the Little Book of Great Research

Getting Great Data
Statistical Analysis
Professional Writing

(no Math)

written and illustrated by
Michael Mittag

1. Introduction(s)

Before we start with the methods, let's introduce each other. But first I want to introduce you to a part of yourself which is about the size of the pea, and very important for all that follows.

Meet your amygdala

***Above:** Your amygdala (you have two of them, and they're each as big as a pea).*

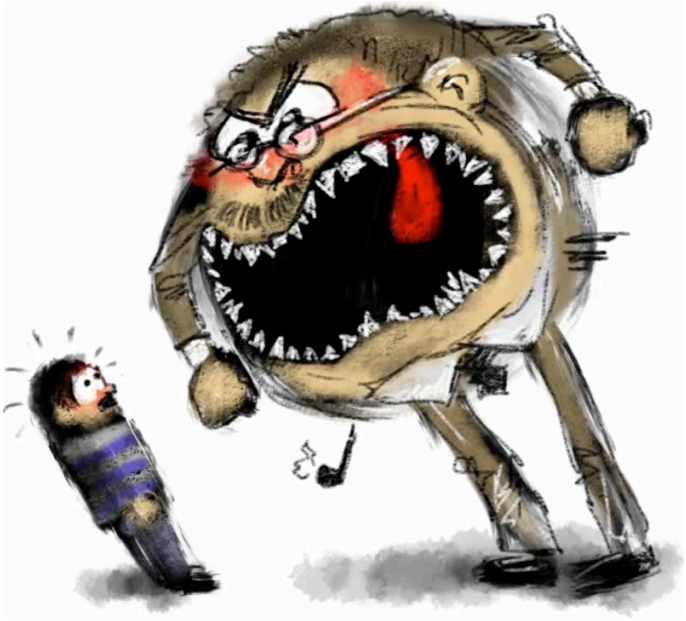
Your amygdala is a pea-sized part of your brain that handles emotional reactions. It's part of the limbic system, which is also called the „reptilian brain", because it's an old brain structure (compared to the neocortex, which allows us to think and speak). As such, it is directly coupled to the emotional systems, and not at all to the cognitive ones. So when the amygdala signals danger, you're afraid, you begin to sweat, and so on. There is not a lot you can consciously do about it, because rational thought does not concern the

amygdala. All that our conscious brains can do is invent stories where it is not so, and sell them to Hollywood where they make movies out of them. But in reality, when your amygdala signals you to run, you run.

Most people's amygdala starts firing pretty good when they see Math or anything that might involve it. That's why I said up front that this book does not contain any. Because I don't want you to run.

What you are feeling is a conditioned response from being bad at Math in school. Especially when you had a Math teacher like mine, who felt that I had large potential, so my mediocre test scores were probably due to abundant flaws in character. At least that was what he gave me (and the class) to understand.

Actually, let's not glorify anything – I had quite a few problems in school, but adding a borderline sadistic, well-meaning Math teacher on top did not solve any of them.



Above: *Mr. Vogelsanger, my high school Math teacher, exploring the finer points of Trigonometry and my incompetence. At least this is how my amygdala remembers it.*

Back to your amygdala: If it fires now, then remind yourself that this is from times past, and that it will stop soon. After a

few minutes, your amygdala will notice that it's just making a fool of itself and quiet down.

And note how it feels, because if you want to do good research, you'll strain your amygdala a few more times.

You see, research involves interacting with people and asking them weird questions. Good research usually involves people you don't know and questions they haven't thought about.

It usually takes guts to talk to strangers on the street, because that's one more situation where most people's amygdala signals danger. Not that there's any real danger involved, but your amygdala doesn't like new people. Confront it a few times and it'll get used to it. Stop doing it for a few months and your amygdala unlearns, and will fire again the next time you go out there with a questionnaire. At least mine always does, and people generally consider me a social person.

So – if you're feeling uneasy about this research thing you're about to learn, that's just a pea-sized brain area giving you weird signals that have nothing to do with who you are today or what you're about to learn. Get used to it and enjoy the rewards. You'll learn new and exotic things, you'll meet new people and talk with them about fascinating topics, and you'll have mastered the key qualification to be a great researcher.

About the book

For some time now, I wanted to write a book about the tools you need to do research. Not an introduction to statistics, and not a 500-page exhaustive work, but the one book that does not exist. The one that's fun to read, easy to understand even if you're not bringing any enthusiasm to the table, that talks plain language and gives you advice you can use.

I know a lot of good researchers who are not into Math, and they currently lack material that helps them understand statistics and methods without going into formula, and that serves as a guide to the methods that are available to answer their research questions.

For demonstrations, I use the occasional textbook example, but try to provide as much real research as possible. As a source, I have used free open-access research from the DOAJ directory, so you can read up all examples for yourself. Which you should totally do, because the stuff is fascinating, especially the articles from the parts of the world you did not know do research. That said, open-access journals not always adhere to the same standards as the ludicrously expensive US-American ones, so I'll note it when an article is written in a way you should not imitate. If you can use university resources, you can find some top quality research in APA journals or on JSTOR. That said, open-access research is certainly less culturally biased, and some of the formal errors the authors commit (such as talking too much about statistics or copy-pasting the output of the statistics program into the article) actually help you understand what they did.

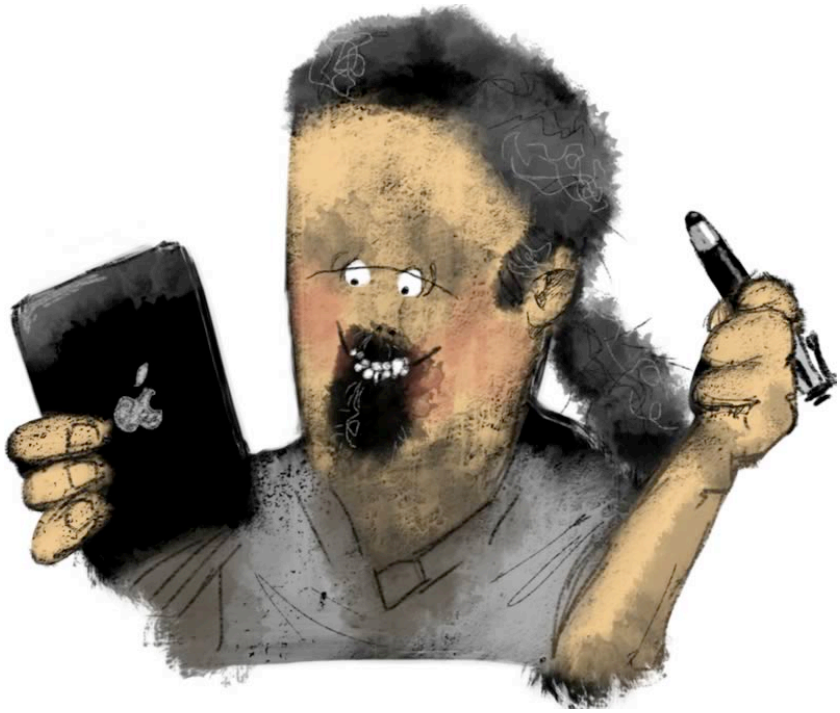
About the author

As you know (if you read all the way until here) I had some trouble with Math in school, mostly because my teacher spread his resources evenly between teaching me how it works and making it hell.

So I was really shocked when I started studying psychology and there was statistics in it (and lots of it). Also, I was slightly better at it than most of my friends, so I spent a lot of time explaining it to them. And, gradually, helping out with theses and academic research.

That said, I haven't published any research papers myself (I went off to pursue a career in multimedia, teaching and writing instead). However, I have helped a lot of researchers with methodological aspects, so I hope that this book can help you too. Let me know how it works out for you.

About the illustrations



Above: *The author and his illustration tools.*

I have created the illustrations to help you understand some of the more difficult concepts and to give you visual cues to remember them.

Also, I hope they're as much fun for you to watch than they were for me to draw. I've created them on the iPad using the Procreate software and a Bamboo stylus.

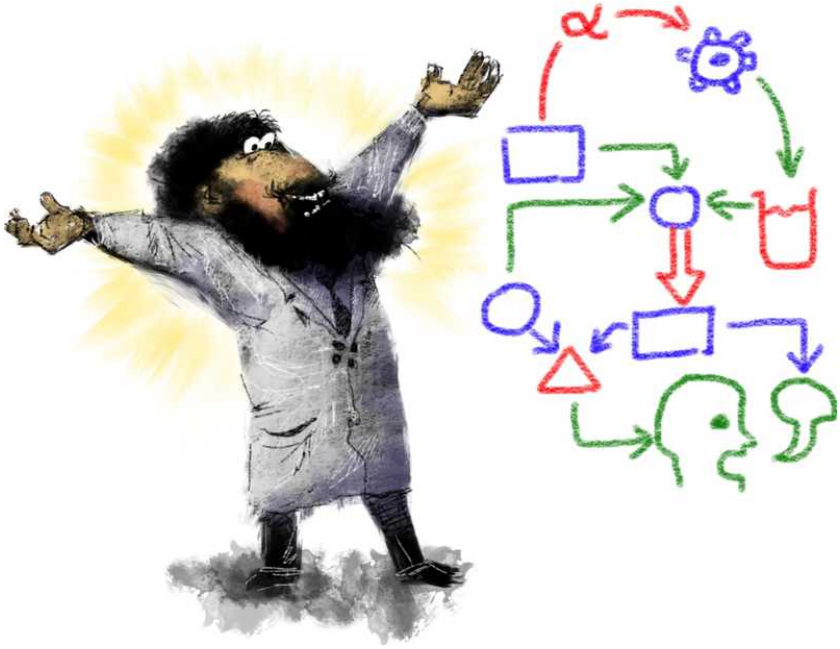
I have tried to be not too biased in what I draw, but it's impossible. I'm a European white male, and simply don't know enough about the rest of the world to provide non-stereotyped imagery for most of it. So the deal is this: If you think that an illustration is biased, send me a sketch or a few non-biased images about the subject, and I'll update the illustration (and post it on my website and give you full credit).

2. *Empirical science in a nutshell*

Empirical science means that you learn about the world from looking at it. That sounds simple, but it's not. A lot of the discussion about it is in the appendix (I originally had put it here, but even a brief overview is a lot of text, and you want to learn how to do it, not what people thought about it in the past).

The old way: Look for theories that explain all of human behavior

Psychologists like Sigmund Freud (and contemporaries) created compelling, fascinating, brilliant theories that tried to explain all of human behavior, from the cradle to the grave. According to Freud, the sex drive can explain nearly everything, and it's as reliable as physical laws.



Above: *In the old days, empirical science was driven by brilliant minds who came up with complex theories that could explain all of human behavior. Theories would spread by sheer brilliance, charisma and eloquence of their inventors.*

This approach has one major drawback: If your theory can explain everything, *you can't test it*. Because whatever you observe, the theory will provide a way to explain it, to the point where nothing you can possibly observe would prove the theory wrong.

And what's worse: The theory *can't predict anything*. If it can explain everything that happens, then it's bad at saying what is going to happen next. Because according to the theory, anything can.

Some modern researchers argue that these theories did not really explain anything, which is certainly too harsh. However, in the long run, it's very difficult to get ahead in science when you cannot test your theories against observation, find faults in them, or find out which of two competing theories is right.

The modern approach: Test predictions (against chance)

So a new type of scientist emerged: One with a more limited view of the world, who was content to create a theory that explains a very narrow field of study, and only some behavior within it, too.

A scientist like Alan Baddeley (to pick just one), who said that human short-term memory is phonological in nature. At this point, let me clarify what short-term memory is (because it's not what you think it is):

Short-term memory stores a few items for a few seconds in your brain.

So far as scientists can tell, there is no separate memory for stuff you learn immediately before an exam and have forgotten a few days later. That's just how long-term memory works, it's the same with newspaper headlines, movie plots, recollecting what you have eaten last week, remembering people's names, and so on. Unless it's really memorable or you go over it time and again, stuff drops out of your memory pretty fast.

When Baddeley says that short-term memory is phonological, he means that when you remember something

for a few seconds, you store the sound of the words in a kind of tape recorder, where you can replay them until they fade away eventually.

To prove it, he had subjects learn word lists. He found that they performed worse when the words sounded similar, presumably because the short term memory has problems keeping the sounds apart. This was even the case when the words were presented in written, so it seems that subjects used sounds for storage, not pictures.

To test his theory, Baddeley used experiments: He put people into identical experimental conditions and compared how they did on the similar-sounding words versus the non-similar-sounding words. He consistently found performance differences, and thus proved his theory.

So how does statistics play into this? – Easy: It tells you whether the differences you find are actual effects or just random noise. Whether what you're looking at is random fluctuation or something you can publish a paper about. But more of that later.

Above: *Modern scientists prove that observations they have made follow the predictions their theories make. That is: They show that there are non-random differences at the places where the theory predicts them.*

Why the new approach is better (although it can be boring)

Now, we have a lot of modern scientists who chase small-scale theories and find their little differences here and there. Acceptance of the theories depends not on the brilliance or rhetorical ability of the scientists, but simply on their ability to predict differences (and prove that the differences they find are not random).

That can be very boring. If you are looking for a good read, then by all means read Freud, or Georg Groddeck (who invented the „Id"), or William James (who studied consciousness and emotions), or basically anybody famous in the old days.

However, modern empirical science has one huge advantage: It is built on proven facts, that is, on observations you know are not random. So even if a theory will turn out to be false somewhere down the road, the facts remain and are free to use for everyone. There is no copyright on scientific fact, so you can use them to build your own theory or apply existing theories to new situations.

And because facts are ultimately so cool and valuable, scientists are careful in creating them. They follow a lot of rules and strict methods (and even the occasional superstition). That's entirely unlike in the old days, where Jean Piaget created an entire theory of development from watching his three children grew up.

Much of the time, scientists' work will also be reviewed by other scientists before publication, just to be sure. For example, most journals apply peer reviews to make sure that everything they publish was read by scientists who can understand it and say it's legitimate.

Above: *The huge advantage of modern empirical research is that the facts are always facts – so whatever you do, there are thousands of empirically proven facts on which you can build.*

2.1. How statistics works

Now that you understand empirical science, let's look at how you (yes, you) can add new facts to it.

A quick recap:

- Alan Baddeley had a **theory**: Short term memory is phonological in nature. When you store something in your mind for a few seconds, you store the sound of that word.
- This theory makes a lot of **predictions**. A really cool one is this: Because the Japanese language uses very short words for numbers and the English one uses comparatively long ones (especially if you pronounce them in the way the Scots do), *Japanese should be better at remembering numbers than Scots*.
- Baddeley (and everyone else) can test this prediction in an **experiment**: Let Japanese and Scottish subjects memorize numbers and see how they perform.

So Baddeley tests the following hypothesis:

Scots perform worse on short-term memory tasks than Japanese.

(Author's note: Baddeley and other researchers did test people from all over the world on short term memory, but they did, to my knowledge, never directly compare Scots and Japanese. However, he would have if he'd have had to illustrate it).

At this stage, there is a problem: *Everyone* performs differently all the time. Some people have slept well, some badly, some are more motivated, some smarter, and so on. More generally:

Whatever humans do, there is always an element of randomness involved.

Let's assume Baddeley starts recruiting the subjects and randomly picks one Scot and one Japanese subject:

Now, any two randomly chosen subjects are probably very different, so how can you compare them? – Essentially, you can't. A naïve, wrong but entirely rational idea would be to pick subjects that are as similar as possible, for example those two:

But how do you measure whether two subjects are similar? There is no real way to do so: Humans are so complex that everyone's different. Also, it would be excruciatingly difficult to find pairs of near-identical subjects, as well as pointless (see below). So he did not do that.

Moving on: Let's assume that Baddeley has recruited two small random groups of subjects, as follows:

We have some random smart guys left and right, some are tired, some motivated, and it's generally a huge mess of different variables that all have an influence on how the subjects perform on a memory test.

Under such conditions, *how on earth* can you make precise measurements? How can you ever come to any conclusions whatsoever? How can you prove that the difference you see between the groups is due to how they pronounce numbers, and not just any random fluctuation?

A solid theory of shaky measurements

As it turns out, you can make a really strong, coherent theory about shaky measurements, which solves all of the problems at once.

The basic idea is that if randomness is everywhere, then we just have to know how big its effect is, so that we know whether the effects we find can be the result of randomness or not.

Let's look at randomness first. Here it is:

***Above**: Randomness (artist's impression)*

Now, some people are better at a task due to random factors, such as how they feel today, whether they have had their morning coffee, how they grew up, maybe even what's in there genes. We summarize all that as follows: Randomness gives some of our subjects a kick that pushes their performance above the statistical average.

Above: *A random kick that boosts somebody's test score above the average.*

Just as often, people randomly perform below par, for example if they have a headache or are absent-minded, or they're just not born for this kind of thing.

Above: *A random bash that lowers somebody's test score below the statistical average.*

At this point, scientists decide that as they already have large effects of randomness in the test scores, they may as well ride with it. ***Scientists do not try to fight the randomness, they just ride with it.***

And they do this as follows:

- By looking at how far individual scores deviate from each other, ***scientists know how large the effect of randomness is*** (they define randomness as anything that is not part of the analysis, in this case, any difference in performance that is not due to nationality)
- ***In a random sample, some of the randomness cancels out***. That is, some of the above-par and below-par scores even out. As a result, any picture scientists get from a sample is much more precise than what they see in individuals.
- ***The bigger the sample, the smaller the effect of randomness***. That is: The more Scots and Japanese Baddeley examines, the less effect there is from one subject's headache, and the less effect randomness has overall.

If you picture the measurement process as a weighting of one group against the other, then randomness gives the overall score a good pull or push (we don't know which).

Above: *Randomness distorting the measurement for the group difference. For small groups, a little randomness can distort the measurements a lot.*

Now, as said, the bigger the groups, the less effect the push and pull of randomness has. Imagine it as follows:

Above*: When examining big groups, randomness has less effect on the results. The bigger the groups are, the*

smaller the effect, and the harder for randomness to cause substantial differences all by itself.

Note that all of this happens according to mathematical laws, so it can be computed (or fed into statistics programs for computation).

Comparing effect and chance

At this point, we have the following information:

- The amount of randomness in each individual measurement (that is: How far the individual measurements are apart).
- The amount of the difference between the two groups.
- How large the groups are (the sample size).

From this, following some sweet Math that Mathematicians have figured out and Programmers have put into software, we can compute *how much randomness pushes and pulls*. In other words, you know now how precise your measurements are, and how much random jitter they include.

There is one small catch: Because randomness is randomness, it does not always push or pull the exact same amount, it's sometimes more and sometime less. All we know is how much push or pull is in the measurement *on average*.

From there, it's one small step to the first of the two figures that scientists almost always tell you, the so-called test value:

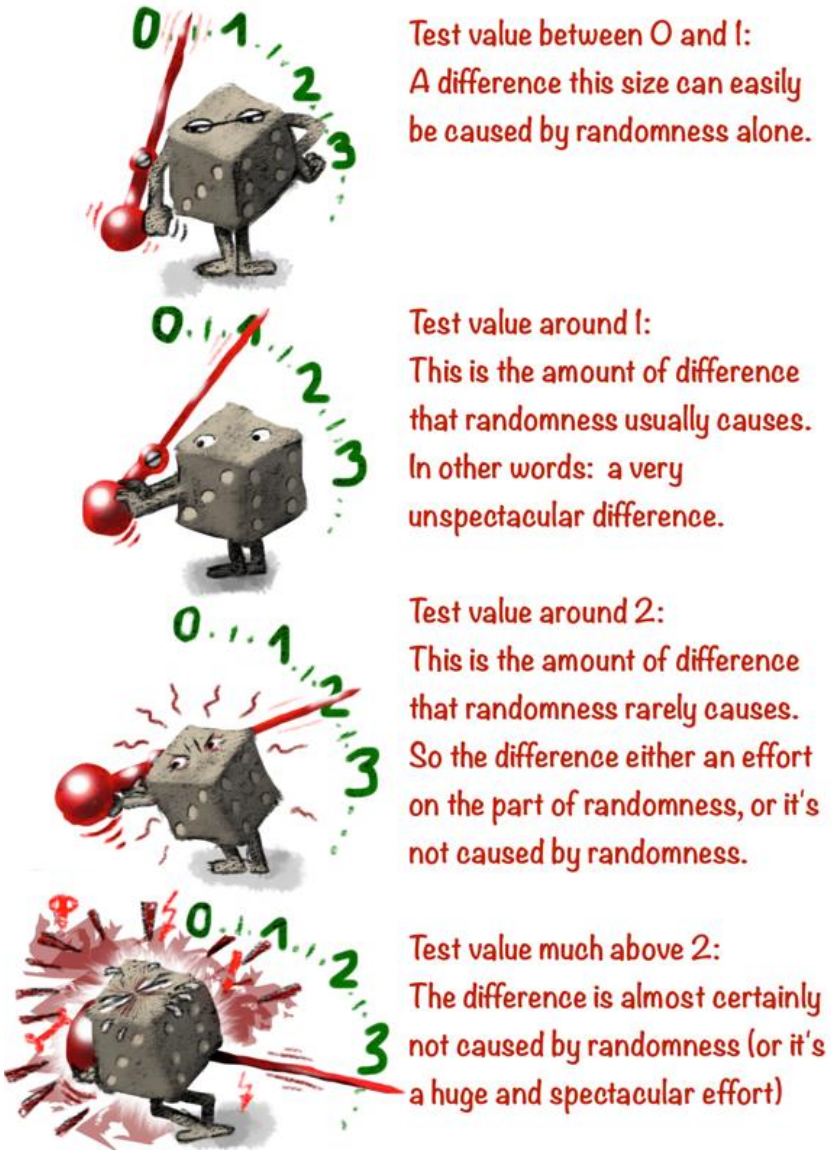
The ***test value*** tells you how much bigger the effect is than the average push and pull of randomness. Most people find this a bit hard to digest at first, but we'll have an entire book for you to get used to it. For starters, let's go over it point by point:

- Randomness always distorts measurements.
- Sometimes, it distorts more, sometimes less.
- We can compute how much it distorts on average. A **test value of 1** means that we have found a difference that is

exactly as big as what randomness normally causes. So it's no big deal.

