数えられる。
- **操作化**。“操作化” という用語は，測定と構成概念の論理的なつながりを指す。あるいは，理

論的構成概念から測定を導き出そうとするプロセスのことである。

- **変数**。最後に，新しい用語を。変数とは，世界に存在する何かに対して，私たちが測定を適用した時の結果である。つまり，変数は実際の“データ”で，最終的に我々のデータセットになるものである。

実際には，科学者でさえこれらの擁護間の区別を曖昧にしようとする傾向があるが，その違いを理解しようとすることはとても有益である。

1.2 測定の尺度水準

前のセクションで示したように，心理学的な測定の結果は変数と呼ばれる。しかしすべての変数が同じ質的なタイプではないので，どういうタイプがあるのかを知っておく必要がある。変数のタイプの違いを区別する大変便利な概念として**測定の尺度水準**というものがある。

1.2.1 名義尺度水準

名義尺度水準の変数は (**カテゴリカルな** (質的な) 変数と言われることもあるが)，異なる取りうる可能性との間に特定の関係がないもののことである。この種の変数は，他の値“より大きい”とか“より良い”といったいかなる意味も持たないし，その平均を出しても全くなんの意味もない。典型的な例として，“目の色”を考えてみよう。目はブルー，グリーン，ブラウンだったり他の色であったりするが，それのどれかが他のものよりも“大きい”とは言わない。だから，“平均的な目の色”について語るのも憚られる。同様に，ジェンダーもそうだろう。男性が女性よりも良いとか悪いとかいうものではない。同じく“平均的な性別”というのも意味がない。つまり，名義尺度水準の変数は，異なる変数の値は異なるということしか言えない。それだけだ。

もう少し注意深く見てみよう。私が人の通勤・通学手段について研究しているとしよう。私が測定しようとしている変数は，仕事に行くためにどういう種類の移動手段を使うかである。“移動のタイプ”変数は，それほど多い値を取るものではない。せいぜい“電車”，“バス”，“自家用車”，“自転車”ぐらいだろう。ここでは，この四つの可能性しかないことにしよう。その上で，100 人の人に対して

どうやって仕事場に行ったかを調べて，次のような結果になったとする。

移動手段	人数
(1) 電車	12
(2) バス	30
(3) 自家用車	48
(4) 自転車	10

さて，移動手段の平均はなんだろう？ 明らかにその答えはひとつではない。愚問としか言いようがない。車で移動することが最もポピュラーな方法で，電車で移動するのがもっともポピュラーでない，それだけである。同時に，私が選択肢を列挙した順番が面白くないことにも注意してほしい。データは次のように表示することだってできるのだ。

移動手段	人数
(3) 自家用車	48
(1) 電車	12
(4) 自転車	10
(2) バス	30

…実質的には何も変わってない。

1.2.2 順序尺度水準

順序尺度水準にある変数は，名義尺度水準の変数より少し構造化されているが，それほどで大きな変化ではない。順序尺度水準の変数は，異なる状態に自然な意味のある順序を与えたものであるが，それ以上のことはない。よく使われる例としては，“レースの最終順位” である。最初にゴールした人に，2 位の人よりも早かったということはできるが，どれぐらい速かったかということはできない。結果的に，1 位 >2 位ということはわかるし，1 位 >2 位ということもわかるが，1 位と 2 位の差が 2 位と 3 位の差よりも大きいかもしれない。心理学的に面白い例もある。私が気候変動に対する人々の考え方に興味をもっているとしよう。私は，その人の信念にもっとも当てはまる文章がどれかを，(以下にリストした 4 つの文章から) ピックアップして人に尋ねる。

(1) 気温は人間の活動のせいで上昇している
(2) 気温は上昇しているがなぜかはわからない
(3) 気温は上昇しているが人間のせいではない
(4) 気温は上昇していない

この 4 つの文章は，確かに自然な序列がついている。“現在の科学に対してどれほど同意するか” という意味でだ。文章 1 は明らかに科学的営みにマッチしているし，2 はかなり適合している。3 はそれほどでもないし，4 は現在の科学と逆の立場だ。だから，私の興味のある用語 (人は科学を支持