- A **test value below 1** means that the difference is even less that what randomness normally causes. In other words: Entirely forgettable. You'd get better results from analyzing random numbers.
- A **test value around 2** means that the difference is twice as big as what randomness typically causes, so it's probably not random.

Let's look at this one more time before we go on, this time with pictures:



Above: *A test value says how big the difference is, compared to what randomness causes on average. Note*

that randomness is not a constant force, it sometimes strikes harder than other times.

You may have noted that all of this is somewhat vague – randomness sometimes puts in a little more effort, some time a little less, and it's hard to judge where exactly to draw the line. So scientists came up with another number that does this for them: The statistical significance.

However, as far as test values go, remember the following:

The bigger the test value, the better. Values below 1 mean that we found nothing but randomness.

The p-value (the most important value in statistics)

The statistical significance (or p-value) is the most important value in all of statistics:

The p-value tells you how likely it is that your result is due to randomness.

That is: Even if you find a huge difference, there is always a tiny chance that it's just due to an extraordinary amount of randomness. Fortunately, we know exactly how small this tiny chance is, we can compute it from the amount of push and pull randomness does on an individual level and from the sample size. From this, the statistics program computes the p-value.

The p-value is actually so cool that scientists use it for nothing less than *the definition of what a scientific fact is:*

If the p-value is below 5%, the result is a scientific fact.

Now, „scientific fact" does not mean absolute truth, because there is no such thing in science. It means that you should assume the result to be true and act accordingly: Either live by it or find a way to prove it wrong. Both is acceptable. But ignoring it is not possible any longer.

There are different ways to interpret the p-value. If you look at the research, then the p-value says how likely it is that the effect is caused by chance. If you look at the researcher, then

the p-value tells you how likely it is that what the researcher says is wrong. In other words:

The p-value is a researcher's sword of Damocles.

Above: *Researchers get recognition (and sometimes even money) from their results. The p-value is the sword of damocles: The probability that they're wrong and their result is nothing but randomness.*

If you are unfamiliar with the tale: Damocles was a courtier and envious of the wealth and power of king Dionysius II, so one day the king made him switch places with him. And while Damocles was enjoying the luxury and the food, he noticed a sword hanging over his head, held by a horse's hair, and ready to drop down on him any second. The king explained that this was what power tastes like: You have all the luxury in the world, but at any second, it can be over.

It's also how research works: As a researcher, you can make any statements on how the world works, but at any time, there is always a chance that you're wrong, and all the research results you present are just random differences that happened to be bigger than on an average day. All you can do is make sure that the sword hanging over you is a small one,

so it won't hurt so much when it comes down. So remember this:

p-values, like swords over your head, should be as small as possible. The smaller the better.

Or, in plain numbers:

p-values must be below 5% (that is: 0.05). The smaller the better.

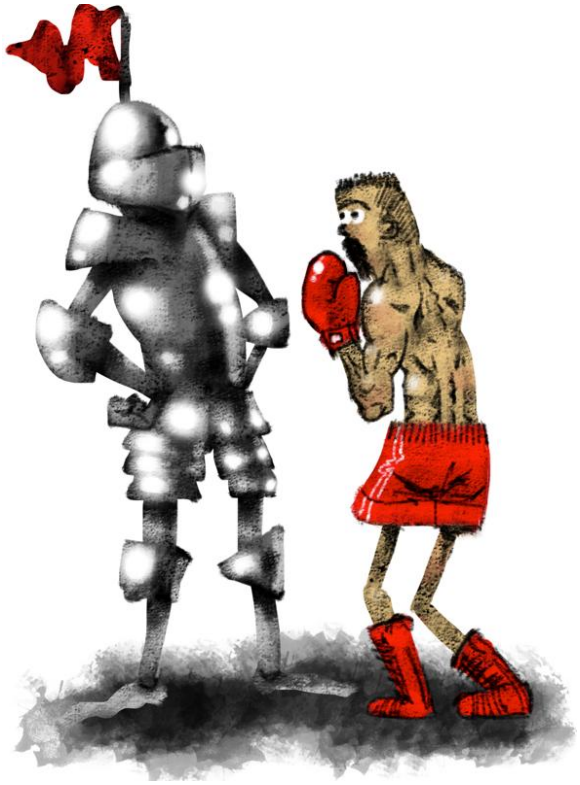
2.2. The ground rules of research

Without further ado, here are the ground rules of all research you'll ever do. Most of the time, you'll follow them very strictly. However, scientist are pragmatic: If it's absolutely impossible to follow a rule, they prefer to keep researching instead of just giving up and doing something else.

Make testable hypotheses

In psychology, Psychoanalysis received a lot of criticism because its predictions are not always testable. For example, Psychoanalysis says that boys are in conflict with their fathers. However, it also proposes mechanisms for how they side with their fathers out of fear, or how they repress the conflict entirely. So when an analyst sees a boy who hates his father, this confirms the theory. A boy who loves his father confirms it, too. As does a boy who seems rather indifferent towards his father. In sum: You cannot observe anything that would contradict the theory.

Modern science is more strict: You must make hypotheses such that they are clearly testable. It must be possible to design an experiment or a survey that proves these hypotheses right or wrong.



Above: *If you present a theory that cannot be defeated by logic or fact, then you're cheating. Whatever you do, it's not science.*

Make hypotheses before data collection

You have to make your hypotheses before you start collecting the data. The reason is that it's *much* easier to come up with a good hypothesis once you have seen the data, so that counts as „cheating".

Above: *Setting your goals after the fact is considered cheating. So: First specify what your research should prove, then get the data, then see if you were right.*

Keep observations independent

This is one requirement from the Math side of things: All your subjects and your observations must be statistically independent. This means:

When you know the result from one subject, you cannot predict the result from any other subject by this.

This is usually not a problem, except in the following situations:

- Any **research you do in a school**. How good school children are at various tasks and what they think about school depends on the school they're in, their classmates,

and the teachers they have. So when you sample one class, you never know if you measure 20 kids or one teacher.

- Any research you do on **text corpuses**. You may have a sample of 200 journal articles about life in Australia in the eighties, but some are probably written by the same authors, created using the same background material, edited by the same editor, using the same style guide as other editors, or inspired by the same then-fashionable literary author.

So, in most cases, you'll have no trouble getting independent observations (and just wonder what all the fuss is about). In the examples above, in contrast, there is no way to guarantee statistical independence, so you'll do the research anyway. And interpret it more carefully.

Above: *Statistical independence means that the results of one person are not connected with the results of another person.*

Possible answers are Yes and No

Science works like a courtroom: You gather evidence and then come to a conclusion. That conclusion is either **Yes** or **No**. The only third option is that you do not have the evidence to prove anything either way.

In a courtroom, they can't sentence a man to half a lifetime in jail because they're 50% sure he killed his wife. It's the same in statistics: Either you can prove beyond reasonable doubt that your hypothesis is correct, or you have nothing.

This is very difficult to grasp for beginners. On the one hand, you're told that scientists are careful and don't jump to conclusions. On the other hand, the methods you use force you to come to a strictly binary conclusion: Yes or No, with little in-between. That is why research articles have very careful wording, which takes some experience to get right.

Above: *Scientific methods always give **yes** or **no** answers, and never „maybe", „very probably, or „rather not".*

Prove difference not equality

One effect of how we do statistics is that you can prove only differences, never equality. So you can prove that men are taller than women, or that women are better at languages than men, but you can never prove that they have the same IQ.

One reason for this is that there is no good definition for „same“. There are only degrees of same: Roughly the same, the exact same number of points, or to the fifth decimal place the same value. It‘s much easier to find a definition for „different“: Any difference you can prove, no matter how large or small.

Above: *When you look closely enough, you may find differences you did not notice before. Because of this, you can never prove that two values are identical. Also, just because you do not find a difference does not mean that there isn‘t one.*

Of course we‘re more interested in substantial differences than in tiny ones. Now, luck has it that most differences that are too small to be of any practical value are also extremely difficult to prove and require huge sample sizes, so

researchers don't usually go to the trouble. Unless they're working for pharmaceutical companies or use very large sample sizes. For example, in a study with 5000 men, researchers found that age-related cognitive decline sets in shortly after age 40 (which was widely reported in the press), at a rate of roughly four percent every 10 years (which was not reported). So: With large samples, you find tiny differences, with normal-sized samples (say, 30 or 100 subjects), any difference you find is probably substantial.

Another thing that beginners often find hard is this: When you find no difference between two groups, the only accepted interpretation is that *you did not find a difference*. There may or may not be one, we may never know (because we can never measure with perfect accuracy). Actually, there almost certainly is a tiny difference between nearly all groups, just below what you can normally measure.

Causality is difficult to prove

Humans tend to think in terms of cause and effect. Often, when we talk about *understanding* something, we mean that we understand what is the cause of the effect we see.

Now, whenever you find that two phenomena A and B are statistically connected, it's tempting to see this as a cause-and-effect relationship (and media will almost always report it as such), but in fact, there are three possible interpretations:

- ***A causes B***
- ***B causes A***
- ***There is a C, which causes both A and B***

For example, assume that you make a study and observe a relationship between video game violence and actual violence: People who play more violent video games are more frequently engaged in fights. This could be interpreted as follows:

- Video games cause people to start fights.
- Being in fights causes people to buy violent video games.

- *There is an external variable that causes both fighting and playing violent video games.*

Such as:

- People with lower income are more often involved in fights, and also play more video games.
- Some people are more aggressive than others, and those seek out both real fights and virtual ones.
- People with bad social skills play more video games (including violent ones), and are also more likely to get into fights.
- People who are frustrated with life get into fights more often and also play more violent video games.
- People who have a lot of spare time play more video games and also get into fights more often, simply because they spend more time in environments where they happen.
- Video games (like most media) lead people to underestimate the dangers of a fight, so gamers are less likely avoid a fight.

In fact, there are **only two ways to prove causality**:

- If there is **no other rational explanation**.
- If you use an **experiment**.

In all other cases, you are supposed to carefully interpret any results as connections between variables, and not jump to a conclusion about causes and effects.

Above: *If you look at the picture above, what is cause and what is effect? Is the kick caused by the swearing or the swearing by the kick? Or both by a hidden external variable (such as the person behind the bush playing a practical joke). – In science, nearly all research faces the problem that it's hard to say what is cause and what is effect.*

3. *Theory of data collection*

Now that you understand how science works, let's look at how you get good data. We'll look at the theory first and then go over the established methods of getting good data.

To understand data collection, you need to learn a few things about humans. There are several dangers lurking when you try to get data from research subjects:

- If you're not careful, you might just measure your own expectations, and not what the subjects actually think or do.
- If you ask the wrong questions, subjects may not understand them, or not be able to answer them correctly. They may fill in those blanks with what they've read somewhere or what they think you want to hear.
- Even if you get accurate, reliable answers, they may just not be what you really need to know. Humans are complex, and it's difficult to understand how they work.

3.1. The psychology behind the data

You can measure two things: What people say and what people do. The technical terms are „behavior“ and „attitude“, and it pays to keep them apart, even though there is a gray area in-between.

Behavior

Behavior is any action we take, such as greeting someone on the street, helping a stranger, smoking a cigarette, picking one product out of the shelf and not the other, pushing a button as fast as possible, and so on.

Usually, you‘ll try to measure behavior as clearly and as directly as possible, ideally in a controlled (experimental) situation while you watch.

Above: *Behavioral data (for this example, buying a flower) is usually the best type of data you can get, but you‘re not always around to observe it.*

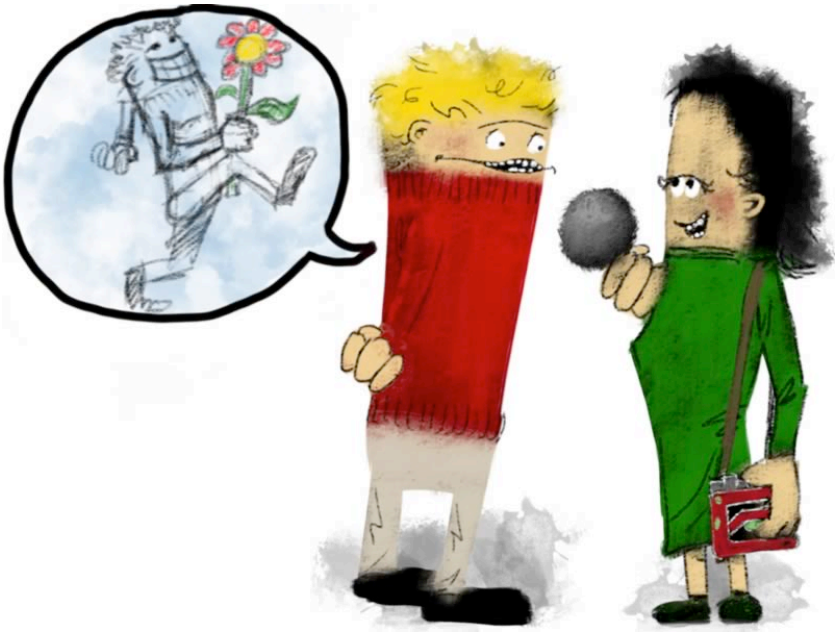
Reported behavior

If you cannot observe behavior directly, you can ask subjects about their past behavior. That's almost as good as direct observation, if done right. There are also a few problems involved.

To get optimal data, consider the following ground rules:

- Subjects can forget things, so **ask things that are easy to remember**.
- **Subjects will not lie** to you. So ask as directly and as clearly as possible.
- **Subjects will try to be nice** to you, so ask the questions in a way that this has no impact on the answers. Also, show interest in all answers, especially ones that do not conform to what everyone says.

If you ask the right questions, reported behavior is as good as observed behavior.

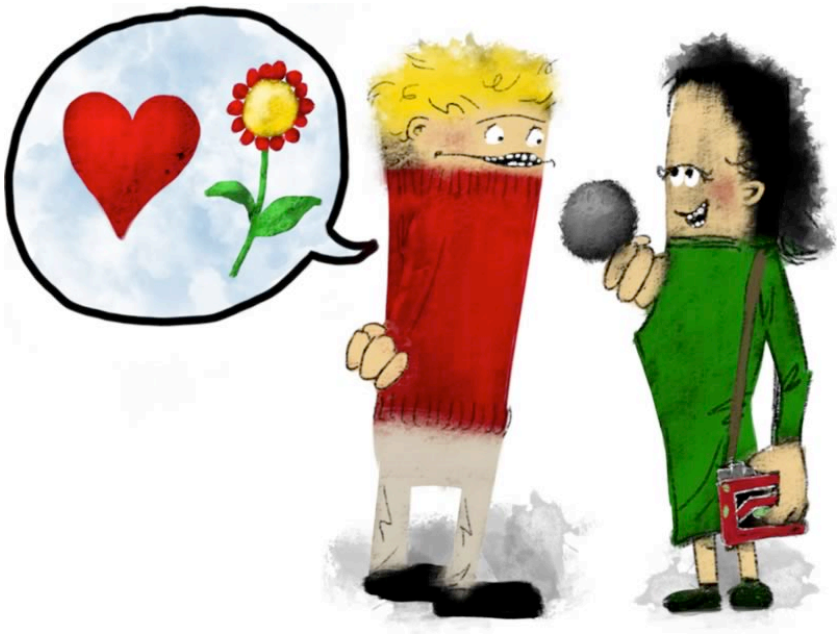


Above: *Reported behavior is good data, just remember that memory can be sketchy and biased, so you need to ask very clear questions.*

Attitude

Attitude is that vague feeling we have towards all sorts of things. Our brain just seems to tell us to like or dislike some of the stuff around us. Which, if you think about it, makes your life a lot easier, because you don't have to analyze all the objects around you all the time. You just know that you can safely mind your own business and your brain will point out all the things that you might want to steer toward or away from.

Measuring attitude is easy, measuring it well is more difficult. While we perceive attitudes as stable and 100% natural parts of ourselves, the reality is much more complex. Attitudes can change depending on the people we're with (and their attitudes), and we constantly filter our attitudes for how appropriate they are at the moment. Still, attitudes are an important area of research, because most of us very much define ourselves by our attitudes.



Above: *Attitudes seem like an easy subject to measure, but they are much more complex than it seems, so measuring them well is difficult. For example, when asked by a woman, most men will probably say that they like flowers. That does not mean they will buy some. Also, when men buy flowers, it's often not because they like the flowers, but because they like the person who receives them.*

How attitudes determine behavior

Usually, they don't.

Attitudes are what you use in private, when you're having a beer, and when people ask you about them. They're a great social tool because you can share them with members of your social group, which strengthens your perception that you belong together.

However, attitudes do not determine what you do.

For example, most of us have a strong attitude against slave labor, yet we buy items that are so cheap that the person producing them cannot possibly have a salary.

One of the most striking proofs for how far people go against their attitudes is Stanley Milgram's experiment. Milgram had his subjects administer increasingly strong (fake) electrical shocks to participants (actually actors) when they failed to answer questions correctly. Most people were prepared to administer the shocks, even at a level where they had to fear that it would kill the subject.

So: Milgram proved that people like you and me can be told to kill someone, and will do it, for no good reason other than somebody told us to. Of course at this point people like you and me say that this has been decades ago and times have changed, and we would never even begin the experiment. However, the experiment has been repeated a number of times all over the world, and it never failed. It's just that we think that our attitudes and core beliefs are what determines our behavior, when actually it's not. Scientists call this the fundamental attribution error.

Above: *Nothing triggers attitudes like alcohol, internet forums and taking part in surveys. Fortunately for everyone, attitudes not always trigger actions.*

How consistent attitudes are

Not much. Attitudes are strong and consistent in most of TV and politics, but not in real life.

For example, a lot of Swiss people think that Germans are typically rather arrogant. A friend of mine was discussing that quite openly in a restaurant with a work colleague, while a German was sitting right next to her – *her husband*.

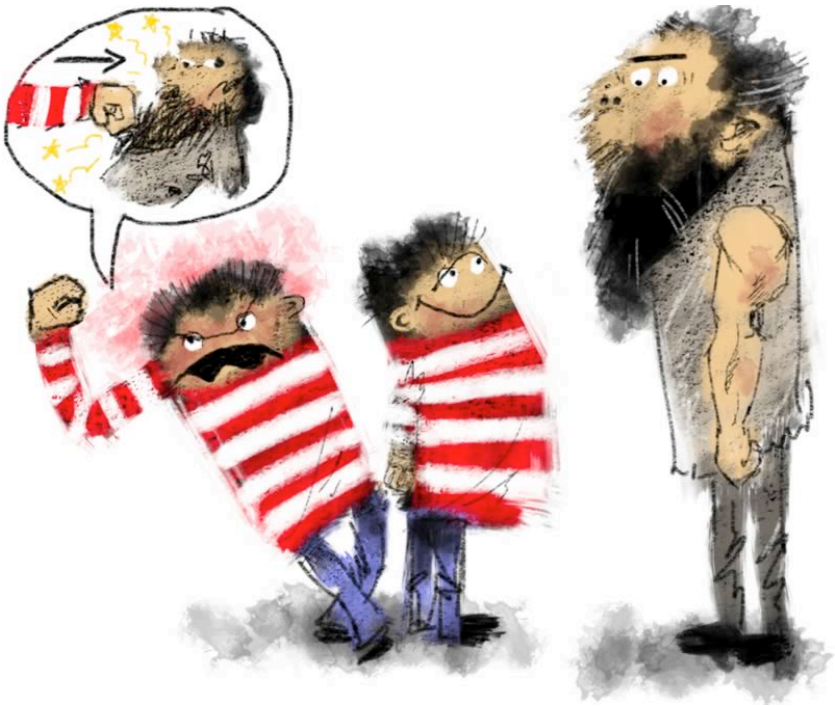
How can this happen? – Actually, quite easily. Attitudes are simply not present all the time in all we do, they are activated depending on time and context. Ziva Kunda and her colleagues have shown this here for racial stereotypes.

Generally speaking, there are some situations that always trigger attitudes:

- when you are confronted with information that is consistent with the attitude
- when you have no other information
- when you‘re drunk

In other situations, what you do has nothing to do at all with any attitudes or stereotypes you may have, for example when...

- you meet someone nice and attractive
- you have personal information about someone
- someone has your respect and/or admiration
- someone agrees with your opinions (or admires you)
- you look into someone‘s eyes and recognize a fellow human
- someone is in a position of authority
- someone is in a position where they can cause you pain, or discomfort



Above: *Attitudes are not always consistent across situations. For example, if you dislike somebody who is more powerful than you, you will usually have different attitudes in the presence and absence of the person. You‘ll prefer to tell yourself that it‘s the same attitude but reason prevented you from following up on it. In reality, reason did have nothing to do with it, attitudes just are not as stable as we like to think.*

3.2. Measuring data quality

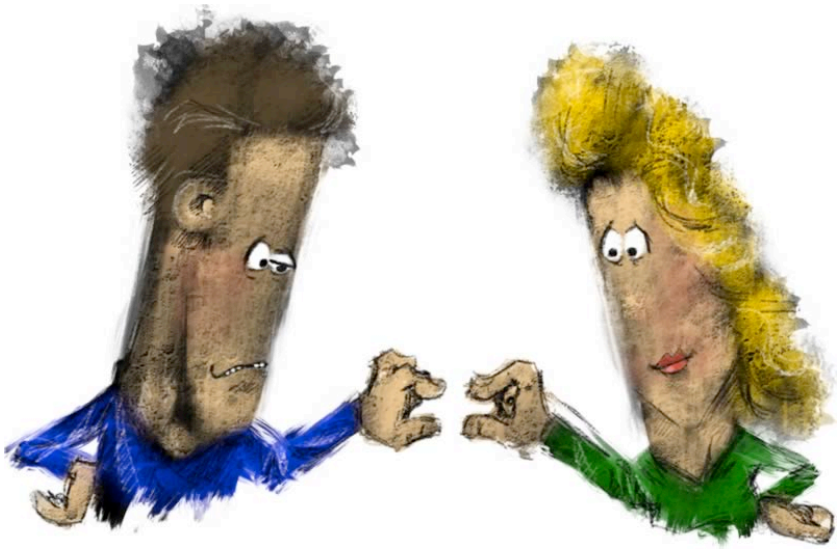
Would it not be nice if you could somehow measure (or even compute) just how good your data is? – Actually you can! The theory for this stems from research on how good psychological tests are (such as IQ tests), but it works well for all areas of empirical research.

There are three core criteria to evaluate any empirical data (and the corresponding research methods):

- ***Objectivity:*** The result does not depend on who performs the procedure.
- ***Reliability:*** If the procedure is performed twice, it yields the same result.
- ***Validity:*** The procedure measures what it pretends to measure.

Let's look at them in some depth.

Objectivity



Above: *Objectivity: Two people measuring the same thing and getting the same result.*

Objectivity means that two people who perform the procedure will come to the same conclusions.

Objectivity is typically a design criterion: You construct your survey or experiment such that it does not matter who conducts it. This is usually not problematic if you do an experiment or use a multiple choice survey.

In cases where objectivity might be a problem, you can use two (or more) people to analyze the results and just see whether they agree.

Reliability

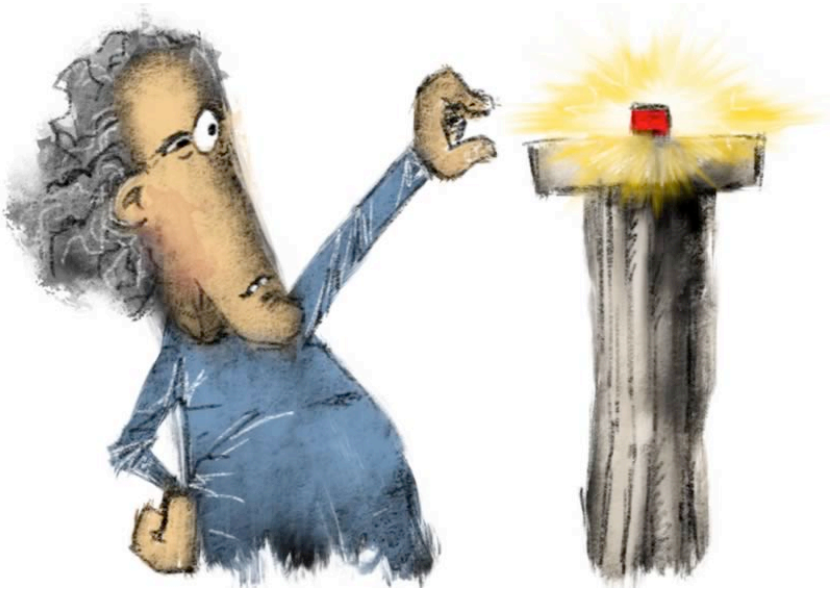
Above: *Reliability: Measuring the same thing twice and getting the same result.*

Reliability means that when you measure the same thing twice, you get twice the same measurement.

Often, it also means that when you measure the same thing in slightly different ways, you get similar measurements. For example, if you use a questionnaire with 10 items measuring social fear, then people should answer those 10 items more or less consistently.

Reliability is the most frequently applied criterion: It's difficult to achieve and easy to measure. Especially for questionnaires, because there is a special (and very easy to perform) procedure to measure scale reliability: Cronbach's Alpha. A high Alpha (close to 1) means that all items of the scale measure more or less the same thing, a low Alpha (close to 0) means they all measure different things.

Validity



Above: *Validity means that you measure the „real thing" (whatever that is).*

Validity means that your procedure measures what it says it measures.

This is the most important aspect, but also the most difficult to check. For example, there is no such thing as „real Intelligence", which you could compare to what an IQ test measures. So scientists usually have to use circumstantial evidence: People with a high IQ are successful in school and business, for example.

Also, it's usually pretty clear what any procedure measures just by looking at it. If you have somebody do complicated mental stuff at speed, you may as well call that intelligence.

One main point to consider is that a procedure can only have high validity if it is objective and reliable. If the procedure does not make reliable measurements, then it measures nothing, so it can't be valid. On the other hand, if a procedure is reliable, then it must measure *something*, so maybe validity is more about understanding what the test measures, and not about tweaking the test until it measures exactly what you think it should.

Example

The criteria are always applied when scientists construct tests, or when they translate existing tests into new languages or taking them to other cultures.

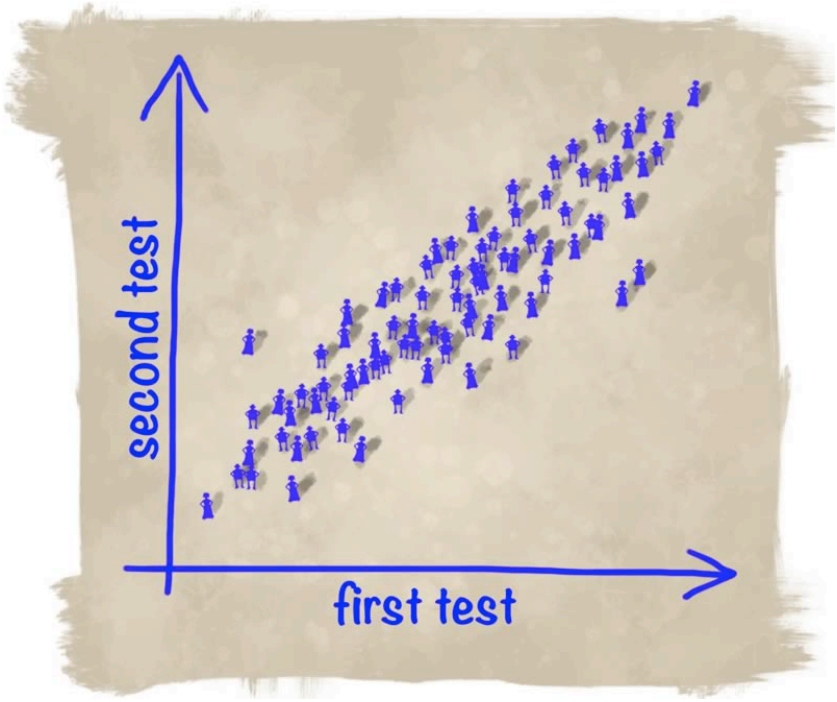
That's exactly what Anatoliy V. Kharkhurin did when he took a language test to the Emirates, in this paper.

Anatoliy's aim was to test the quality of an internet-based vocabulary test. In the test, participants are shown a number of pictures of common objects, and have to type the correct English word for them. The question is whether this procedure accurately measures a person's language ability.

Above: *In Anatoliy's test, non-English-speaking subjects saw pictures of common objects and had to name them.*

Because an internet-based test does not require an examiner, there is no debate about the **objectivity** of the test. You can always nitpick (for example, maybe the test is slightly easier on a fast computer than on a slow one), but for all practical purposes, objectivity is granted.

To test the **reliability**, Anatoliy administered the test to 130 students. Then, 35 days later, he administered the exact same test to the same 130 students, and compared the scores. He found a correlation of 0.83, which is very high (0 is no correlation, 1 is a perfect correlation). That means: The students answered the test in very much the same fashion both times, so the test is highly reliable.

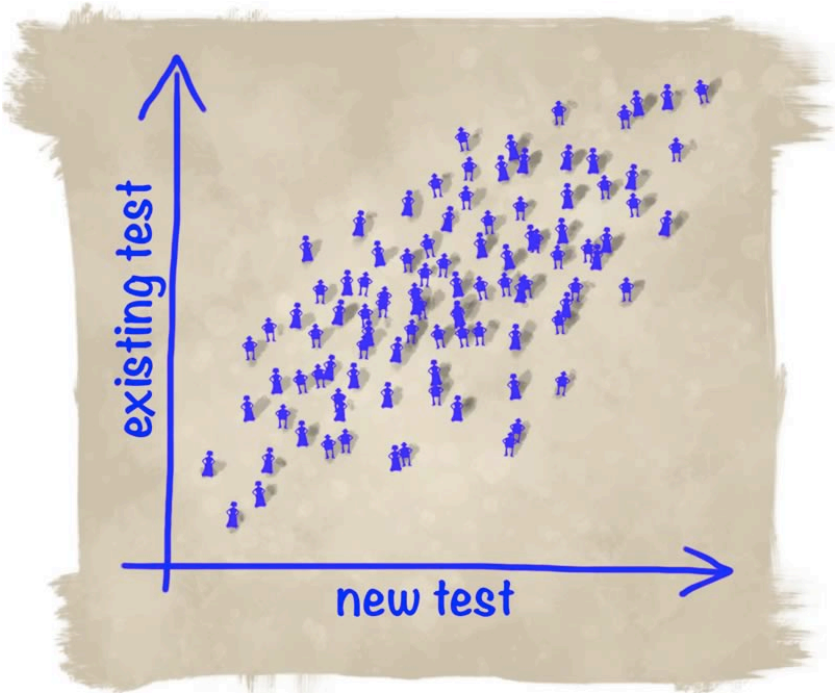


Above: *Anatoliy found a correlation of 0.83 between the first test and the second test 35 days later. This means that subjects' scores were very consistent over time, so the test is reliable.*

To check the **validity**, Anatoliy gave students a set of established tests, and compared the scores between his new test and the established ones. He writes:

> Although correlations between the iPNT and other tests were highly significant, the correlation values ranging from . 52 to .68 indicate that this test provides a partial assessment of the linguistic abilities measured by other tests in this study. – *Kharkhurin, 2012*

Correlations between 0.52 and 0.68 are not bad. They mean that there is a strong connection between what Anatoliy's test measures and what the other tests measure. However, it's far from perfect (that would be 1.0). So it seems that the new test does measure something that is related to language ability (as defined by the established tests), but either not quite the same or not all of it.



Above: *The correlation between Anatoliy's new test and existing tests was around 0.6. This means that the new test often agrees with the existing one, but by no means perfectly so.*

Note that this is one of the few cases where you can accurately determine validity: If you create a new measure for something where you already have established measures. That is rare, however, because most of the time you are creating a test exactly because there isn't one already.

3.3. Sampling

You often hear about „representative surveys“, and it‘s best to get the idea out of your head right now:

- You can‘t have a sample that exactly represents the population.
- Sampling more people certainly gets you better results, but there is no magic number for how many people you have to sample so that they represent the population.
- It‘s certainly not a percentage. Statistical results become more precise with the number of people in the sample, not the percentage of the population they represent.
- Actually, asking people at random is often your best bet to get a sample that represents the population.
- In most contexts, a „representative survey“ is simply a survey where the researchers have made sure that they really ask people randomly (that is, where everyone has the same chance of being asked, no matter where they live, how old they are and what their lifestyle is).

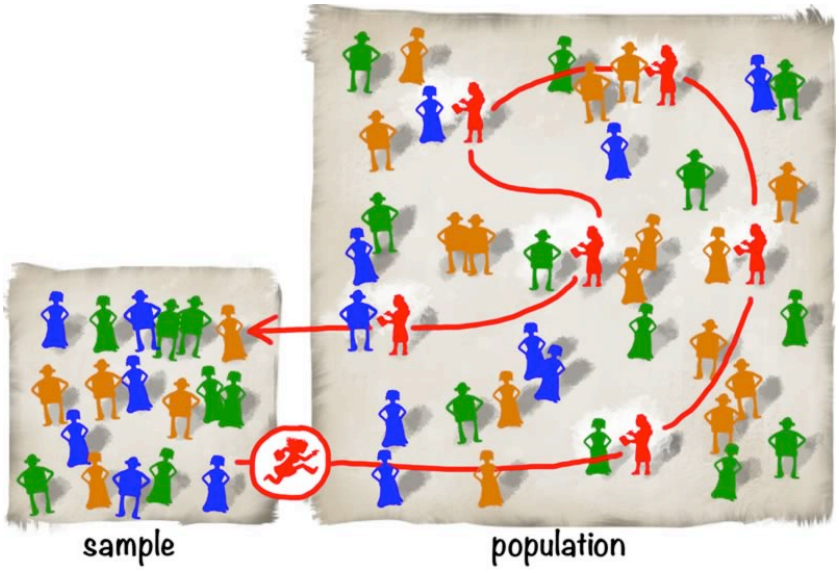
As in all research, there are different methods to get a good sample. Picking people at random is probably the most obvious (and usually the best), but there‘s more.

Sampling methods

First, let‘s look at a few methods for sampling. That is: At different ways to recruit your subjects from the population.

Random samples

Practically all research is done with simple **random samples**. You just get people that have no connection to you or each other.

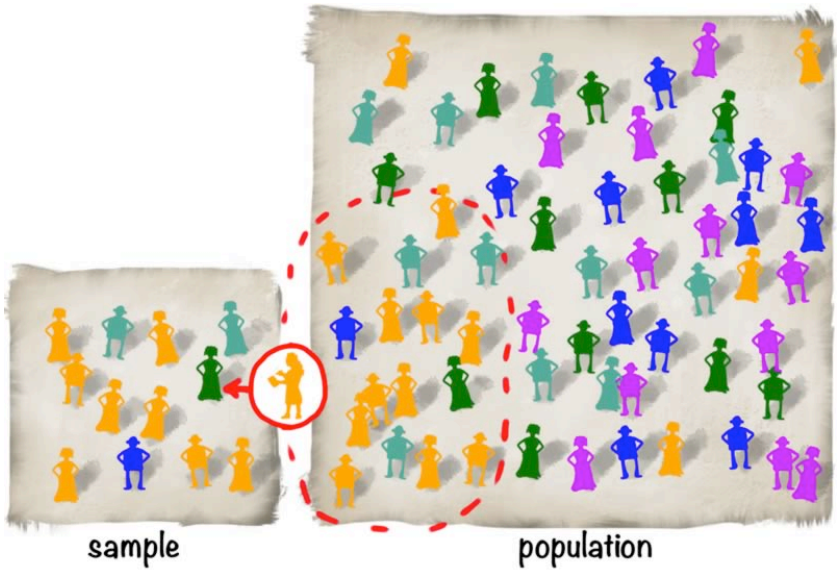


Above: *In random sampling, you just get random people from the population into your sample. It usually involves effort or money to make sure that everyone in the population has an equal chance of being included.*

That is the theory, anyway, and if you do market research (where you can pay subjects), that's basically all you need to know. Now go out and get people.

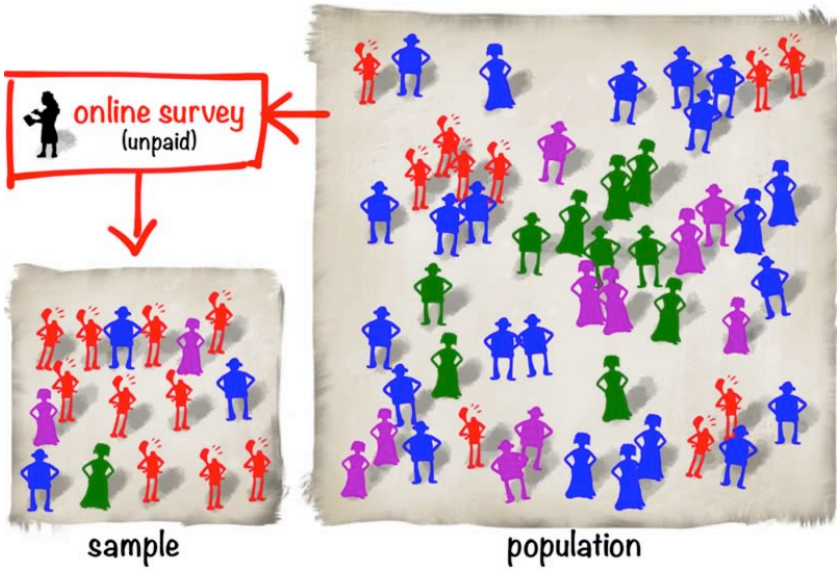
If you do academic research, you'll often have to rely on friends of friends, on the kindness of the people on the street or on students who are fulfilling some formal requirement. That is not *exactly* random, but often it's good enough. Just watch out for the following:

- Don't ask your immediate friends. They're almost certainly a very biased subset of the general population.



Above: *If you use friends and family, then your sample will share a lot of characteristics with you, and not represent the population well.* ***Don't do that!***

- Don't ask people who are knowledgeable in the field you're researching. They will provide expert opinions, not personal opinion. If you want expert opinion, you can read a book.
- Don't just put your survey on the internet or mail it around. You'll get responses from people with an urge to tell you their opinion, which are usually not a good indicator for what most people think. You can do this if you pay for the answers (because then you get replies from people who like money, which is an even representation of the population).

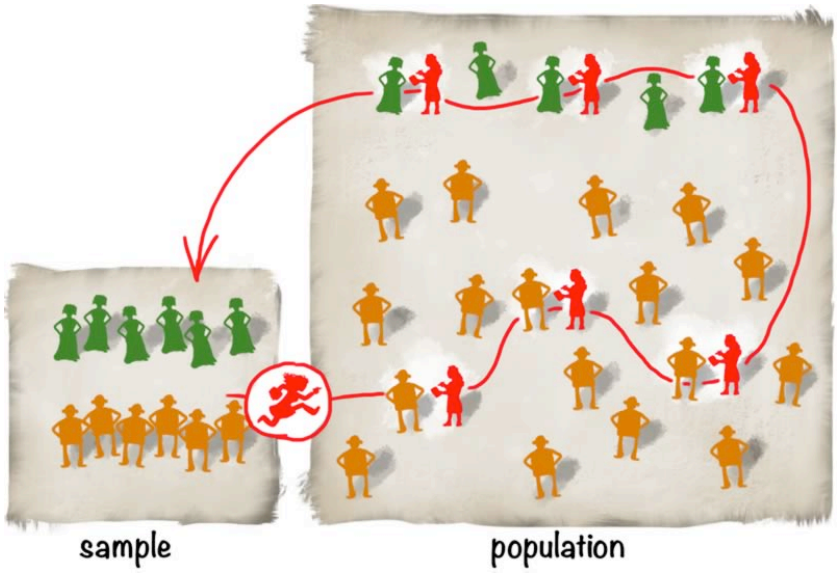


Above: *If you use an online survey, there will be a tendency that your sample mostly includes people who want to make their voice heard. That voice may not be the general opinion of everyone. If you use online surveys, either pay people, or use a recruitment method that creates a balanced sample.*

Stratified samples

Random sampling has only one major disadvantage: Anything that is rare in the population is also rare in the sample.