する度合い) でいうと，項目を 1 ＞ 2 ＞ 3 ＞ 4 の順に並べることができる。この順序づけがある以上，次のように選択肢を並べると奇妙なことになる。

(3) 気温は上昇しているが人間のせいではない
(1) 気温は人間の活動のせいで上昇している
(4) 気温は上昇していない (2) 気温は上昇しているがなぜかはわからない

…これは項目の自然な“構造”に違反しているように見えるからだ。では，100 人にこの質問をして，以下のような答えを得たとしよう。

反応	度数
(1) 気温は人間の活動のせいで上昇している	51
(2) 気温は上昇しているがなぜかはわからない	20
(3) 気温は上昇しているが人間のせいではない	10
(4) 気温は上昇していない	19

このデータを分析するときは，グループ (1),(2),(3) を一緒にするのは合理的だと思われるし，100 人のうち 81 人が*少なくとも一部は*現在の科学を支持していると言えそうだ。また，グループ (2),(3),(4) をまとめることもあり得る話で，100 人のうち 49 人が現在の科学的観点に，*少なくともいくらか同意していない*といえる。しかし，(1),(2) と (4) を一緒にすることはおかしなことで，100 人のうち 90 人がどうこうしてるとは言えない。これらの反応を一緒くたにする合理的な根拠がないからだ。

つまり，私たちがこれらの項目が，意味のあるグルーピングで自然に序列づけ*できる*ということはできるが，その平均を求めるということはできない。例えば，ここでの例でいうと，“平均的な”反応は 1.97 である。もしその意味がわかるというのなら，ぜひ知りたいのでどうぞ教えてください！

1.2.3 間隔尺度水準

名義や順序の尺度水準に比べると，**間隔尺度水準**の変数や比率尺度水準の変数は，かなり意味を持った数値である。感覚尺度水準の変数は，数字の間の*差分*が解釈可能だが，“自然な”ゼロの値をもつものではない。間隔尺度水準の良い例は，セ氏で測られる気温である。例えば，昨日は 15° で今日は 18° だというとき，この 3° の違いに意味がある。もっというと，この 3° の違いは，7° と 10° の間の 3° と*全く同じ*である。つまり，間隔尺度水準では足し算引き算に意味があるのだ[*3]。

[*3]実は，かなり物理学の知識がある読者に教わったのだが，この温度は厳密な間隔尺度水準ではない。というのも，3° 温度を上げようとするのに必要なエネルギーの量は，その時の気温に依存するらしい。だｋら，物理学的に注意していうなら，温度は間隔尺度とはいえないのだ。しかし例としていいものなので，このちょっとした不都合な真実は無視しようと思う。

しかし注意してほしいのは，0° が “温度がない” ことを意味するものではない点だ。これが意味するのは “水が凍る温度” であって，これはちょっと恣意的なものだ。だから，温度を掛け算したり割り算したりしようとするのはポイントがずれていることになる。20° は 10° の二*倍*熱いとは言えないし，20° が −10° のマイナス 2 倍というのも意味がないことだ。

あらためて心理学的な例を見てみよう。大学一年生の態度が時間とともにどう変わっていくかに興味があったとしよう。もちろん，各学生が入学した年を記録するだろう。これは間隔尺度水準の変数だ。2003 年に学生になった人は，2008 に始めた学生よりも 5 年早い。しかし，2008 を 2003 で割って，後者の学生を前者の学生に比べて “1.0024 時間後の人” とは言わないだろう。そんなの全く意味がないじゃないか。

1.2.4 比率尺度水準

四つ目の，そして最後の変数の種類は**比率尺度水準**の変数といわれ，0 がゼロという意味を持つものであり，掛け算や割り算を許すものである。比率尺度水準にある変数の心理学的な例としては，反応時間 (Response time:RT) がいいだろう。人が問題を解いたり質問に答えたりするのに要する時間を図ることは，様々なタスクでよく見られるものである。というにも，それがタスクがいかに難しいかを示す指標になるからだ。アランが質問に回答するまで 2.3 秒かかり，ベンが 3.1 秒かかったとしよう。感覚尺度水準と同じように，足し算や引き算はどちらもここで意味のある操作である。ベンは実際，$3.1 - 2.3 = 0.8$ 秒長くアランより時間を要した。しかし，今回は掛け算や割り算も意味があるのである。つまり，ベンはアランの $3.1/2.3 = 1.35$ 倍長く時間がかかったとも言える。比率尺度水準にはなぜこんなことができるかというと，反応時間はまさに “0秒” が “時間が経過していない” ことを表しているからだ。

1.2.5 連続か離散か

あなたが考えておくべき第二の分類方法がある。それはあなたが分析に持ち込もうとしている変数の種類に関するものだ。連続変数と離散変数という区別があるのだ。この区別は以下の通りである。

- **連続変数**はどの二つの値を取っても，論理的には必ずその両者の間に値がありうるもの。
- **離散変数**とは，連続的でない変数のこと。離散変数は，中間点に対応するものがない。

この定義はおそらくちょっと抽象的なのだが，例を見たらすぐにシンプルなものだとわかってもらえるだろう。たとえば，反応時間は連続変数である。アランが 3.1 秒，そしてベンが 2.3 秒反応に要したとして，キャメロンがその間の 3.0 秒になるかもしれない。デイビッドが 3.031 秒，つまりキャメロンとアランの間に反応時間が来ることもありえる。現実的には，反応時間をそんなに正確に測定することはできないかもしれないが，原則的には確かに可能なのだ。反応時間の 2 つの値の間に，常に新しい値を見つけることができるのであれば，RT は連続変数として考えられる。

Table1.1　尺度水準と離散/連続の関係。チェックマークが入っているセルは可能であることを意味する

	連続	離散
名義		✓
順序		✓
間隔	✓	✓
比率	✓	✓

離散変数はこのルールを破った時に現れる。例えば，名義尺度水準は常に離散変数だ。移動手段の種類の例を思い出してみると，電車と自転車の “間” に何かがあるとは言えないし，数学的に 2 と 3 の間に 2.3 があるという意味をもたない。移動手段は離散変数なのだ。同様に，順序尺度水準も常に離散変数である。“2 位” は “1 位” と “3 位” の間に常に入るが，“1 位” と “2 位” の間には論理的に何も生じ得ない。間隔尺度水準と比率尺度水準は事情が違う。上で見たように，反応時間は (比率尺度水準の変数だが) 連続変数である。セ氏で測られる温度も (間隔尺度水準の変数だが) 連続変数である。しかし，学校に通い始めた年 (間隔尺度水準) は離散的である。2002 年と 2003 年の間には X 年が存在しないのだ。正誤判定できる問題の正答数も (比率尺度水準の変数だが)，離散変数である。正誤がわかる質問は，“部分的に正解” とすることができないので，5/10 と 6/10 の間というのが考えられない。表 **??**は尺度水準と連続/離散の区別の関係を要約したものである。チェックマークは可能であることに対応している。強調しておきたいのだが，(a) テキストによっては間違えているものがあって，(b)“離散変数” は “名義尺度水準のこと” と言ったりするのだ。残念なことだ。

1.2.6　複雑なもの

いいかい，これを聞いたらショックを受けるだろうことはわかるけど，現実世界はこの小さな分類法が提案するよりもずっと厄介なものがある。現実世界においてはこの綺麗なカテゴリーにぴったり当てはまる変数はないので，測定の尺度を堅苦しいルールのように扱わない方がいい。そういうものではないのだ。

これはガイドラインに過ぎず，違う変数の種類を違うように扱うべき状況について考える助けになるようなものにすぎないのだ。それ以上のものではない。

古い例を取り上げてみよう。多分*ほんとうに*古典的な例の，心理学的測定ツールである，**リッカートスケール**のことを考えてみる。リッカートスケールはあらゆる調査デザインにつかえるものだ。あなた自身，何百，何千回と回答したことがあるだろうし，もしかするとあなた自身も使ったことがあるかもしれない。以下のような調査の質問をしたとする。

> “海賊はみんなすごい” という発言に対するあなたの意見を最もよく表しているのは，
> 次のうちどれですか?