If that is a concern, you can use a so-called stratified sample. For example, if you want to research how people of different age groups use the internet, you may stratify the sample for age. That means you'll try to get the same number of young, middle-aged and old people in your sample.



Above: *In a stratified sample, you balance the probabilities of including some categories of subjects. The result is that those categories are more evenly represented in the sample than they are in the population.*

As far as I can see (I've never used a stratified sample myself, and know only one person who has), it does not matter whether you go about the selection process more or less methodically. You might go for a set number of people in each group you sample, or just generally make more effort towards the people that are more difficult to recruit.

However, you have to take care that all subjects are randomly sampled from similar groups. For example, you can't compare 20 young university students with 20 seniors from a bird watching club.

Getting back to the population

Once you have defined the sample, it's usually time to re-think what it represents. As already mentioned, no sample represents the population in general. At best, it represents all the people willing and able to participate. Most likely, not even that:

- If you do a telephone survey, it's all the people who are at home and don't have anything immediately urgent to do. Which may mean that students are overrepresented and mothers underrepresented.
- If you do a street survey, it's people who are walking through the city at any given time (and are not in a hurry).

Whichever you do, make a note of what you think your population is and mention it in your research report if it may have *any* effect on the results. Remember that this does not invalidate your results in any way, it just means that they are valid only for most people, and not for all of them.

Selection criteria

Most research is not concerned with all people in the first place, but only with a subset. For example, when you want to find out how young shoppers think about a product, you're only sampling young people.

Other selection criteria might be:

- Males aged 20 to 40.
- Random young people from the street.
- Mothers with young children.
- Long-time customers of your company.
- Native English speakers who have been living in Switzerland for more than five years.

Note that you can always exclude people from your study later on – you may even have to, for example if you discover that they gave only nonsense answers. However, if you do so, you have to make a note of it in your report, such as:

„The sample consisted of 19 shoppers age 21 to 36. One person had to be excluded because there were reasons to doubt the truthfulness of the responses."

Sample size

Now for a crucial and very difficult question: How many people do you include in your sample? The following answers are all valid:

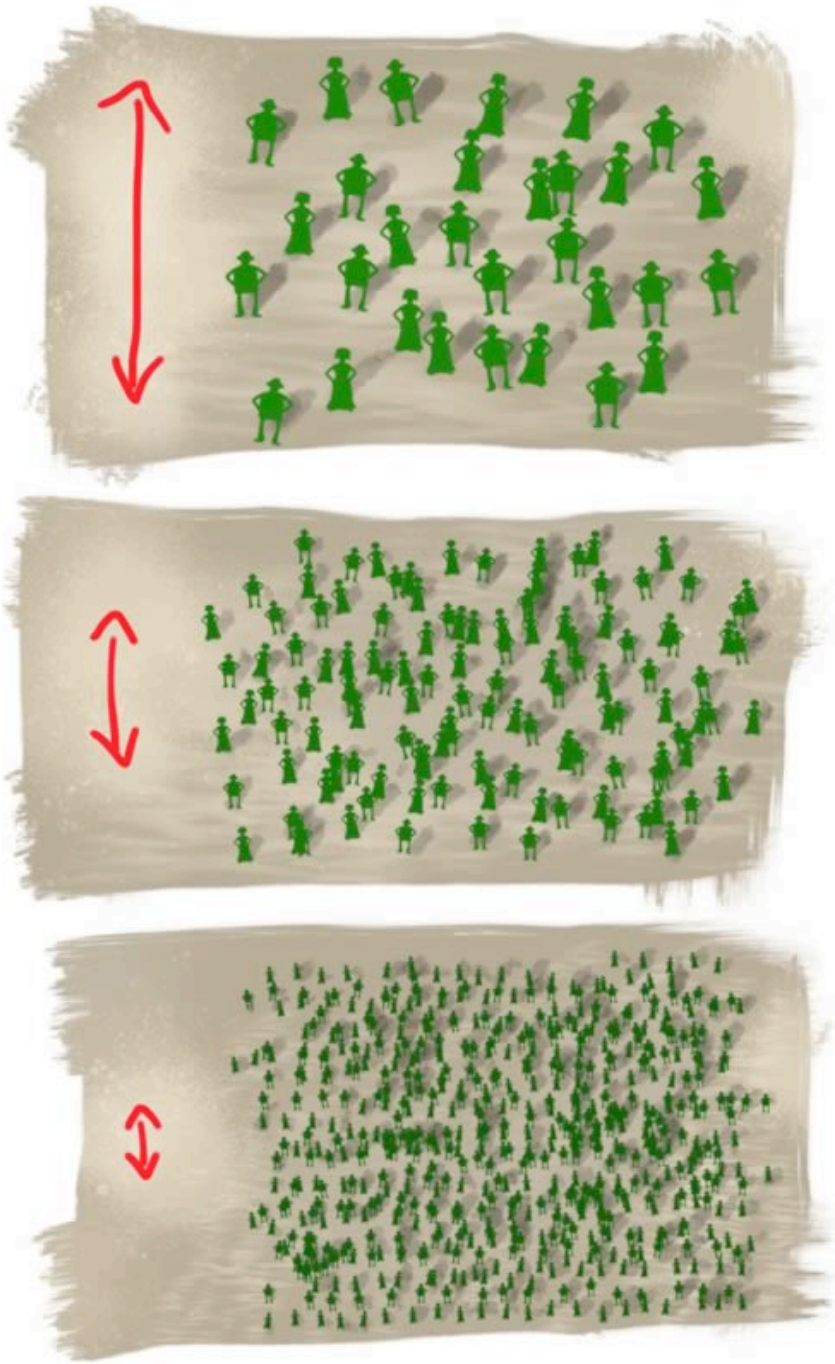
- 20 to 30
- As many as you can get
- As many as the other people who do similar research
- As many as you've computed you need (using special formula and basic assumptions on the size of the effects you hope to find)

More is (slightly) more

There is a simple law that describes what happens when your sample size increases:

Your statistical precision doubles every time your sample size quadruples.

So: Let's assume your usual sample is 30 subjects. This is typically enough to find an average-sized difference. If you want to find a differences half that big, you'll need 120 subjects. A quarter the size, you need 480. And so on.



Above: *The relationship between the size of the effects you can find and the size of the sample you need. Half the effect size equals four times the sample size. In other words: To double the precision, you have to quadruple the sample size. More precision requires much, much, much more subjects.*

If you want 30 times the precision of your little experiment with 20 to30 subjects, you need a sample size that's as big as the world's population.

This is why it's perfectly acceptable to just use the same 20 to 30 people that a lot of research is based on, or whatever other authors in the field have used: Because using more people does not increase your precision all that much anyway, at least not compared to how much it increases your effort.

That said, it's often not much effort to sample a few more people (say, 35 instead of 25), and it's generally a good idea to do so. Not because of the overall precision, but because you may have to exclude people later, or you may want to compare subgroups of your sample.

Computing the required sample size

Because all of statistics is just mathematical equations, you can solve them for any variable. So by twisting the formula around, you can put in the size of the effect you hope to get, and find out how many subjects you need in order to prove a statistical significance for it.

So far, I have rarely used this, mostly because I've felt that the practical use is often limited:

- I can read up how many subjects other researchers have used for similar tasks.
- In order to compute the required sample size, I need to make assumptions on how my data looks. That's easy for textbook examples and difficult if I use slightly more complex tests.
- It's usually more efficient to increase precision by using more complex designs than by increasing the sample size. And increasing the sample size is limited anyway.

However, computing sample sizes has become a fashion lately, in my opinion because of the following reasons:

- It's also called ***Power Analysis***, which sounds good and not too difficult to do. The term means that you check if there is a good possibility of getting significant results before you start the procedure (The power is the probability that you get a significant result).
- It's easy to do a textbook example for very simple applications. So most textbooks have a chapter on the subjects (which then turns out to be entirely unhelpful with real research situations).

- *„Did you do a power analysis?“* is always a valid question to ask researchers, even if you don‘t know much statistics. So reviewers often ask it.

Don‘t get me wrong: Computing the required sample size is a great idea, and you should do it if you can. It‘s just that it will frequently not be worth the effort, because the computation is difficult, it involves a lot of guessing, and it won‘t affect what you do.

There is one area of research where you should be serious about power analysis and sample size estimation: That is medical and pharmaceutical research. When you deal with patients and give some of them a treatment that might improve or worsen their condition, while others receive a placebo, then you want to spend some time figuring out how many patients you really need in your sample.

4. Methods of data collection

This section lists several methods of data collection in no particular order, and following no strict categorization. This is because in practice, you'll mix up the methods for nearly any research you do. For example, when you do an experiment, you'll typically present your subjects with a survey questionnaire as well, or you'll analyze the experimental outcome along with a quasi-experimental question (such as: did men perform better than women).

Anyway, these are all methods you can use. Whichever works, use it. If you want advice, it's this:

- Give the experiment special consideration. Proof by experiment is the best proof you can get.
- Also, give the repeated measures design special consideration, because it's the most efficient tool you have (when you can apply it).
- Do use smoke testing. You'll learn and you'll avoid mistakes.

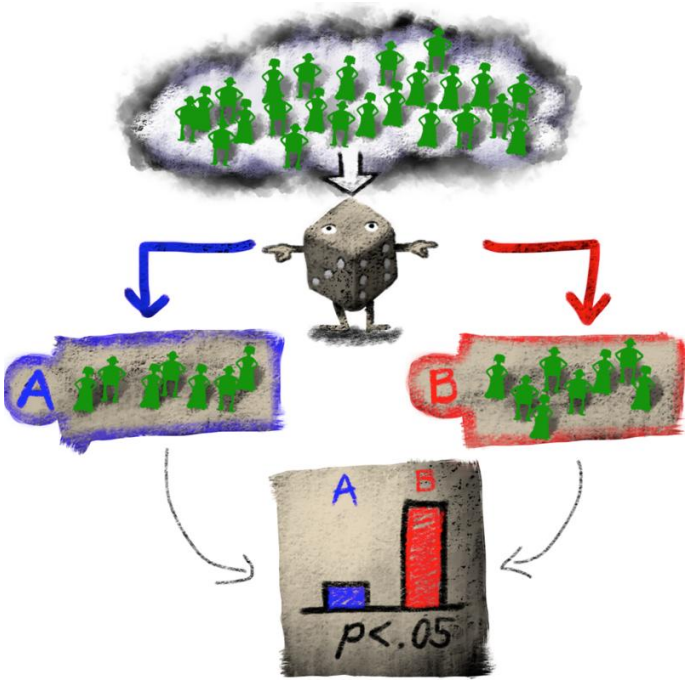
4.1. The Experiment

The experiment is the high road to scientific knowledge. It works as follows:

1. You create two (or more) experimental conditions, let's call them A and B. For example, you could see if subjects learn better when listening to classical music (A) or to rock music (B).

2. You randomly assign half your subjects to condition A and half to condition B.

Why is this so great? – Because in an experimental setting, everything is completely random, with the exception of the experimental condition. For example, if you have a few very smart people among your subjects, they will be randomly assigned to either condition. That does not mean that both groups end up exactly equally smart, merely that any differences between the groups are due to randomness.



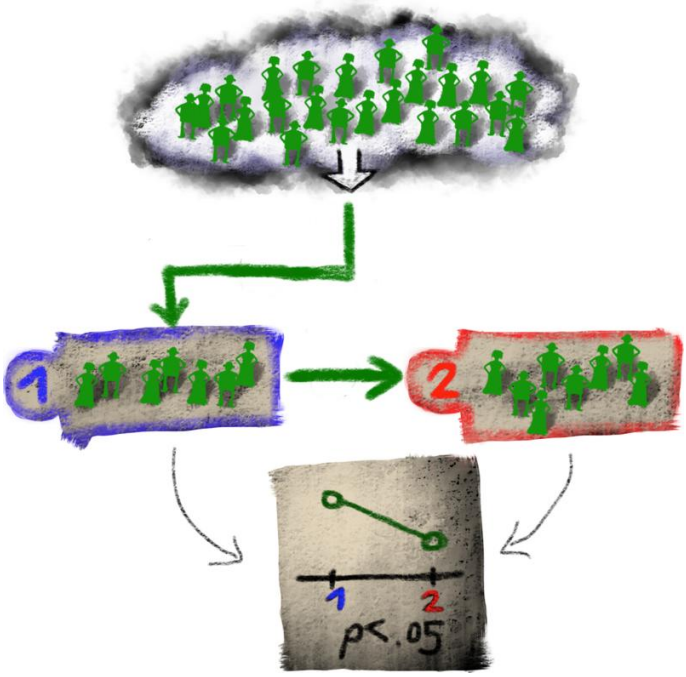
Above: *In an experiment, you randomly assign subjects to different tasks. Because of the random assignment, any group differences must be due to the different experimental conditions.*

That means in turn that all differences that are not due to randomness must be the consequence of the experimental condition. And because you'll use statistical methods to discard any random differences, it means this:

All statistically significant differences you find in an experiment are caused by the different experimental conditions, and nothing else.

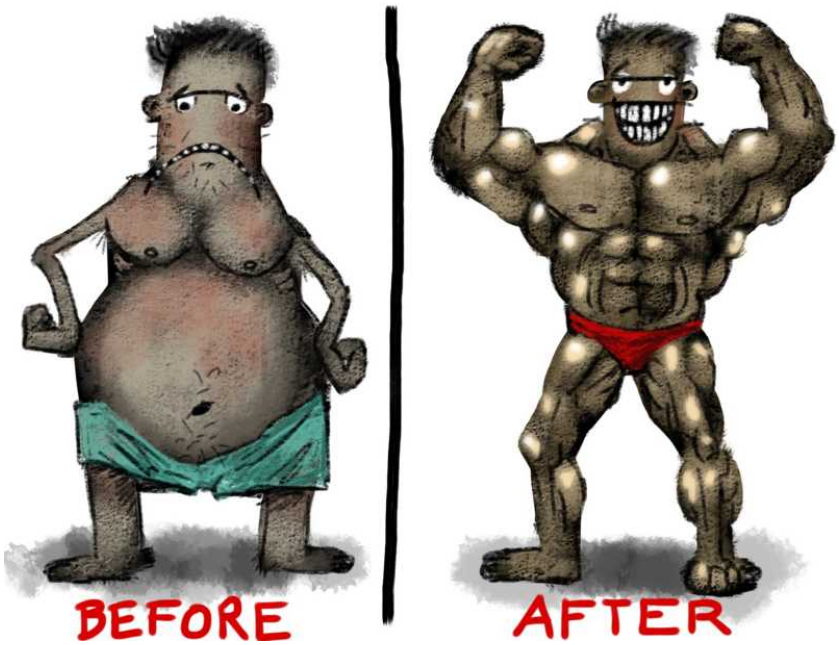
4.2. The Repeated Measures Design

In a repeated measures design, you measure the same variable multiple times per subject. This is the most efficient type of research available: Because you compare subjects with their own past selves, you have very little variation in the sample, and get significant results very quickly.



Above: *Sometimes, you can design an experiment such that the same person performs both tasks. This is a bit more difficult to set up, but statistically much more powerful.*

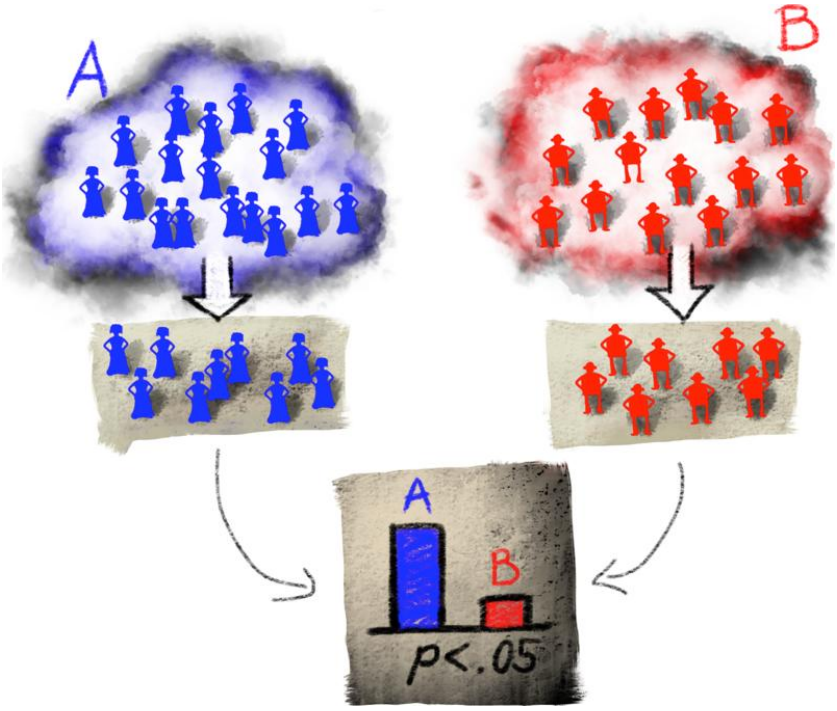
The most frequent application of this is a before-and-after design, which you know from advertising:



This type of advertising is used exactly because it's so convincing: If we see the same person before and after, then we can directly compare, and there are no other variables involved. For example, the above comparison would not be very convincing if it was „somebody who does not use the product" versus „somebody who uses the product". No, the fact that it's twice the same person is what makes this convincing.

4.3. The Quasi-Experiment

The Quasi Experiment uses the same design and procedure as the experiment, but you do not randomly assign who is in which group. A good example is if you compare men and women in a certain task. Because you can't randomly assign subjects to be „men" or „women", it's called a quasi experiment.



Above: *In a quasi-experiment, you compare existing groups in an otherwise experimental setup. The results are not as clear as those from an actual experiment.*

Because you did not assign subjects randomly, you cannot know exactly what causes the difference. For example, when men perform better than women at a Math task, you don't know what causes it: It may be in the task itself, or in society, maybe it's a choice girls make when they decide that other things are more worth pursuing, or it's in the genes (which it probably is not, but you can't know).

Still, the experimental setup means that you usually get good solid data, they're just not quite as convincing than if you use an actual experiment.

4.4. The Observational Study

The observational study is a term that describes any study that is not an experiment (that is, where you don't randomly assign people to the experimental groups).

This includes quasi experiments and much more, such as surveys, tests, population statistics and so on. Basically, any collection of data where you look at subjects instead of having them go through an experimental setup.

Above: *In an observational study, researchers observe subjects without using an experimental setup.*

Observational studies are typically easier to do than experiments. Their results are more difficult to interpret, but usually easier to apply to the real world.

4.5. The Survey

Above: *When you do a survey, you get lots of answers from many different people. It's a great way to gather much evidence in a relatively short time.*

Measuring attitude

The survey is ideal for measuring attitudes. You can quite directly tap people's opinions and there is no catch and no hidden problem. All you have to do is make really sure that you ask one thing at a time and that each question can be understood in only one way.

If you need inspiration about what to ask subjects, there is an easy trick: Go to the internet and read people's comments on a subject. You can get them from forums or (even better) from reader comments on newspaper websites. Let's look at a few example comments I've grabbed form the internet, and see how you can transform them into survey items. I've picked some comments from an article about eyewitness testimony:

„Eyewitness testimony has long been recognized as often unreliable. However it varies from one witness to another and from one situation to another. Identification procedures can and should be improved, but there will always be some errors. The 99.9% certainty level expressed is actually quite high and certainly "beyond a reasonable doubt". However studies show that witness estimations of

certainty in their identification has no correspondence to accuracy....“ – says reader „Bronx Justice“

This could be made into the following survey items, to which your survey respondents can agree or not:

- Eyewitness testimony is often unreliable.
- The quality of the testimony depends on the witness.
- In some situation, eye witness testimony is very reliable.
- When relying on eyewitness testimonials, there will always be errors.
- Witnesses can correctly state how confident they are about an identification.

„i still remember the guy's face that shot me over 30 years ago.“ – says reader „John Dillingerr“

- Memories formed in life-and-death situations are much stronger and clearer than normal memories, and can be preserved over many years.

„The police know this. My pharmacy was robbed and the guy used a 12" bladed knife. The police were called. We were taken in separate vehicles to the police station to give our immediate and individual testimony. The only thing we agreed on was the size of the knife.“ – says reader „Susan Shaffer“

- When examining a crime, authorities should question the value of eye witness testimony.
- Eye witnesses often disagree on important facts.

„Visual memories are funny; they're time series, not static. Even if he tried to change his posture etc., if the prosecution had him re-enact by approaching you and demanding money, your memory of his movements would enable you to pick him out of a "lineup" of re-enactors basically 101% certainly.“ – says reader „jf12“

- Lineups are more efficient than photos for identifying suspects.
- Eye witnesses can trust their gut feeling when confronting a suspect in a lineup, even if they do not have accurate memories.
- When eye witnesses have been able to observe a suspect from up close and for more than a few seconds, they can identify the suspect with 100% certainty.

And thus, from scrolling down one comment section, we have a survey on people's attitude towards eye witness reliability. All that is left to do is to format it properly and add answer boxes.