そして選択肢が次のように提示される。

> (1) 全く同意できない
> (2) 同意できない
> (3) どちらとも言えない
> (4) 同意できる
> (5) 強く同意できる

これは 5 件法のリッカーと尺度の例で，いくつかの (ここでは 5 つの) 順序づけられた可能性の中から一つを選ぶ。一般的にはそれぞれのケースに言葉で説明が加えられる。しかし，全ての項目が正確に記述されているとは限らない。以下のような表示の仕方も，典型的なリッカートの 5 件法である。

> (1) 全く同意できない
> (2)
> (3)
> (4)
> (5) 強く同意できる

リッカートスケールはとても便利なツールだ。ちょっと限定的だとしても。この質問はどの種類の変数になるだろう？ 明らかに離散的だ。2.5 に反応することができないのだから。名義尺度水準でないことも明らかだ。順番通りに並んでいるのだから。そして比率尺度水準でないことも明らかだ。自然なゼロがないのだから。

順序尺度水準か間隔尺度水準なのだろうか？ 一つの意見ではあるが，“強く同意できる” と “同意できる” の差分が，“同意できる” と “どちらとも言えない” の差分と同じサイズだということを，証明するのは不可能に思える。実際，日常生活の感覚では，これらは全く同じなはずがない。これに従うと，リッカート尺度は順序尺度水準として扱うべきだということになる。一方で，実践的にはほとんどの参加者が “1 点から 5 点までの尺度” 全体をかなり真面目に扱っているようで，5 つの選択肢の違いが互いに似ているように振る舞う傾向がある。このことから，ほとんどの研究者がリッカート尺度のデータを間隔尺度水準として扱っている[*4] これは感覚尺度水準ではないが，実際は十分近似できていると考えて，**疑似的間隔尺度水準**として扱うのが普通だ。

[*4] 嗚呼，心理学…は何に対しても応えるのが難しいなあ！

1.3 測定の信頼性を査定する

ここまで，理論的な構成概念をどのように操作化し，それによって心理学的測定に変えるかについて，少し考えてきた。そして心理学的測定によって，多くの異なる種類の変数を得る。そこで，次の質問を検討してみよう。測定はうまくいったのか？ これをふたつの関連する用語で表現するなら，*信頼性*と*妥当性*の話になる。単純にいうなら，**信頼性**とはあなたが測定したものをどれぐらい*正確に*測っているかについてのものであり，妥当性はその測定がどの程度*精度があったか*についてのものである。このセクションでは信頼性について論じよう。妥当性についてはセクション**??**でみていこう。

信頼性は実にシンプルな概念だ。それは測定の反復可能性，あるいは一貫性を表す言葉である。私の体重を “バスルームの体重計” で測定したものは非常に信頼性が高い。もし私が体重計に乗ったり降りたりを繰り返しても，同じ答えを示し続けるだろう。私の知能について，“母親に尋ねる” のは非常に信頼性が低い。ある非，彼女は私がちょっと賢いと言い，別の日に彼女は私を全くの馬鹿者といった。信頼性の概念は測定が正しいかどうかとは異なるものだ (測定の正しさは妥当性に関係する)。私がバスルームの体重計に乗り降りする時に，ジャガイモの袋を持っていたとしても，信頼性は高いままだ。つまり，同じ答えを出し続けるという意味で。しかし信頼性が高いからといって，私の体重と一致しているかというとそうではなく，間違った値になっている。これを専門的にいうと，*信頼性はあるが妥当ではない*測定ということになる。同様に，私の母が私の知能についていうことは信頼性がないが，彼女がいってることはいくらか正しい。たぶん私はものすごく聡明というわけではないし，彼女が私の知能を推定するときは日によって乱高下するものの，基本的には正しいだろう。これは*信頼性は低いが妥当ではある*測定ということになる。もちろん，もし私の母による推測があまりにも信頼性が低いものであると，彼女が私の知能についていうところの数多くのクレームのうち，どれが実際に正しい表現なのかを見極めることは難しくなるだろう。だから，ある意味で，信頼できない測定というのは実践的な目的において妥当でないものになってしまうのだ。だから，多くの人がいうように，信頼性は妥当性にとっての必要条件 (しかし十分条件ではない) ということになる。

オーケイ。では信頼性と妥当性の違いがはっきりしたところで，違うやり方で信頼性を測定する方法を考えよう。

- **再検査信頼性**。これは時間が経っても一貫しているかどうかに関するものである。もし後日同じ測定をしたら，同じ答えが得られるだろうか？
- **評定者間信頼性**。これは人が違っても一貫しているかに関するものである。もし誰かが同じ測定をしたら (たとえば，別の人が私の知能について評定したら)，同じ答えが得られるだろうか？
- **平行検査信頼性**。これは理論的に質が等しい測定を使っての一貫性に関するものである。もし