By the way, if you were wondering: The scientific consensus is that eye witness testimony is horribly unreliable. Our brain is not good at learning when under stress, so it will not record much information while your life is threatened (or the contents of your wallet). However, after the event, eye witnesses spend a lot of time thinking about it, and form very strong memories based on the little information that is available.

Measuring behavior

It's a bit tricky to get good behavioral data with a survey. If you're not careful, you'll measure how people think they behave. As in all research, the following two rules of civilized communication apply:

1. People want to help you and appear nice in general.
2. People won't lie to you.

This means: Do ask questions that people can (and have to) answer honestly. Avoid questions where they have wiggle room and can nudge their answers into what they think is socially appropriate, what they think helps you, or what they think makes you like them. Specifically, construct your items according to these rules:

- Rule number one: **Get rock-solid data**, such as: How much did you pay for traffic fines last year, how many cigarettes do you smoke a day, do you give money to beggars, did you read the horoscope yesterday, how often do you go shopping in an average week, what was the biggest purchase you made last year, how long is it since you last tried to learn a new language, and so on.
- Rule number two: **Get lots of it**. If you want to research shopping behavior, ask lots of questions about shopping. For example: How often do you do it, how much do you spend, how long does it take, do you do it alone or with others, how often do you buy things you don't use afterwards, how often do you hesitate because you can't afford something you want, and so on.

Remember: It's not important that you ask deep questions that cover the essence of shopping and the human soul. Ask questions that have clear answers, and ask lots of them.

Once you have the data, you can use statistics to tell you which questions are connected with each other and which are just random data points in the desert.

Writing great items

There are two basic formats for questions (or „items"), and both work well. You can pose them as actual questions or as statements, to which subjects agree or disagree (in various strength):

- **Question**: Do you think that French is a beautiful language?
- **Statement**: French is a beautiful language.

I have a slight preference for statements, as they are typically a bit shorter, easier to understand, and I find them easier to answer truthfully. When I hear or read a statement, some automatic filter in my mind just automatically tells me if I agree with it or not. However, questions follow the natural flow of conversation better, so you may prefer to use them, at least for the first few items.

Sometimes, asking people to decide between alternatives can work, too:

- **Decision**: Do you prefer the sound of French or Russian?

This is usually tougher to answer, but it also makes sure that you get a good answer. In a decision, subjects are forced to make a call, and if they are biased (for example, because they think you have to like all cultures equally), the bias often applies to both sides of the decision.

It is usually good to include some decision questions in your survey, because they create rock-solid data that has very little room for interpretation.

There are a great many formats for answers, and none are perfect. There are two considerations for the answers:

- How many answers you offer.
- The range you include in those answers.

Number of answers

When you give a multiple choice question, you can give any number of boxes for the answers, so we'll go through all options one after another:

1. One box means that participants just check the boxes that apply to them, so it's very fast, but when they omit an answer, it'll count as „no".
2. Two boxes have the same content, but you'll see whether somebody answers „no" or leaves out the question entirely.
3. Three is a good number for simple, clear questions.
4. Four is the right number when you want to force people to make a decision (as in, „rather yes" and „rather no"). As such, it's great for telephone interviews.
5. If you use five boxes, there is a danger that people go for the second and fourth option, because they want to avoid the middle and the extreme.
6. Offering six boxes also means that people cannot choose the middle – which is a bit cheap given that you offer so much choice in the first place.
7. Seven items (or more) are cool if you have a paper survey and want to make sure that there are ample options.

Range of answers

As a rule, you want to restrict answers to the realistic range, so that average people can give all answers more or less equally.

To be totally honest, I'm running a bit late in writing this text, but fortunately this is not so difficult. It just means that

when you write the items, don't think of including all the answers that are technically possible, so you **don't** have an item like the following:

Question: How good are your driving skills?
A) Perfect
B) Good
C) Mediocre
D) Bad
E) Non-existent

This practically forces people to answer B, good. It's much better if you just include what people might realistically answer. A much better scale would be:

Question: How good are your driving skills?
A) Good
B) Fairly good
C) Average
D) Slightly insecure
E) Insecure

Note how this scale tops at „good", because people are reluctant enough to give themselves the top rating, and they won't do it at all if it's too high. Also, the below-par items are worded carefully, so people are encouraged to check them (even though technically they're not an exact continuation of the scale).

Often, you can improve answers if you ask people to rate themselves in relation to others:

Question: How good are your driving skills?
A) Better than those of my friends
B) Slightly better than those of my friends
C) About the same as those of my friends
D) Slightly below those of my friends
E) Worse than those of my friends

However, if you really want good data, ask about behavior. It does not matter if it's perfect or not, behavior questions are always good, and it's always surprising to see how they compare with the attitude data:

Question: How many traffic fines did you get last year?

Essentially, a solid far-from-perfect behavior measurement is easily worth as much as any attitude question, so use it. Actually, use both, and then compare what people think of themselves and what they really did lately.

The anatomy of a great survey

A good survey has an anatomy, that is, some things are first, some in the middle, some at the end. We'll go through each part one after another.

Introduction

Briefly introduce yourself and the survey. **Do not give people any hints about what you're trying to find out.** Tell them you'll answer all their questions afterwards.

If you work for a university, you should say so in the first sentence, so people know you're legit. Also, if you can offer a reward, say so early, before people can form a negative opinion towards you.

Warm up

In a street survey, find out first if your subject fit the selection criteria. Typical questions are: „Where do you live?", or „What is your mother tongue?".

Try to start with some easy, but compelling questions. Things you are interested in, and things people will like to tell you. Also, you can use **framing** (in psychology, that's not something negative): If you do a survey about shopping, ask people a question or two that has them recall actual shopping experiences. This will improve the quality of the answers for the rest of the questionnaire. It will get people to answer based on their own personal experience, not based on what they saw on TV or have heard friends complain about lately.

Questions

This is the main body of the survey, and should last for anything between 5 and 30 minutes. Generally, people are prepared to invest some time – if not, they'll simply not participate.

The all-important question is: How do you group all these questions. There are two ways to do so:

- By topic
- By answer type

It's usually tempting to group the questions by topic, but there is no real reason to do so. Try to group the question by answer type, so subjects may start with a few yes/no questions, then some statements to which they agree on a scale from 1-5, and so on. This usually speeds up the process a lot. Also, mixing the topics gives a fresh feel to the questions and makes it more difficult for subjects to guess what you're up to.

Sometimes, researchers deliberately mix up the questions, just so subjects do not build some overall attitude on which they base all their answers. Also, it's usually a good idea to switch the polarity of the items. This means, sometimes „good" means more points, sometimes less points. This makes sure that people read all the questions carefully, and not just fall into a general answer tendency.

Once you have the blocks of questions, order them strategically. There is no real method to it, just see what you'd like or dislike. For example, you probably don't want lots of complicated questions up front.

Questions that are boring or intimate

If you have any questions that people probably do not like to answer, place them at the end. By this time, subjects will trust you, probably like you, and in any way have invested enough time and effort to not jump off immediately. Such questions include:

- Anything that is long and boring, such as: „I'll read a list of publications and you tell me which ones you read, and how often".
- Anything intimate, such as relationship status, income, or age. Note that most subjects will slightly hesitate, not because they don't want to tell you, but society trains us to not straight-out answer these.

Cool down

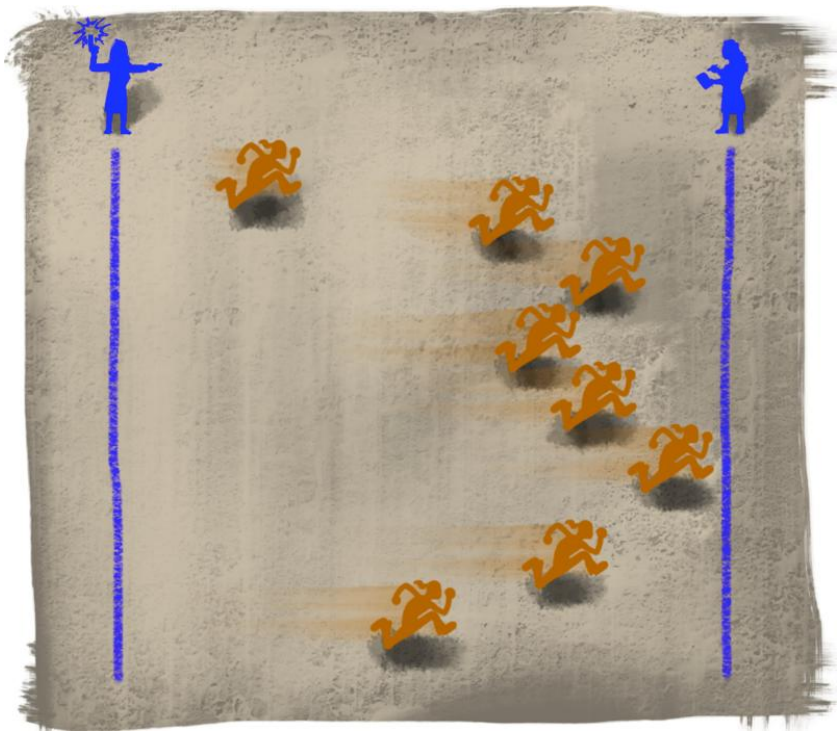
After all is said and done, thank your subjects and offer to answer any questions they have about the survey. Depending on the content, it can be smart to ask them whether they can think of anything important that you forgot to ask. Also, you can offer to send them the results.

Making your life easier

A few things can help to make your life much easier:

- Write down everything you mean to say on a sheet of paper or on the survey itself. It's much more easy to be natural when you know what you want to say.
- Number all the questions and answers. This will make it much easier to put them into your statistics program.
- Try out the questions and answers on your friends. Time how long it takes. For a very quick test, read the questions and answers aloud.
- If you need a reminder, write SMILE somewhere you can see it while you conduct the survey. You'll get much better results if you occasionally smile

4.6. The Test



Above: *Tests check how people perform when they really try. They're a great way to find differences.*

Science is all about getting reliable data. And one way to get them is to tell people to perform a task as best as they can. Which is what they'll usually do anyway – most people want to impress, given the possibility to do so.

Also, tests are a big part of our school system, so participants will be immediately familiar with a test situation. In fact, it may be easier to tell a participant to solve some Math problems than to rate a list of statements on a scale from 1 to 10. Simply because they've been in hundreds of Math tests already and know what to do.

Tests are anything (long or short) that asks for an effort, such as:

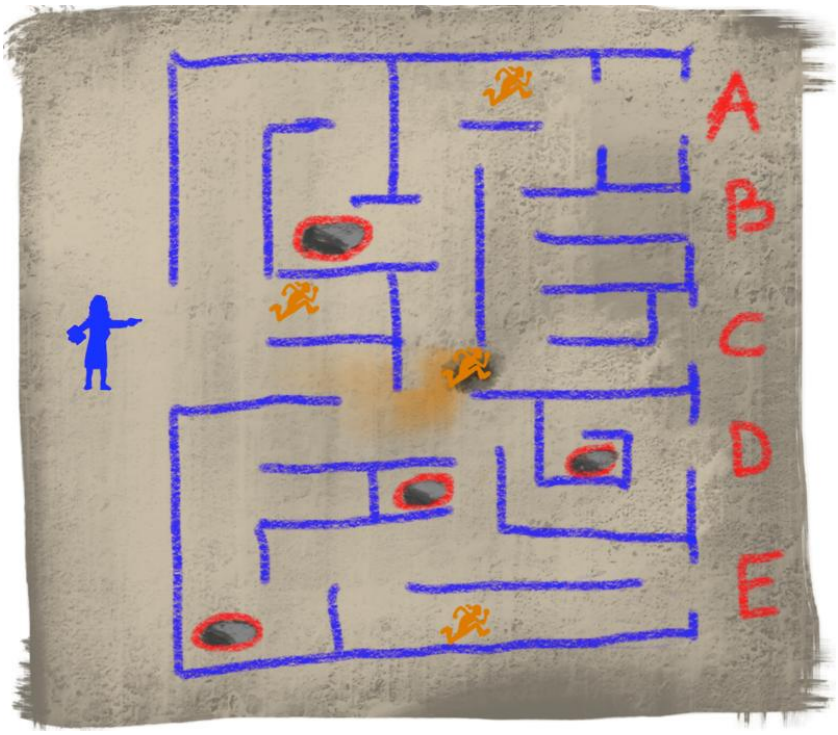
- Name as many French nouns as you can.
- Remember as many products as you can, for which you have recently seen advertising on TV.
- Name as many animals beginning with the letter „M" as you can in one minute.

- Draw a line that is exactly 10cm long.
- Answer some questions about a short movie you have just seen.
- Can you name five Swiss politicians?

You can also ask people for top behavior they have shown in the past.

4.7. The Usability Study

The idea of usability testing is quite new. It started some time in the eighties and has grown widespread only recently. When I was young, engineers would just try to pack as many features into a computer program as possible, and adverts for programs would feature long lists of all the things you could do with a program (90% of which you'd never use because you never bothered to figure them out).



Above: *In a usability study, subjects are given a number of tasks to perform and goals to reach, and then you watch how well they do, what dead ends they meet and what obstacles they encounter.*

More recently, engineers do usability testing on almost any software that goes out. That is, they take actual humans and use them to test whether a piece of software is in a state where those actual humans can use it efficiently.

For most tests, you only need around five subjects and a computer. It's usually a good idea to have some sort of recording device ready (anything will do), if only to play it back to the programmers, who won't believe the results otherwise.

If you have any experience in empirical methods, doing usability is fairly easy. And this is how it works:

1. Come up with a number of tasks that you want subjects to perform. They should be relevant to what people would normally do with the software (or website) and require as little extra knowledge as possible.

2. Sit the subjects in front of the computer and give them the tasks. If you make a recording, inform them and get consent. You can take pretty much any subjects except programmers, web designers and usability experts. If you can, use students and academics, because nothing motivates programmers as much as seeing smart people fail on stupid tasks.

3. Tell subjects to think aloud, and tell them that you won't answer any questions. Tell them you're testing the software, not them, so it's okay if they are unable to complete all the tasks the first time around, and that you are interested in seeing what they think is the right thing to do.

4. Now, sit back and watch, and never do anything except 1) reminding people to think aloud, 2) telling them that you won't answer questions, and 3) give them general encouragement to continue trying if they seem stuck.

5. When they're done, thank them and answer any questions they might have. Note any general feedback they give.

It's best to write down a protocol of the things you plan to say. You'll get used to it very soon, and it makes your life much easier. For example, this can look as follows:

1. „Hello, Mr./Mrs._____. Thank you for participating. I will ask you to perform a number of tasks on our website. Your aim is to complete them to the best of your abilities. We are interested to see how easy or difficult these tasks are on our new website, so I will not give you any indications on whether you are on the right track or not. Do you have any questions?"

2. „Great. One more thing: Is it okay for you if we record your actions and anything you say during the session?"

3. „Okay, so your first task is: Please find a hotel in Amsterdam that is in 10 minutes distance from the

Rembrandt house. Can I ask you to think aloud on what is going through your mind when you do this?“

4. *Rules for feedback:*
   When asked about the task: „Your task is to find a hotel in Amsterdam that is in 10 minutes distance from the Rembrandt house“.
   When asked if an action is correct: „Please continue.“, „I cannot give you feedback“, „Go on“, „It is important for us to see how you do on your own, without my feedback“.
   When a subject becomes silent: „What are you thinking right now?“, „Please think aloud“, „Can I ask you to think aloud?“
   If a subject seems stuck or wants to give up: „Please continue“, „Go on“, „Please try to complete the task“.

5. „Okay, now your second task is: Please find a hotel in New York that has a room for two people available. You‘ll be arriving in New York on June 12 and departing on June 14.“

(and so on)

Usability tests are extremely efficient because of the following reasons:

- You do not need many subjects because you are looking for problems that many people will have, and not for statistical values. You can get valuable insight even from testing one person, and will find most of the large problems by testing 5 subjects.
- Any problem you find will really improve the efficiency of the software or website, so hundreds of people will benefit from it later.

If you‘re interested, read Jakob Nielsen‘s website (and his book). He‘s the pope of usability and one of the very few people who get most predictions right.

Note that there are also some usability methods that are much more demanding, but they are used only rarely, and only by big corporations.

4.8. The Smoke Test

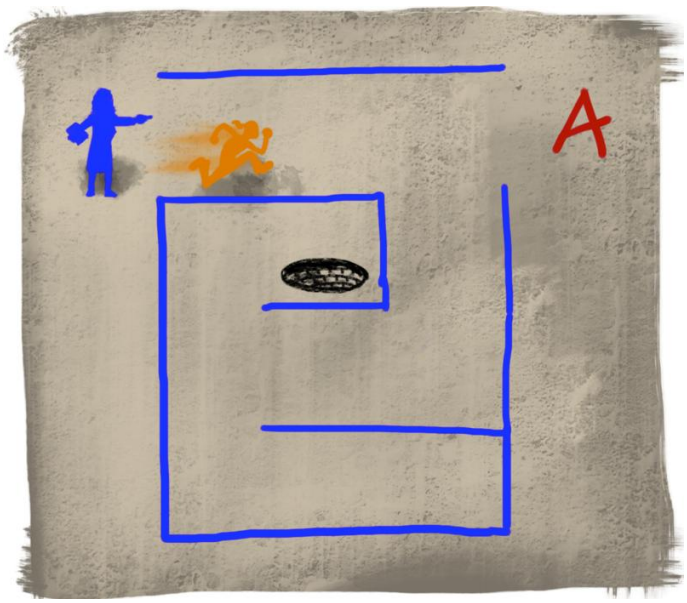
In engineering, a smoke test is a quick test to check the basic functioning of a component:

> „You plug in a new board and turn on the power. If you see smoke coming from the board, turn off the power.“ – *Kahner, Bach & Pettichord (2002).*

I‘m not aware of an equivalent term for empirical research or social sciences, but every other industry uses the term „smoke test“ in one way or another, so I‘ll go with it, too. My definition is as follows:

In a smoke test, you show a real person some of your material and see if they get it.

„Material“ meaning anything you plan to communicate to a wider audience, such as a corporate slogan, a diagram, or a research questionnaire. „Real person“ meaning anyone not involved in designing the material.



Above: *In a smoke test, you give one person one simple task and see if they can do it. If they can‘t, then the task was not simple, except to your eyes. If it‘s important (such as a marketing campaign), you can always use more people and more sophisticated measures.*

This area is usually under-researched, mostly because of the following factors:

- Your ego.
- Your guts.
- The consequences.

That is: Your ego tells you that your material is great. It takes some guts to think otherwise (or tell your boss otherwise), and then some more guts to ask strangers what they think about your material. And if it turns out that your material does not work, then that may mean that nobody understood what your corporate website was about in the past six months, or worse.

So: If you work in an environment where you create materials that other people need to understand, do yourself a huge favor and ask somebody if they do understand it or not.

If it helps, then let me tell you that *all* comedians do it. They're not funny by nature, but they can think up potentially funny material, and they have the guts to ask somebody they trust which of the material actually works. For example, I occasionally do cartoons, and when I started out, about 50% of my cartoon ideas were any good. Now it's more like 90%. Still, I always ask somebody I trust before I publish a cartoon.

And it's always tempting to skip this step, because somehow it's easier to show something to the entire internet than to ask one person face to face.

4.9. Mixing it up

Some methods of data collection are very sophisticated, others are very simple and require almost no effort (apart from getting the subjects to participate).

Do not forget about those effortless methods just because you focus on the sophisticated ones. They are basically free extra data, and can be extremely useful to contrast with your sophisticated data.

- For example, if you use an experiment to find out which type of advertisement has the best recall, then ask people about what types of ads they like, if they like them at all, if they think they're visual types, and so on.
- If you use a survey that has a ten item scale that measures objectively and based on exact behavioral measures how friendly people are, then ask them also in plain text how friendly *they think* they are.
- If you do a psychometric test that measures language ability, ask people about their school grades, whether they read books, whether their job involves writing, and so on.
- Always make sure that you measure both behavior and attitudes, if at all possible. This is almost guaranteed to provide interesting (read: diverging) results.

5. Getting the data into a statistics program

This chapter could be very short, because there is only one strict rule:

All the data from one subjects goes into one row of data.

Let me show you:

ID	age	sex	before	after
J.D.	34	2	35	98
A.T.	25	1	24	26
J.J.	22	2	18	36
A.S.	63	2	45	49
D.F.	51	1		52
W.Q.	33	1	40	58

Above: *When entering data, the data of one subject goes on one line. If you have missing values, just leave the cell empty. Your statistics program will know how to handle this.*

Then, there are a few rules that make your life much, much easier (and reduce the amount of errors you make).

Number any sheets of paper you get back

If you use a paper survey, then put a number on any survey that is completed. Or an ID number (such as the initials of the person conducting the survey plus a number). Do this with a fat red pen on the front page, so it's visible, and type the ID into the data sheet.