違う体重計を持ってきて私の体重を測定したとして，同じ答えが得られるだろうか？

- **内的一貫性信頼性**。もし測定が同じ機能を持つ異なるいくつものパーツから構成されているとしたら (たとえば，性格検査の質問紙の結果はいくつもの問いを通じて迫っていくものだが)，個々のパーツが同じ答えを出す傾向にあるかどうか。

全ての測定が全ての形式の信頼性を満たすというものではない。たとえば，教育評価は測定の一種として考えることができる。私が教えているある科目，*計算論的認知科学*は，研究課題と試験 (プラス，その他少し) の要素から評価することになっている。試験は研究課題の評価とは幾分異なる側面を評価しようとしているから，評価全体としては内的一貫性が低い。しかし，試験はいくつかの質問から構成されていて，それは同じものを (近似的に) 測定しようとしているから，同じような結果を出す傾向にある。つまり，試験そのものはかなり高い内的一貫性をもっているのだ。これは当然のことである。信頼性を求めるのは，同じものを測定したい場合に限るべきだ!

1.4 変数の “role”： 予測変数と結果変数

変数の話から移る前に，最後にもう一つ用語を説明しておこうと思う。普通，我々は研究の結果として多くの変数を手にすることになる。そこで，我々がデータを分析する時に，他の変数に関連づけてある変数を説明しようとすることはよくあることだ。このとき 2 つの役割，つまり “説明する” と “説明される” という役割の違いを意識することが大事だ。今からこれについて明らかにしていこう。まず，何度も繰り返し参照することになるので，変数を記述するのに数学的な記号を使うことにしよう。“説明される” 変数を Y で，“説明する” 変数を X_1, X_2, のように表記することにする。

X と Y という異なる名前をつけて分析するのは，分析において異なる役割を演じるからだ。これらに対して，古くは**独立変数** (Independent Variable,IV)，**従属変数** (Dependent Variable,DV) と名付けられていた。The IV is the variable that you use to do the explaining (i.e., X) and the DV is the variable being explained (i.e., Y). 独立変数は説明に使うもの (つまり X) で，従属変数は説明される変数に使うもの (つまり Y) である。この名前の背後にある意味は次のようなものだ。もし X と Y の間に本当に関連があるのなら，Y が X に依存している・従属していると言えるし，研究が “適切に” デザインされていたら，X は他の何物からも独立しているはずである。しかし，個人的にはこのネーミングはマズイと思う。これだと，ミスリーディングであることを忘れそうになるからだ。なぜなら，(a) 独立変数は実際に “何物からも独立である” というわけにはいかないし，(b) 変数間に関係がなかったら，従属変数が独立変数に従属することはないのである。現に，独立変数と従属変数というのがマズいネーミングだと考えるのは私一人ではないので，もっと良さそうな他の名前がいくつもある。この本で使うのは**予測変数**と**結果変数**である。このアイデアは，変数 X(予測変数) を使っ

Table1.2　データセットで分析に使われる変数が担う異なる役割について，区別を明確にした用語。この本では古典的な用語は使わず，新しい用語を使うようにしていることに注意してほしい。

変数の役割	古典的な名称	モダンな名称
“説明される”	従属変数 (DV)	結果変数
“説明する’	独立変数 (IV)	予測変数

て Y(結果変数) について何らかの推測をしようとするから，というところから来ている[*5] 。これについて，表 **??**にまとめておく。

1.5 実験的，あるいは非実験的研究

あなたが気にするべき，大きな区分基準は “実験的研究” なのか，“非実験的研究” なのか，という違いだ。この区別をする時，本当に話をしているのは，実験者がその研究において人と出来事全体をどの程度コントロールしているか，ということについてである。

1.5.1 実験的研究

実験的研究の特徴は，研究者が研究のあらゆる側面をコントロールしていること，特に被験者がその研究の途中でどういう経験をするかをコントロールしていることにある。実際的には，研究者は予測変数 (独立変数) を操作したり変化させたりするが，結果変数 (従属変数) は自然に変化するに任せる。このアイデアは，予測変数 (独立変数) を意図的に変化させ，それが結果に何らかの因果的な効果を与えるかどうかを見るためにある。もっというと，予測変数以外に結果変数の原因となるものが存在する可能性がないことを保証するため，他のものは一定に保つか，ある種 “バランスの取れた” やり方で，結果に影響を与えないことを保証するのである。実践上は，他の何ものも実験の結果変数に影響を与えないと*考える*のは不可能だし，一定に保つというのも難しい。これについての標準的な解決方法は，**乱打マイゼーション**である。つまり，異なる群にランダムに人を割り付け，群ごとに異なる処置をする (すなわち，違う予測変数の値をもたせる)。ランダマイゼーションについては後ほどもっと詳しく説明するが，ここではランダマイゼーションが群間にみられるあらゆる系統的な変

[*5] しかし腹立たしいことに，別の名前もよく使われる。それをリストアップしようとは思わない – そんなことをしたって何にもならないからだ。ただし，私が “結果変数” としたところが時折 “反応変数” とされることがある，ということだけは指摘しておこう。やれやれ。この種の用語の混乱はとてもよくあることで，まいっちゃうよね。

化の可能性を (無くすことは無理だとしても) 最小化するために行われるものであることを確認できれば十分だろう。

ごく単純で，まったく非現実的で，非倫理的な例をみてみよう。喫煙が肺がんを引き起こすことを検証したいとする。これを実現する一つの方法は，タバコをお吸う人と吸わない人を見つけ出して，タバコを吸う人の肺がん率が高いかどうかを見ることだろう。これは全く実験的ではない。なぜなら，研究者がタバコを吸う人と吸わない人に対してコントロールしていないからだ。そしてこれは大問題なのである。たとえば，タバコを吸う人は食生活が良くない傾向があるかもしれないし，アスベスト鉱山で働く傾向があるといったことが考えられるからだ。ここでのポイントは，群 (喫煙者と非喫煙者) は多くの点において異なっており，喫煙習慣の違い*だけ*ではないということだ。だから喫煙者の方が肺がん率が高いということは，他の何か別の原因があるかもしれず，タバコが原因ではないかもしれないのだ。他の要因のこと (たとえば食習慣とか) は，専門的には “交絡変数” というが，これについてはまた後で話すことにしよう。

今はとりあえず，適切な実験がどのようなものかを考えたい。ここでの問題点は，喫煙者と非喫煙者は多くの点で異なっているということだった。解決策として，もしあなたに倫理観がなかったとしたら，誰が吸うか吸わないかを*コントロール*すればよい。具体的には，若い非喫煙者を無作為に 2 つのグループに分けて，その半数を強制的に喫煙者にする。そうすると，半分が喫煙者であるということ以外の点で，量グループが異なるところは非常に低くなる。こうして，もし喫煙者群が非喫煙者群よりも肺がんになる確率が高ければ，(a) 喫煙が癌の原因であり，(b) 我々は殺人者だ，というはっきりした確信を得ることができる。

1.5.2 非実験的研究

非実験的研究は，広い意味で “研究者が実験で行うようなコントロールをしないあらゆる研究” を指していると言えるだろう。もちろん，科学者はいつもコントロールしたがるものだが，上の例にあったように，コントロールできないとかするべきでない状況というのもある。がんになるかどうかを見るために，人に強制的にタバコを吸わせるのは論理的に大きな問題 (というかほとんど犯罪) だから，実験者がコントロールすべきでない状況の良い例といえるだろう。しかし別の問題もある。倫理的な問題を横に置いたとしても，“喫煙実験” は他に幾つかの問題があるのだ。たとえば，“強制的に” 喫煙させようと言った時，私は非喫煙者のサンプルが喫煙を*始めさせる*という意味で，喫煙者にするという意味だった。これはマッドサイエンティストが好みそうな，確実かつ悪どい実験デザインのように聞こえるが，現実世界での効果を検証する方法としては，それほど健全ではないだろう。たとえば，喫煙が肺がんの原因になるのは，その人の食生活がわるいからで，普段から喫煙をする人は食生活が悪いのかもしれない。しかし我々の実験における “喫煙者” は “普通の” 喫煙者ではないから (つまり，非喫煙者を喫煙者にしたとしても，それは他の普通の，現実の喫煙者が持つ特性を持ってないので)，食生活は良いのかもしれない。そうだとすると，彼らが肺がんになることはないから