Why this helps you: Because somewhere down the road, you'll find a number in your data sheet that is wrong or just weird, and you want to check with the original survey. If by that time, your surveys are a large stack of paper you can't tell apart, then you're in trouble.

From experience: Tracking a single answer in a stack of survey responses is a pain, especially if that value probably does not exist there in the first place, and you find a few more questionable data points while searching.

Make a note of everything (preferably on an empty survey)

While typing the data into your data sheet, put an empty survey sheet next to you for note taking. If you have followed my earlier advice, then the survey itself should already contain notes on how to translate the checkboxes into numbers. If you did not, then mark them on the survey now.

Why this helps you: The worst thing you can do while entering data is to switch around the codes, for example if you code all men as „1" and women as „2", while your friend does the reverse. So: Create a sheet that contains everything anybody needs to enter the data, and which answers all the questions that come up.

From experience: Noting codings on the survey is better than putting them on post-it notes with which you frame your screen (I've tried both).

Translate responses one-to-one (and transform later)

Sometimes, you know that you'll have to translate responses at some point. The most frequent case is if you want all responses to go in the direction of „more = better", and have a few questions that you need to turn around for this.

However, while entering data, just type everything as it is on the survey, as closely as possible.

Why this helps you: It'll make data input much faster, and statistics programs are very good at changing data later.

From experience: If you ask yourself halfway through whether you have correctly reversed every number before input, then it takes some work to check this. Also, when typing data, you want to be able to put your brain into auto pilot. Data input usually does not take huge amounts of time, but if you have to spend a few hours trying not to miss the same five items on a hundred possible occasions, it becomes work.

Use long variable names (or labels)

When I started out, statistics programs allowed a total of 8 characters for variable names. So variables had names like „amda1141", which read like:

- a: Time point (a = before, b = after treatment)

- md: Marlington's Depression Inventory
- 114: Item number 114
- 1: Checkbox one (because the item is multiple choice)

Now, modern statistics programs usually allow you to use longer variable names, and you should by all means use them. So a perfectly acceptable variable name is:

before-mdi-selfmedication-alcohol

This is much easier to read. If you have not guessed already: Does a subject drink alcohol to combat effects of depression. Which, by the way, is a horrible idea. If you suffer from any disease, get professional help (you can even look up what the available treatments are in Wikipedia). If you have *any single mental problem*, there is a good chance that effective treatment is available. If you have any mental problem plus substance abuse, treatment is more difficult.

By the way (and returning to the subject we're actually discussing): Often, a good name for the variable is the full wording of the survey item.

Why this helps you: A lot of your time in data analysis will be spent selecting the variables you want to analyze, and then looking at the output. If you need to consult your notes every time, *everything* takes longer.

From experience: In most projects I do, I help somebody else analyze their data. As such, it's always good if I can understand which variables are which.

Document your data

Okay, to be honest, practically nobody ever does this. Which is usually okay, if you use a survey and nobody ever uses the data again. However, if you think that there is any chance of anybody doing further analysis on your data, or if you work in a large-scale project that spans multiple research centers or generations of students, do everyone a favor and spend an afternoon documenting what everything represents in the data set.

Why this helps you: Not much at all, actually. But everyone after you will be extremely happy (or extremely not on the brink of desperation because they have to get results from numbers they don't understand).

From experience: I have worked on a side project to a longitudinal study that followed kids through their school career and tracked if they were bullied. That's huge, because you can see if kids who are bullied early will develop a career of being bullied, or not. So you know whether to make interventions early or how concerned you need to be when your kid does not get along well in school. Except that we did not know with any confidence which data was from which time point, and it seemed that nobody else knew either. Finally, I spent a few hours finding out which data belong together (that is, which questions were answered by mostly the same people, so they probably belong to one time point). Then I compared this with the number of children that were examined at each time point (which was also faulty, but hey) and finally knew what was what.

Insta-analyze the data

Once the data is in the statistics program, get yourself an empty survey (or print one out if you did an electronic survey) and get some basic statistics: Means, number of responses per answer, and so on. It's pretty obvious, but statistics programs offer no one-click obvious analysis function, so you have to do this manually.

Then, just put the rough elementary statistics into the paper survey using a pen. It's low tech, but very effective (especially because you're already familiar with the survey).

The more advanced version is to create a correlation matrix for all (yes, all) pairwise correlations. This results in a huge table that tells you if any two variable are connected. This is a very fast way to see where effects are and to rule out alternative interpretations for the effects you find.

Why this helps you: You'll quickly become familiar with the data, and you have your quick look-up for all basic data.

From experience: Bosses always want to see results, so it's great if you have them fast. Also, they are not so familiar with statistics programs (and sometimes not familiar with what you were doing, too). Showing them a survey with all the answers written in is a great way to give them first results.

As for the correlation matrix: I usually print it out as small as possible and then glue all the pieces together (it never fits on one page anyway). I've created some that were several meters long. In turn, I could immediately comment on any

hypothesis my professor had regarding which variables influence each other.

6. *Descriptive statistics*

Descriptive statistics are methods to describe your data. That is: You try to use an image or a few numbers to give people an idea of what your data is.

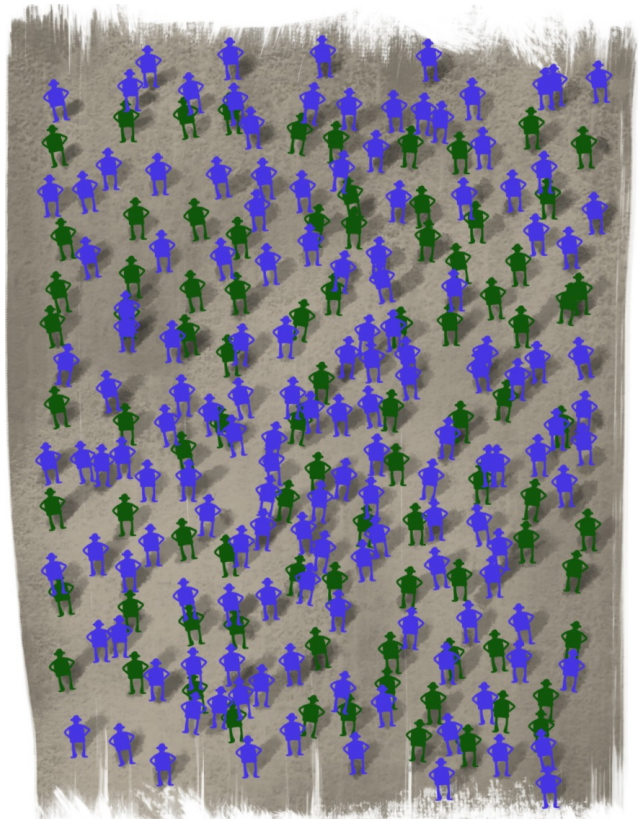
You'll see.

6.1. Frequencies

The most basic operation you can do is to count which values occur how many times. For example, you may find that your sample contains 27 women and 31 men. That is already your first frequency analysis. It's nothing spectacular and probably does not feel a lot like you're doing statistics, but it is a totally valid analysis.

You typically find frequencies in the sample description, as here in the article by Yoram Barak, Moshe Tishler, and Dov Aizenberg:

> 223 physicians (131 psychiatrists, 92 primary care) completed the survey.



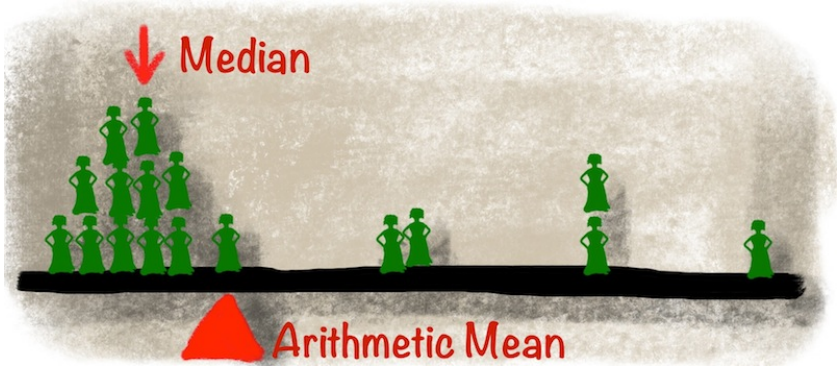
Above: *that is what 131 psychiatrists (blue) and 92 primary care physicians (green) look like in one picture. As you can see, in this case, the numbers are easier to understand than the picture.*

6.2. The „Middle“

Funnily enough, the term „middle“ does not have a precise mathematical meaning. For statistics, there are two preferred ways to describe the middle of a variable:

- What you know as „middle“ or „average“ is the **arithmetic mean**. It‘s used extensively in grading and in communicating statistics to people whose only experience with statistics is the grade point average.
- The **median** is defined so that 50% of the values are below and 50% are above it.
- (There is also the **mode** – it‘s the value that occurs the most. It‘s never actually used, because it is neither any stable nor necessarily in the middle of the distribution).

Overall, the median is more robust, more democratic and often more representative than the mean. That is because the median is not affected by how far above and below it the values are. As such, it is also unaffected by single very large values.



Above: *If you imagine a scale as a seesaw, then the arithmetic mean is the point where the seesaw is balanced. It is more affected by observations that are far out. The median, in contrast, is always in the middle of the observations.*

In short: The arithemtic mean is the value you‘re familiar with and it works well most of the time. The median has a lot of advantages, so ***do use it***. Really! If you feel uncomfortable, then at least compute both values, and see if they are different (which typically means that the median is right and the arithmetic mean is wrong).

The **only reason not to use the median** is if your scale has very few levels. Let's say you use a questionnaire with scales from 1-4, nothing in between. Because the median is always a scale value (or, very rarely, the value exactly between two scale levels), all scale medians will have very few values. Actually, most of the medians will be either „2" or „3". In this case, do use the arithmetic mean, because it's more precise and because with limited scales, there cannot be any outliers that bias the arithmetic mean.

Examples

As an example, I use this study from Suhaila H. Khan about maternity health costs in Bangladesh. Suhalia examined how much money women had to pay for various services in maternity wards. The costs are grouped by expenditures that the women had to pay, such as travel and a hospital admission fees, and medical expenditures, which are provided free of charge.

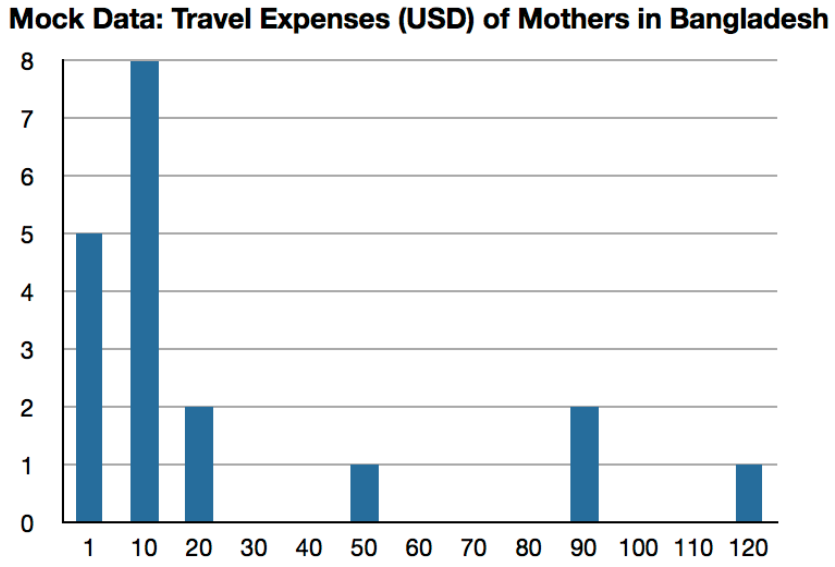
	Expenditures on items NOT supposed to be provided from hospital	
	Travel	**Fee**
NVD (n = 19)		
median	12.50	0.25
mean	29.72	0.23
range	1–127	0.10–0.33

***Above**: Health costs (in US Dollars) for travel and hospital admission fee for women giving normal birth (NVD = normal vaginal delivery) in Bangladesh.*

The first column shows the cost of traveling to the hospital – which is a fairly large factor in Bangladesh and often requires families to sell jewelry or take up a loan. As you can see, the mean travel cost is $29.72, the median is $12.50. This means that half of the patients pay less than $12.50, half pay more.

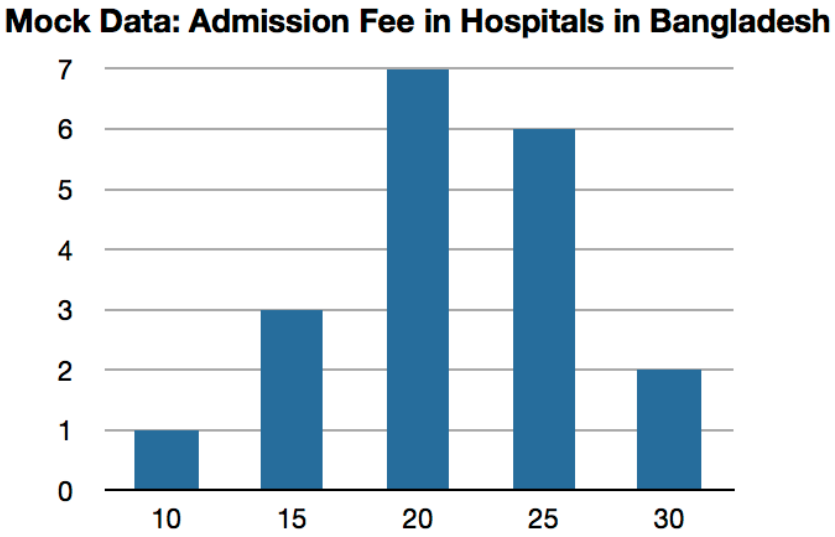
Mean and median are widely different, because some mothers pay up to $127 (as you can see in the „range" row), while you can't pay less than zero (and in fact the minimum recorded is $1).

To show you better, I have created a mock histogram. The actual data is made up, but it gives the same mean and median as the study:



Above: *An example with mock data that gives the same mean, median and range as in the study. As you can see, the mean of $29.72 is largely meaningless – almost all mothers pay less, and only a few mothers pay much, much more. In contrast, the median of $12.50 correctly shows what most mothers pay.*

The second column in the table shows the admission fee. From the range, we can see that this is fairly similar for everyone: between 10 and 33 cents. Consequently, the mean and range are also similar, at 23 and 25 cents, respectively.



Above: *If the distribution is roughly symmetric, mean and median are close to each other, and both adequately represent most of the population.*

Let's look at some more examples from the same study. This time, it's about items that are supposed to be provided for free, but which patients or their relatives still had to pay for (or did pay for voluntarily).

Expenditures on items supposed to be provided free from hospital

Medicine	Food	Tips	Other	Tests
11.25	0.88	1.25	3.88	0.00
18.35	2.24	2.11	5.99	3.42
1–70	0–23	0.75–5	0–25	0–23

Above: *Costs (in US Dollars) paid by patients for various supposedly free hospital services for women giving normal birth. The rows are: Median, mean and range (as in the previous piece of the same table).*

As you can see, for all these items, the mean (middle row) is substantially larger than the median (top row), which indicated that a few people paid relatively high prices. As an example, let's look at the last row, the tests: A median of zero indicates that at least half the patients did not pay for tests. However, some patients did (up to $23), so the mean is above zero. In this case, the median is more representative to most people's situation (most people pay nothing for tests).

Note that **money** is one prime example for a variable where the mean does not mean much: Some people have immense amounts of it, placing the mean value at some arbitrary position that does not have any connection to what most people have.

The other textbook example is **reaction time**: You can't be faster than instantly (or the few hundred milliseconds your nerves need to relay the impulses), but you can easily be much, much slower.

6.3. The „Spread“

For some reason, most of us have an extremely simplified view of describing how far things are apart. For how far two things are apart, it‘s easy: It‘s the distance between them. Duh. Then again, most of you have probably never considered how to talk about how far away from each other three things are.

Range

The range is the easiest measure and the one that is instantly familiar. It‘s the distance between the largest and the smallest observation.

It‘s also a fairly stupid measure, because no matter how many observations you have, the range relies on only two of them. The two most extreme ones. In other words: It‘s unreliable and not representative of the sample at all. So it‘s never actually used for statistical computation, but only to give readers an extremely rough estimate of where all the data points are.

Standard Deviation

The proper measure for how far apart observations are is the standard deviation. It‘s how far the data points are away from the mean. On average.

If that sounds easy, it‘s because it‘s also not quite correct. You see: Because mathematicians base a lot of of statistical computations on the standard deviation, they have customized it to work really well with Math, no matter whether you like the resulting formula or not (and you would not, which is why I don‘t show it, but point to Wikipedia instead).

The deal is this: Either we compute the standard deviation in a slightly weird way, or everything else in Statistics becomes horribly complicated (even compared to how not trivial it is already).

To sweeten the deal, you can have a somewhat simplified idea of what the standard deviation is and you won't be far off:

- You can think of it as the mean distance between each data point and the sample mean.
- You can think of it as the mean distance between any two data points.

These simplified ideas are some 20-30% lower than the actual standard deviation, but they give you a solid idea of what the standard deviation is: How far the observations spread, or in other words, how far away they are from the sample mean and from each other.

Variance

The variance is the squared standard deviation. If you see one, use your pocket calculator and take the square root.

Quartiles and Quantiles

Quartiles and quantiles sound more complicated than they are. In fact, you already know one of them: The median.

The median is where 50% of the observations are lower and 50% are higher. That's the 50-eth quantile or the second quartile (because two quarters of the observations are lower).

To understand quartiles and quantiles, you only need to translate them into a percentage value, and that's how many observations are lower than the value given. Let's look at some examples:

- If the 10-th quantile is at 15 seconds, then 10% of the observations are less than 15 seconds, 90% above.
- If the first quartile is at $12, then 25% of observations are less then $12, and 75% above.
- if the 80-eth quantile is at 36 points, then 80% of people have below 36 points, 20% above.

That said, quartiles and quantiles are not used all that often (mostly for psychometric tests, where you have to know exactly what a score of 21 points means, and not just whether it's above or below the mean).

6.4. The Distribution Type

Why are statistical distributions important? – The basic idea is that reality is Math, so you can replace the jaggy lines of your histogram with a mathematically exact curve.

That means: If you know the form of the distribution, you have an *exact mathematical description* of where your data is, which (obviously) makes statistical computations a lot easier.

The catch is that if you're wrong about the distribution, then you have a totally wrong mathematically exact description. For now, let's assume that we're right, and let's look at the most frequently used distribution: The normal distribution.

Normal distribution

The normal distribution is the most frequently used distribution. It's very common in nature, and many researchers just assume that pretty much everything is normal distributed unless there is ample evidence to the contrary.

Also, the normal distribution has a few nice properties:

- It's symmetric.
- Roughly two thirds of all observations are within the mean plusminus one standard deviation.
- Roughly 95% of all observations are within the mean plusminus two standard deviations.

Some people go so far as saying that the mean plusminus one standard deviation is „normal" – that's certainly too simple, but let's say that anything outside of plusminus one standard deviation does stick out.

Everything else

There are a lot of other distributions, such as the binominal distribution you get when you toss a coin, then the more mathematical Poisson distribution, the t-, F- and Chisquare-distribution we use in statistics a lot, and many more.

However, most of them are fairly close to the normal distribution to begin with, and get even closer the bigger the sample size is. So generally, you can assume that you can treat anything that looks like a normal distribution as if it were one.

6.5. Figures and Tables

Figures and tables are a great way to structure your text and to make it more accessible. Experienced readers will typically look at the tables and figures before they read anything else (except maybe the abstract), so the following things are important:

- Figures and tables ***must either be comprehensive, or they must make a point***. Don‘t just use figures and tables because you have some numbers floating around.
- If you go for comprehensive, try to pack as much as you can into the figure or table, so that a reader feels that he or she has really understood a significant chunk of your work once they have understood the figure or table.
- If you make a point, then make the table or (more often) the figure as straightforward as possible, so that it is quickly clear what you want to say by it. If that does not work, drop it and use words.
- ***Provide good captions and labels***, so that readers can understand without going back into the text. If necessary, write a long(ish) caption; reading five lines of caption is still easier than hunting through the text to find out what the „QRPB“ group is.
- In figures and tables, you can ***use simplistic labels***, even if you make them up on the spot. For example, „old“, „success“, „treatment“ are great labels – provided they‘re explained in the caption. „Group A“ or „People with above average task fulfillment score“ are bad.

The Pie Chart

Pie charts are used very, very rarely in actual research. They work well in slide shows and company brochures, but are considered amateurish in scientific publications.

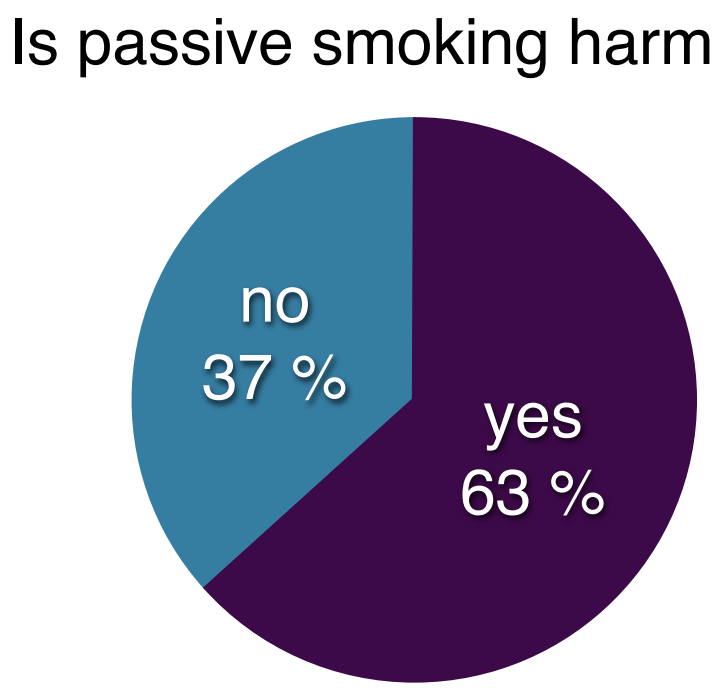
For example, in the directory of open access journals, which includes 850'000 articles, *I've found one article that uses pie charts* (and not for basic descriptive statistics, either).

Why? – Because pie charts have a *very low information-to-ink ratio*.

For example, Deborah Barnes and Lisa Bero have examined review articles about passive smoking. That is: Articles that review a lot of other articles to come to a summary conclusion. As it turns out, they do not all come to the same conclusion – some articles find that, according to the existing literature on the subject, passive smoking is harmful. Other find that it's harmless. Barnes and Bero write in their article:

> Overall, 37% of articles concluded that passive smoking is not harmful. – *Barnes & Bero, 1998*

They could have put the same information into a pie chart (and probably have, if they ever did a presentation or poster on the subjects):



Above: *Pie charts get people's attention and waste space.*

The pie chart contains *less* information than the sentence they used, but takes up much more space. And because publishing is a publisher's game, authors don't get that space. Some publishers even have authors pay for charts, because of

all the extra work and cost that creates (or because they are greedy).

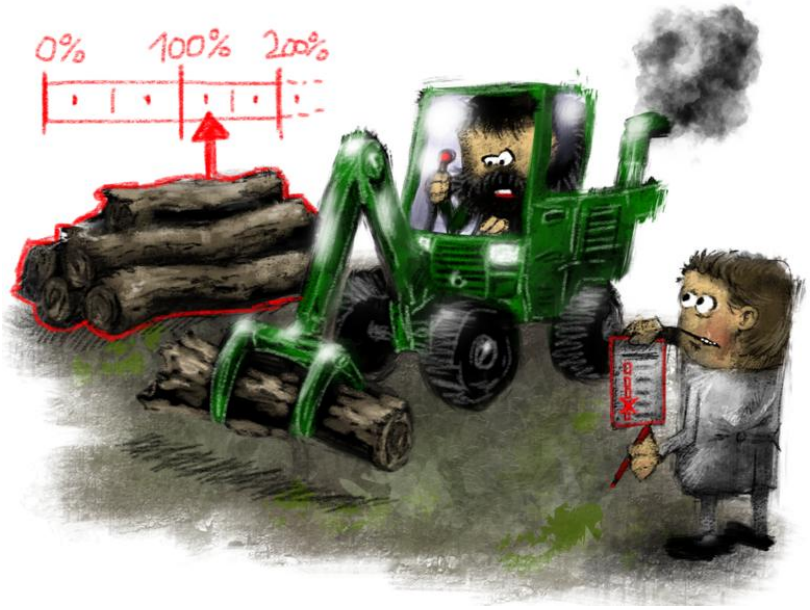
Maybe this changes with electronic publishing, and we'll get more (and more colorful) figures, but I think it's a long way until the pie chart can re-claim any scientific authority.

For now, use pie charts to convince people of scientific facts when you do a poster or power point presentation. When you write an article and want to look like a serious scientist, avoid pie charts.

The Histogram

The histogram shows the distribution of one variable. For each value or range of values, the histogram shows how many observations were made.

As an example, I use a study from Thomas Purfürst and Ola Lindroos about harvester operators. Thomas and Ola had operators perform harvesting tasks while two expert watched and gave marks (from 1 = best to 5 = worst). They also measured the operator's output in the past month and compared it with how everyone else was doing (100% = the operator has an exactly average output).

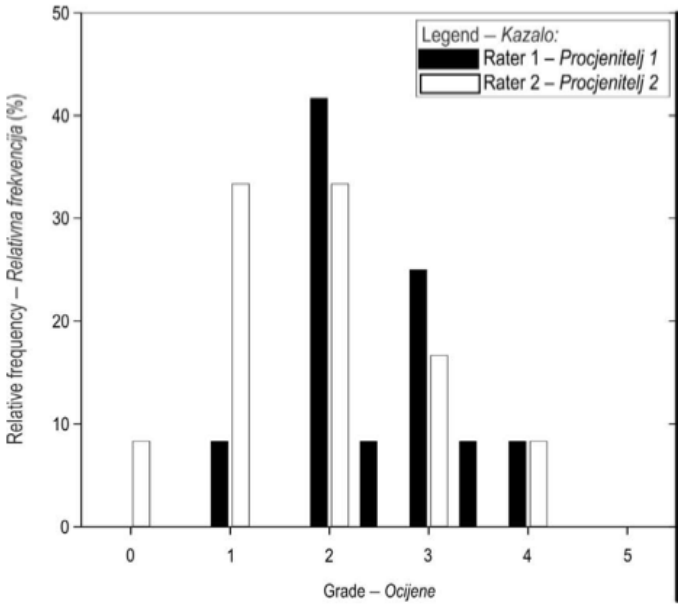


Above: *Experts were asked to rate how well operators performed on a scale from 1 (best) to 5 (worst). Also, the operator's actual performance was measured (100%=average performance).*

Variable has few levels

If a variable has a comparatively small number of levels (say, up to 20), a histogram is the same as a basic frequency analysis, and very easy to create.

Here's a histogram about the grades the operators received from the two raters:



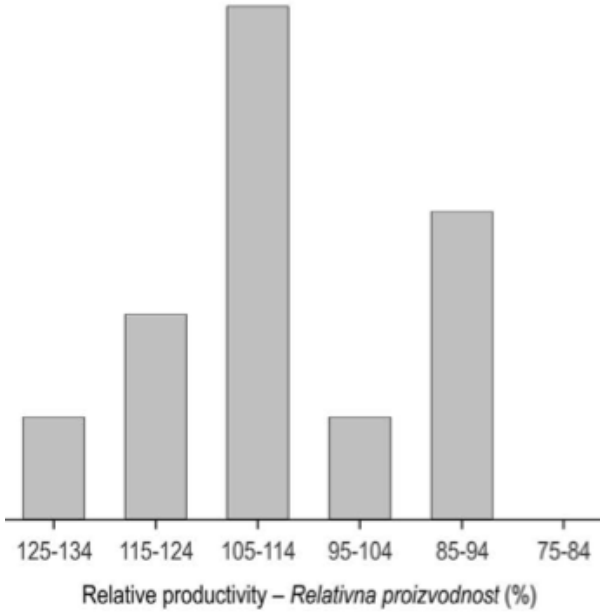
Above: *Two raters were asked to rate the performance of harvester operators. The height of the bar represents how many operators received the respective grade. Lower grades indicate better performance.*

For some reason, Thomas and Ola have chosen to give us the frequency in percentages, so we can see that roughly 8% of the operators received a grade **4** from both raters.

In this type of histogram, there are simply as many bars as there are categories, although here the raters have taken a bit of a liberty with the categories (Rater 1 used values of 2.5 and 3.5, and rater 2 used the value 0, all of which are not intended by the authors).

Variable has many levels

If a variable has many levels or even an infinite number of levels, you cannot just make one bar per level. Instead, the statistics program will cut the scale into even segments. You can make your own segments, such as the 10% intervals used by Thomas and Ola. Whatever you do, note that the histogram will always look a bit different depending on your choice, so don't interpret a single gap or spike as meaningful. It may disappear with other choices for the interval width.



Above: *Relative productivity of the examined harvester operators compared with average productivity of all operators. Most operators are in the 105-114% span, meaning that they are 105 to 114 percent as effective as the average operator. There is a dent at 95-104%, which may however disappear if you select a different interval for the segments.*

The Bar Chart

The bar chart is probably the most widely used type of chart. Still, most bar charts can be expressed just as well (and more precisely) as tables, which is what the more prestigious academic journals prefer. However, if you have to write a poster, advertising materials, or a long academic paper, then bar charts are a great way to get the essentials across fast.

Mohammad Ali Salmani Nodoushan examined different rhethorical moves (according to Désirée Motta Roth's typology of 1995). The bar chart elegantly shows which moves were used the most.

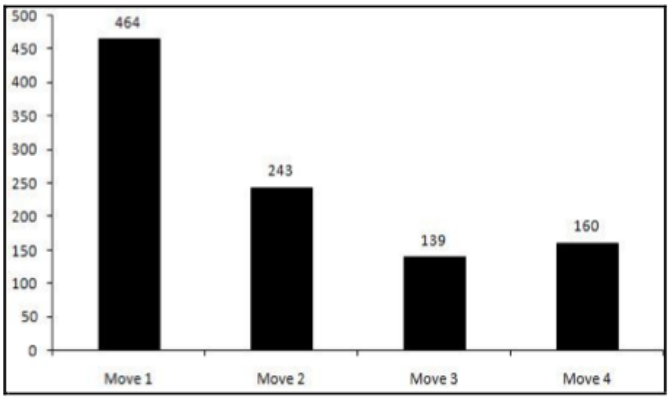


Figure 1. Frequencies of the four moves in the corpus.

Above: *Mohammad Ali Salmani Nodoushan's bar chart shows the number of rhetorical moves in book reviews: 1) Introduction, 2) Outline, 3) Highlights, 4) Evaluation.*

Note: In some journals, the above chart would have been replaced by a table or a few lines of text, because it does not convey a lot of information compared to how much space it uses.

The Line Chart

The line chart can display the same data than a bar chart. However, the line implies a connection (whereas the bars imply separation). There is no strict rule, but you'll typically use a line to show:

- Different measurements over time (such as before-and-after studies, especially if you use multiple time points)
- A profile, that is, different aspects of one overall characteristic
- Passage of time in general

In a study about snake bites in Nigeria, Dr. Oluwagbenga Ogunfowokan checked how long it took until snake bite victims arrived at the hospital for treatment.

Above: *Oluwagbenga took blood samples and noted how long it had taken from the snake bite until the subjects arrived at the hospital. If the blood coagulated (bottom vial), the bite was light. If it did not (top vial), it was severe (as this indicated that poison was active in the blood).*

Because this involves the passage of time, Oluwagbenga used a line chart. To fully understand it, we have first to learn about snake poison (or specifically: viper poison). One main effect of viper poison is that it turns blood to jelly, which is called *coagulation*. You can see a very nasty youtube video about it here (don't click – and if it's not available, search for *„ Snake Venom Clots Human Blood In Scary Russell's Viper Video"*). So if you know how much coagulated blood somebody has, you know how much poison they received. The way to measure this is that you take a blood sample and see if you can get it to coagulate. If you can't, then all the coagulating agents in the blood have been used up by the poison, so you know that the person received a severe dose.

To put it in simple terms:

- If somebody's blood coagulates normally, they are lightly poisoned (if at all).

- If somebody's blood does not coagulate, they are severely poisoned.

Now Oluwagbenga created a line chart from this, where he shows how many lightly and severely poisoned subjects arrive at the hospital, and how long it takes them to get there after they have been bitten:

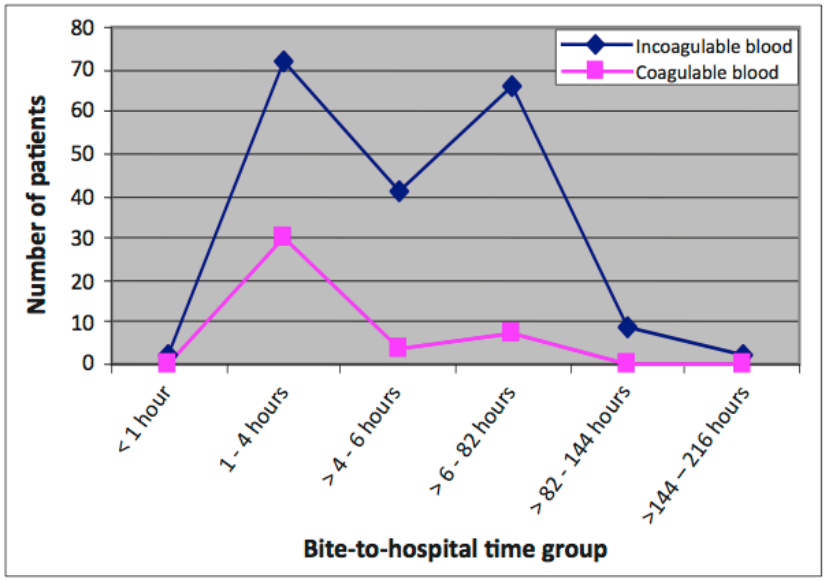


FIGURE 2: Proportion of patients with coagulable/incoagulable blood at presentation vis-a-vis bite-to-hospital time after carpet viper bite.

> ***Above:*** *Severe (blue line) and light (purple line) cases of snake poisoning, and how long it took patients to get to the hospital after they had been bitten.*

As you can see, almost nobody arrived under one hour. 30 light cases arrived within 1 to 4 hours after having been bitten, after which the light cases pretty much stop. The severe cases keep coming in until around 82 hours.

What does this mean? – First of all, that it often takes *days* until a severely poisoned person gets to the hospital (which also means that probably, some victims die before getting there). Also, light cases only go to the hospital when they are near, and probably stop coming after a few hours because by that time, they have figured out themselves that they are essentially okay.

The severe cases, on the other hand, really take their time. Probably, some go to local healers first, and only reluctantly go to a hospital after hours or days of intense pain. Remember that the blood clotting test was done upon arrival

at the hospital, so these people had active venom in them *after* all that time.

One note about the presentation of the chart: This one is straight from Excel, and would need reformatting for some of the more serious journals. For one, those journals are printed in black-and-white, so the colors would have to go and be replaced by line types. Also, gray backgrounds are discouraged – there is no reason to use anything but white for the empty areas in a chart.

The Scatter Plot

The scatter plot offers a good overview of how two scale variables are connected. Let‘s just look at one, and you‘ll see:

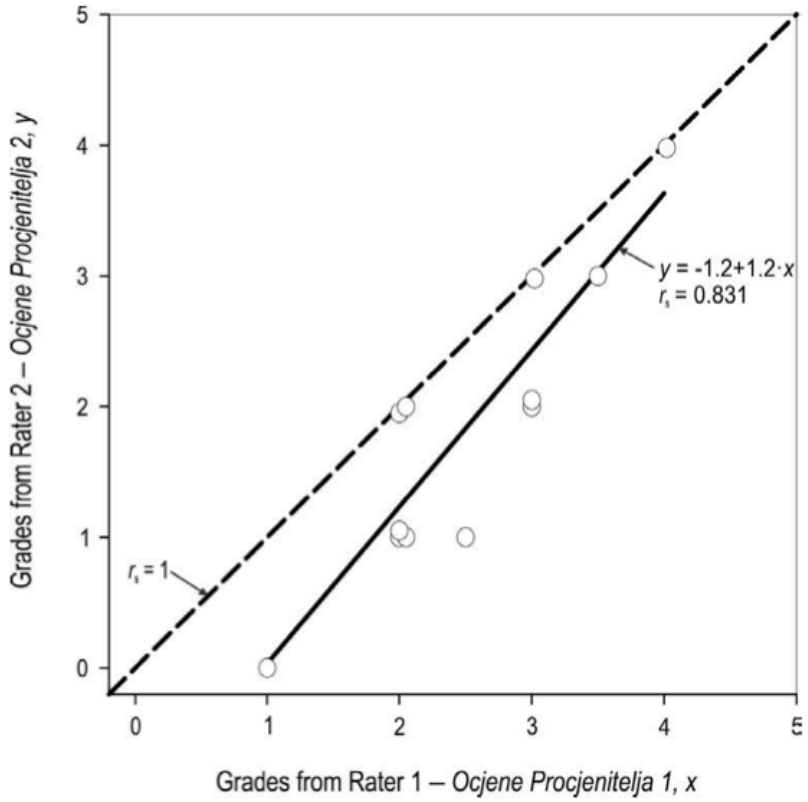


Fig. 3 Relationship between the two raters' grading of harvester operators (N=12; ratings for four operators coincide)

Slika 3. *Odnos ocjena vozača harvestera dvaju procjenitelja (N=12, procjene za četiri operatora koincidiraju)*

> ***Above:*** *A scatter plot showing how two raters judged operator performance. Every dot represents one operator. The horizontal and vertical values represent the ratings by the two raters, respectively.*

The example is taken from study from Thomas Purfürst and Ola Lindroos which I've already discussed here. Thomas and Ola have made two raters judge the performance of a number of harvester operators. In the figure, every dot represents one operator. If you look at the top right of the image, you see one dot where both raters gave a value of 4 (the second-worst value possible). As you can see, the raters pretty much agree on the scores – when rater 1 gives a low score, so does rater 2, and vice versa.

For further information, Thomas and Ola have added the diagonal (the dashed line). If both raters were equally strict, then the dots would lie around this diagonal. They are below, which means that rater 2 overall gave lower scores. The solid line indicates the mathematical line where that is closest to

the points. Note that for most scientific work, it's more important that raters agree in relative scores, and not that they give the exact same values. For exams where you need a certain score to pass, it's of course important that the scores are on the same level and do not depend on the examiner.

So: Scatter plots are great tools to learn where your data is. Then again, they are usually not included in articles. This is because a more space-efficient way exists to tell readers how closely two variables are tied together: The correlation coefficient (discussed here).

The Table

Tables sound easy and boring, but creating a good table often takes a lot of effort, and in turn a good table can compress a lot of research into a small space. Also, the scientific table format is somewhat special, so a professional can look at the tables to both see what your research is about and whether you've got a good grasp of the scientific methods.

A lot of the time, scientists use no figures, and present all the data in table form. That's because tables are precise, comprehensive and you can easily put in a lot of extra information.

So let's look at a few tables from Barnes and Bero's article about passive smoking now. Barnes and Bero have analyzed a number of review articles about passive smoking, and found that some came to the conclusion that it was harmful, others that it was harmless. Then they looked for factors that would explain why scientists disagree in the face of lots of evidence.

Please note that Barnes and Bero are not on a crusade against smokers or tobacco companies here. They have looked at all sorts of variables that might have an effect on the result, such as how well it was executed, which type of effect it was looking at, and the year of publication. It just so turned out that the only variable that had any effect at all was this: Whether the authors had received funding by the tobacco industry. The result is in the following table.

The simple table

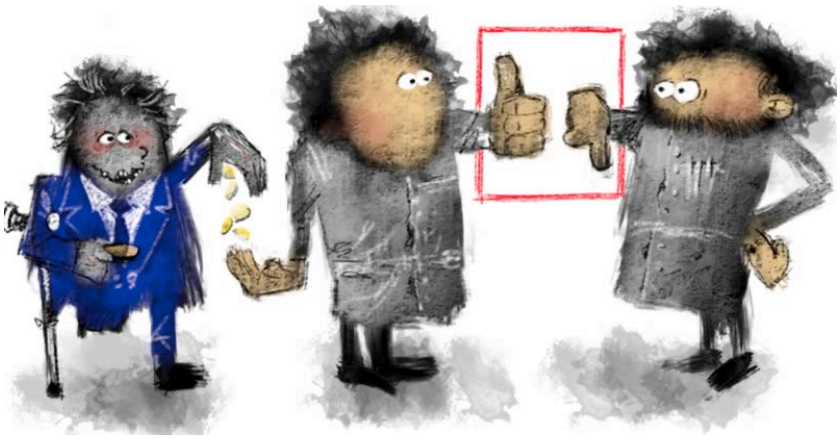
Table 3.—Relationship Between Article Conclusions and Author Affiliations

Article Conclusion	No. (%) of Reviews	
	Tobacco-Affiliated Authors (n = 31)	Non–Tobacco-Affiliated Authors (n = 75)
Passive smoking harmful	2 (6)	65 (87)
Passive smoking not harmful	29 (94)	10 (13)
Significance	$\chi^2_1 = 60.69$; $P<.001$	

Above: *A simple table showing the relationship between the conclusion of review articles about passive smoking and the affiliation of the author.*

To sum up: If authors receive funding from the tobacco companies, they find that passive smoking is harmless. Authors who receive no funding find that it is harmful. This is some very nice and scientific language that the authors use. What they mean is:

Scientists find that passive smoking makes people sick and kills them, unless those scientists are paid by tobacco companies.