我々の研究は失敗だ。これは “普通の” 世界の構造を壊してしまったから (専門的には，これは “人為的な” 結果といわれる) である。

非実験的な研究をふたつのタイプに分ける区分として注目すべきは，**準実験研究** と **事例研究**である。上で述べた例である，喫煙者と非喫煙者の肺がん発症を検証するために喫煙者と非喫煙者を統制せずに行う研究は，準実験デザインといえる。つまり，これは実験と同じだが，予測変数 (独立変数) を操作していないのである。それでも結果を統計的に分析することはできる。ただし，より注意深く，慎重であらなければならない。

別のアプローチである事例研究は，一つ，あるいはごく少数の事例をとても詳細に記述することをねらう。一般的にいって，事例研究の結果に統計的な分析を適用することはできないし，わずかな独立した事例から “人は一般に”，という一般的な結論を引き出すことはかなり難しい。とはいえ，事例研究は状況次第で有用なのである。まず，他の手法がない状況。神経心理学では，この問題はよくあることである。特定の脳の領域にダメージを受けているひとをたくさん集めることはできないし，そうするとできることはいくつかの事例について，できるだけ詳細に記述することだけである。しかし，事例研究にも素晴らしい利点はある。多くの人を研究対象にできないからこそ，各事例における特定の要因が果たす役割を理解するために，時間とエフォートをかなり注ぎ込むことができるからだ。これは十分にやる価値のあることである。結果として，事例研究はあなたが実験，準実験デザインで見られるような，より統計学的なアプローチを補うことができる。この本では事例研究についてそれほど多く触れないが，紛れもなく価値のある手法なのである!

1.6 妥当性を検証する

他の何よりも科学者が欲するものは，研究が “妥当” かどうかである。この**妥当性**という考え方の背後にある考え方は，とてもシンプルなものだ。あなたは自分の研究が信用できるか？ もしできないのなら，それは妥当ではないということだ。しかし，そう言ってしまうのは簡単なのだが，信頼性をチェックするよりも妥当性をチェックすることのほうが難しい。本当のことをいうと，妥当性が実際なんなのかについて正確で明確な同意を得ることはできない。現に，様々な種類の妥当性があって，それらは個々別々の問題を扱っている。また，全ての妥当性が全ての研究に当てはまるというものでもない。ここでは 5 つの異なる妥当性について論じようと思う。

- 内的妥当性
- 外的妥当性
- 構成概念妥当性
- 表面的妥当性
- 生態学的妥当性

はじめに，何が問題なのかについての簡単な紹介をしておこう。(1) 内的・外的妥当性は最も重要なものである。なぜなら，それはあなたの研究鵜が実際に意味があるかどうかという基本的な問いに直結しているから。(2) 構成概念妥当性は，あなたが考えていることをあなたが測定できているかどうか，を問う。(3) 表面的妥当性は “見え方” が重要であること以外，それほど重要な問題ではない。(4) 生態学的妥当性は表面的妥当性の特殊ケースで，注意すべき多くの側面の表面的なものに対応している。

1.6.1 内的妥当性

内的妥当性というのは変数間の因果関係について，正しい結論を導き出せるかどうかということに関わっている。

ここで “内的” といってるのは，これが研究の “中における” 関係性を指し示しているからだ。Suppose you're interested in finding out whether a university education makes you write better.

簡単な例で考えてみよう。大学教育によって文章が書くことが上手くなるかどうかを明らかにしたいとする。

これを検証するために，あなたはまず一年生の学生グループを集め，1000 語程度のエッセイを書いてもらい，スペルや文法ミスの数を数えるとする。

次に，明らかに一年生よりも大学教育を受けてきているであろう 3 年生の学生を見つけてきて，同じ実験をする。

そうすると三年生の方がミスが少ないことがわかった，としよう。

そこであなたは，大学教育が文章を書く技術を向上させた，と結論した。あってるかな?

この実験の考えうる大きな問題点は，三年生は年長なので，書くこと以外にも色々な経験を積んでいるだろうということだ。

だから，因果関係を導き出すのは難しいだろう。年長者は作文がうまい? もっと作文の経験をしている人は? もっと長く教育を受けていたら?

このいずれかが真であれば，三年生のパフォーマンスを向上させる*原因になる*のではないか? 年齢? 経験? 教育? 答えられないでしょう。

これが内的妥当性の失敗の例である。なぜなら，あなたが異なる変数間の*因果的*関係を適切に切り分けられてなかったからだ。

1.6.2 外的妥当性

外的妥当性は知見の**一般化可能性**や**応用可能性**と関係している。