Above: *It's not as if you needed an illustration for this, but anyway: From up close (red frame), some researchers say that passive smoking is harmful, others say it's*

harmless. Only when you analyze more factors, you see that there is a method to this: Researchers funded by the tobacco industry say one thing, everyone else another.

Back to the table. It's a simple table, so it's just rows and columns, but note how the formatting is still quite advanced (and not entirely intuitive). We'll go through this point by point:

- First and most importantly, ***there are no vertical lines***. A thick horizontal line is above and below the table, and a thin one separates the heading from the data.
- The **title** is extensive, and if you have any idea about the research, it's enough to understand what the table is about. If there are details (such as abbreviations that need explaining), you can write them in a footnote under the table.
- What you put into the table is up to you. If there's room, put in more. Expect readers to spend some time with the table, and give them as much information as you can. In the above example, the authors added percentage numbers, total numbers of the subjects as well as the result of the statistical test.

The complex table

Table 2.—Descriptive Characteristics of Review Articles on the Health Effects of Passive Smoking

Characteristics	No. (%)* of Articles (N = 106)
Conclusion	
Passive smoking harmful	67 (63)
Passive smoking not harmful	39 (37)
Type of review	
Systematic	11 (10)
Unsystematic	95 (90)
Peer review status	
Peer reviewed	64 (60)
Non–peer reviewed	39 (37)
Missing	3 (3)
Author affiliation	
Tobacco industry	31 (29)
Non–tobacco industry	75 (71)
Topic	
Lung cancer	27 (25)
Heart disease	10 (9)
Respiratory disorders	17 (16)
Multiple health outcomes	44 (42)
Miscellaneous	8 (8)
Years of publication	
1980-1986	16 (15)
1987-1992	47 (44)
1993-1995	43 (41)

*Percentages may not sum to 100 because of rounding.

Above: *A stacked table. The authors have included headings and indentation to segment the rows. Also, they use a footnote (if you use multiple, you can number them). Here, they show how many articles of each type they have examined (for example, they examined 67 articles that came to the conclusion that passive smoking is harmful).*

Stacked tables allow you to present a lot of data in an orderly fashion. Still, the main characteristics are the same:

- There are **no vertical lines**. A thick horizontal line is above and below the table, a thin one below the heading.
- The **title and note** explain everything you need to know.
- What you put into the table is up to you. More is more.

Note that most readers (and all publishers) prefer if you make a few big tables rather than several small ones. That's why Barnes & Bero have used one complex table and not six small ones (one per characteristic).

You can use several tables if you write a long text, such as a dissertation. In that case, it's cool to open each section with a table, so readers can quickly see what the chapter is about (especially if they are already familiar with the table format).

7. Three choices you have to make all the time

So now you have the data and some descriptive statistics, what next? – For most people, that's asking somebody who knows statistics what to do with them. Which is perfectly okay.

Then again, you are only three choices away from doing the analysis yourself, and it's the same three choices for every analysis. So it's not actually very complicated. You can totally do it. Especially if you stick to comparing two variables at a time (which is 90% of what people do, usually).

7.1. Parametric vs. non-parametric tests

There are two kinds of statistics, and they both do essentially the same things and usually come to the same conclusions. They just use different means.

Non-parametric statistics uses the numbers as they are, and get results from crunching them. This is good honest stuff, where few things can go wrong, and which always works. Also, it only gets you so far, so you cannot do any of the really complex analyses. If you're reading this, you probably were not planning to do so anyway, so non-parametric tests are good for you.

Parametric statistics live in a magical kingdom where the data are normal-distributed. So they replace all the data points with a mathematically exact rainbow-colored normal distribution curve and do all computations with that curve. This is a tiny bit more precise than just crunching the numbers, and you can do a lot more with it, especially the complicated stuff. Also, it can be very wrong, if it turns out that the data is not normal distributed in the first place.

Above: *If you do parametric statistics (top), you can soar high, but it only works correctly in that fantasy world where everything is normal distributed (and there are happy rainbow unicorns, too). If you do nonparametric statistics (bottom), you crunch raw numbers and get honest results that are always correct.*

To sum up (again):

- **Feel free to use non-parametric statistics.** *It works. All the time.*
- **If you have to use parametric statistics** (because there are no non-parametric ways to test what you need to know), **do so.**

- If you want to, you can of course check if the data are normal distributed (although you can never be sure), and then go with that. A lot of people do so. Personally I do not see a reason to use parametric statistics when non-parametric statistics does the same job without any of the problems.

7.2. One-sided vs. two-sided hypotheses

First, what is a one-sided and a two-sided hypothesis?

- A **one-sided hypothesis** states that **A is bigger than B**. Or alternatively, that B is bigger than A. Each of the statements is a one-sided hypothesis.
- A **two-sided hypothesis** states that **A and B are different**.

In academic research, almost all hypotheses are one-sided. For example: Women earn less than men, smoking decreases health, yoga helps against neck pain, people remember better when the testing environment is the same as the learning environment.

In practical research, you more often have tow-sided hypotheses: Americans perceive online shopping differently than Europeans, women have different expectations towards customer service than men, and so on.

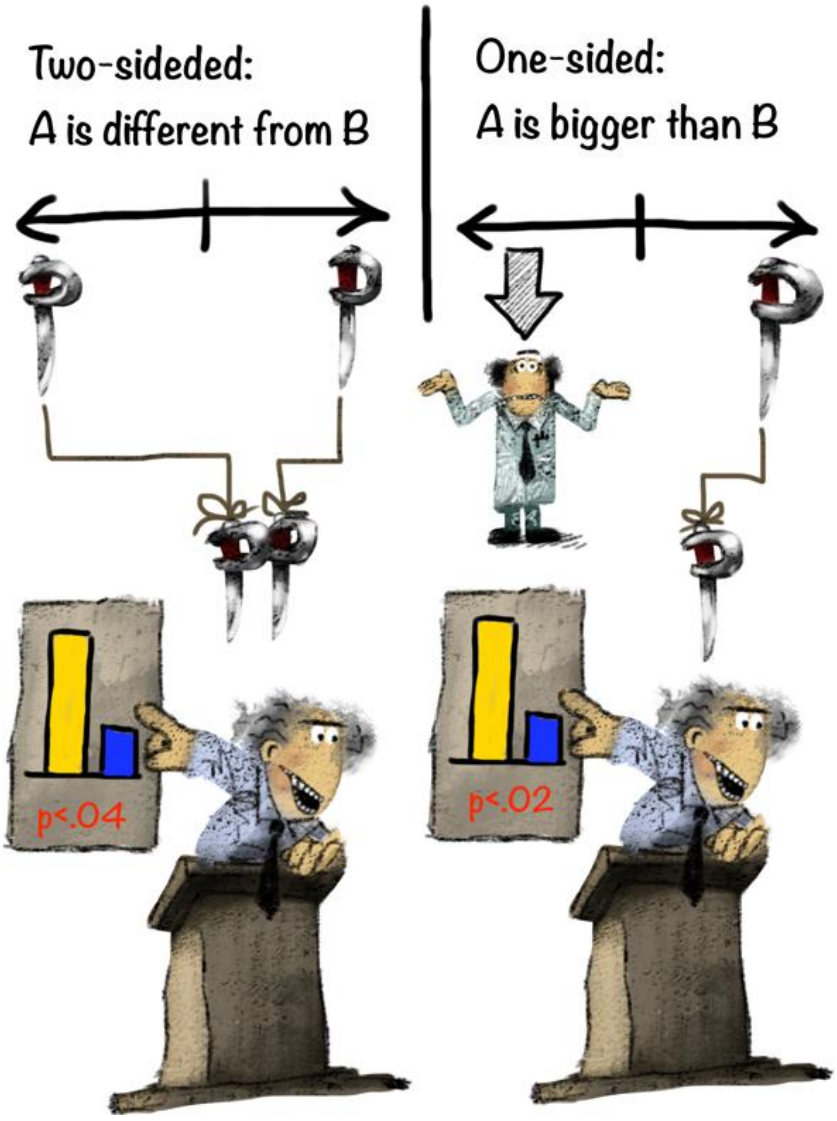
And why is this important? – Because statistics programs often give two-sided p-values (also called *two-tailed* values). If your hypotheses are one-sided, then you can *cut the two-tailed p-value in half*.

That's right: If your statistics program returns „p=.082 (two-tailed)", and your hypothesis is one-sided, then you can cut this in half, so it becomes:

p=.041

That is: You get a significant result (which you would not have otherwise).

Okay, it's not that this is so hugely complicated in itself, it's just that it's really difficult to explain it without making it complicated in the process. So I'll try again with pictures:

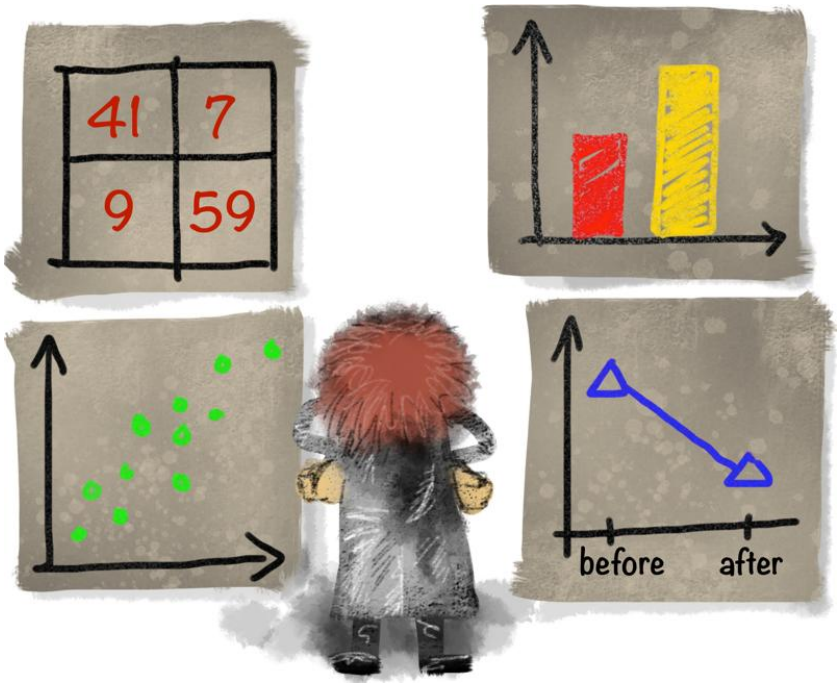


Above: *In a two-sided test (left), there are two swords of Damocles: One for randomly getting a large positive difference, one for getting a large negative difference. Both would lead you to make false statements. In a one-sided test (right), you won't publish a large negative difference anyway, so only the sword for the large positive difference applies. Which translated to half the risk of making a false statement.*

7.3. Picking the right test

This sounds like a tough decision, but if you compare two variables, **there are only four types of test**. It looks like more because of the following:

- Sometimes, there exist two different tests that do the exact same thing. Usually, one is better and one is easier to compute, so in this day and age, you can always use the better one.
- Sometimes, you have one test that compares two groups and a different test that compares any number of groups. Of course, „two“ is „any number“, so a test that compares two groups is technically obsolete. However, most statisticians still use the old test if they have only two groups.
- For most tests, you have a parametric and a non-parametric version, with totally different names.



Above: *When you want to compare two variables, there are only four types of tests you can use.*

The tests themselves are described later – for now, just remember that there are four types of test you need to understand, and then a lot of variations and different names

that are entirely unimportant (and which I describe in great detail and with practical examples, so do not worry).

8. *Analyzing one variable*

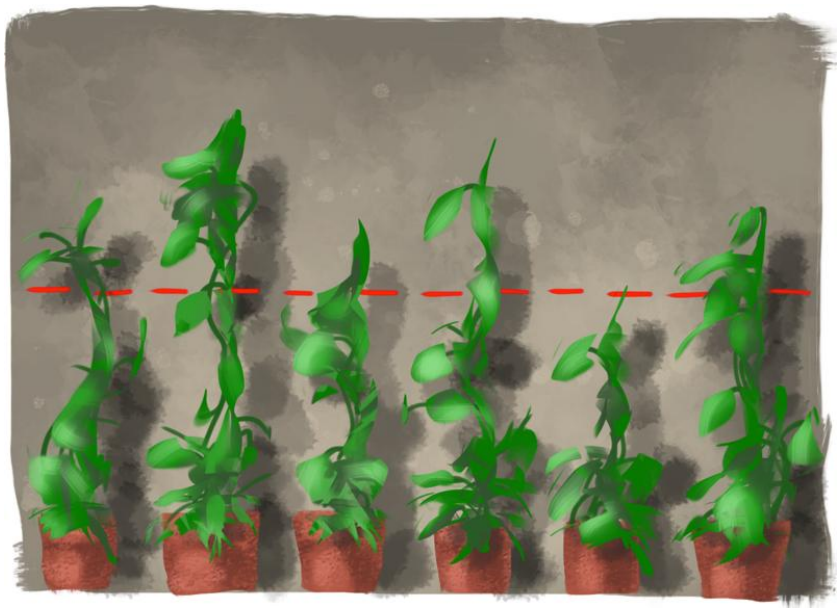
For the sake of completeness, let's talk about analyzing single variables. That's not done very often, but methods for it do exist.

8.1. Testing whether the middle is zero (or any other value)

Stevia Rebaudiana is a plant that has leaves that taste sweet, but contain no sugar. The sweet taste was first described by Moisés Santiago Bertoni, a Swiss botanist, in 1887. Today, a lot of people are looking for a sweet taste that does not involve sugar, so a lot of people want to go out, get themselves some Stevia plants, and start cultivating.

But how do you find out which plant to use for cultivating? And does it even matter? – In the wild, all plants grow differently depending on how much light and water they get, how much competition they have, and of course how good their genes are for what you want to do. You only know whether you have a good plant after you grow it under controlled conditions.

Raji Akintunde Abdullateef and Mohamad Osman went and collected 10 wild Stevia all across Malaysia. From these, they grew 10 offsprings each, under identical conditions. From these 100 total plants, they computed the mean plant size, leaf number, leaf size and a few other things.



Above: *Raji and Mohamad compared samples of Stevia plants against the average height of all their Stevia (red dotted line).*

In the next step, Raji and Mohamad compared the mean of each group of 10 plants with the total mean and checked for differences. They found two groups that were statistically taller than the average, or in their words:

> MS012 and SBK were significantly taller than other collected accessions at p<.001 and p<.002 respectively.

So: Different Stevia you find in the wild do have different properties for cultivation, and they had just located (and grown) two very promising strains (three, if you consider other variables as size).

9. *Analyzing two variables*

There are four basic tests you can do with two variables. Which one you use depends on the type of variables you use, and each has a specific flavor. So it's not just a question of naming the right one for each occasion, but about getting the feel how each of them works.

With these four basic tests, you can do about 90% of all statistical analyses you'll ever be required to do. These four tests are:

1. **Frequency tables:** In a test, do men more often fail than women?
2. **Comparing groups:** In a test, do men have higher scores than women?
3. **Correlations:** Do older subjects have higher test scores?
4. **Repeated measures:** Do subjects have higher scores the second time they take a test?

For each type of test, there are a number of different actual tests, which are confusingly named and arranged. This is because of historical reasons:

- Most tests carry the names of their inventors, or a random letter of the alphabet, or both. For example, take the *Mann-Whitney-Wilcoxon U Test*. Sometimes, they're also named by what they do („Compare two means") in a statistics program.
- The naming is not canonized, so you'll find the same test under multiple names (the *Mann-Whitney-Wilcoxon U Test* is also called *Mann-Whitney U Test*, *Wilcoxon Rank Sum Test* or just *U-Test*). Worse, there are several Chi-square-tests that do very different things (using similar Math, however). And you can write *Chi-square, chi-squared, Chi²* or χ^2.
- When new and better tests have become available, the existing ones were usually kept. So you typically have one test to compare two groups, and one to compare any number of groups. The first one is technically unnecessary („two" is „any number"), but still used.

For this section, I will start with describing the test by what it does, and then go into the details of what it is named and which flavors exist.

9.1. Frequency tables

A frequency table shows how often things happen together, as in the following (simplified) example from Barnes and Bero, already discussed here.

Deborah Barnes and Lisa Bero examined review articles about passive smoking. Now there are plenty of studies about the health effects of second-hand smoking. So many, in fact, that there are over 100 review articles that just sum up all the other articles, so you don't have to read them all. The problem is that these review articles come to different conclusions. Deborah and Lisa have solved this problem. They checked if the authors were known to have received funding by the tobacco industry, and found the following connections:

Table (after Barnes & Bero): Relationship between article conclusion and author affiliation.

Article Conclusion	Tobacco-Affiliated Authors	Non-Tobacco-Affiliated Authors
Passive smoking harmful	2	65
Passive smoking not harmful	29	10

Barnes and Bero show two variables here:

- Affiliation (top row): Whether the authors have received money from the tobacco industry.
- Article conclusion (left column): Whether the article came to the conclusion that passive smoking was harmful or not.

The research question is:

Is there a connection between author affiliation and article conclusion?

We can state the question also as a Null Hypothesis. In this case, the aim of the test is to disprove the following Null Hypothesis:

Null Hypothesis: There is no connection between author affiliation and article conclusion.

Reading frequency tables

How do you read a frequency table? – There are two ways, either by rows or by columns. Depending on the table, one reading is usually easier to understand than the other. Let's start by column:

Article Conclusion	Tobacco-Affiliated Authors	Non-Tobacco-Affiliated Authors
Passive smoking harmful	2	65
Passive smoking not harmful	29	10

First, look at the odds in each column:

- In the first column, it's 2 : 29 (in favor of „not harmful").
- In the second column, it's 65 : 10 (in favor of „harmful").

Then, compare the two odds. Because they are very different, it looks as if affiliated authors and non-affiliated authors come to very different conclusions (we'll see in a minute how to properly test this).

You can also look at the rows and come to the same conclusion:

Article Conclusion	Tobacco-Affiliated Authors	Non-Tobacco-Affiliated Authors
Passive smoking harmful	2	65
Passive smoking not harmful	29	10

Again, let's look at the odds:

- In the first row, it's 2 : 65 (in favor of „non-affiliated").
- In the second row, it's 29 : 10 (in favor of „affiliated").

This time, the conclusion is that most articles who find passive smoking harmful are written by non-affiliated authors, while most article who find it harmless are written by affiliated authors.

That is another way of saying the same thing: There is a connection between author affiliation and article conclusion. Now let's look whether this holds up in the light of a statistical test.

The Chi-square test

The Chi-square test checks if the odds are the same in all columns (or rows, it does not matter). For the above example, the test asks:

Is 2:29 the same ratio as 65:10 *(plus random variation)* ?

The Chi-square test checks this. Like all tests, it delivers two values: A test value (how much bigger than chance the effect is) and a p-value (how likely it is that the effect is pure chance). In an article, the values are cited as follows:

> Authors with industry affiliation more often find that passive smoking is harmless than authors without industry affiliation ($\chi^2 = 60.69$, $p < .001$).

What does it mean? First of all, chi-square test values are squared for some reason. So as the square root of 60 is 7.7, the observed difference is nearly 8 times as big as what you could expect by randomness alone. That's a lot; you don't usually get results this clear, especially if there should not be an effect in the first place. Consequently, the probability that the result was randomness is less than .001, that is less than one tenth of a percent, or less than one in one thousand.

Or, in plain words:

The tobacco industry buys research that shows that smoking is harmless.

To be fair, I don't think that the researchers consciously forged their papers. However, there are a number of psychological effects that can explain how intelligent, honest people come to conclusions that are heavily biased toward their own beliefs and the beliefs of the people around them. Such as: The confirmation bias, the publication bias, conformation to the norm, groupthink, the availability

heuristic, the halo effect, or simply the fact that scientists like money and recognition, and hesitate to go against the people who give it to them.

Numerical example 1: No difference

To give you a feeling of how different ratios work towards statistical significance, here are a few examples with made up numbers. For the following computations, all p-values are one-sided, meaning that your hypothesis or research question stated in which direction you expected the effect to go.

First, let's look at a table where the odds are practically identical, with just some tiny random variation:

Table: Mock data to illustrate how the Chi-square test works with different numbers

Answer	Men	Women
Yes	16	15
No	14	15
Significance	χ^2=0.01; p=.89 (n.s.)	

For this table, the test value is nearly zero: The difference is much, much smaller than what you'd expect by chance alone. Consequently, the probability that the difference is due to chance is near 100% (p is .89, which translates to 89%).

Numerical example 2: No difference either

Now, let's look at a different table with near-identical odds:

Table: Mock data to illustrate how the Chi-square test works with different numbers

Answer	Men	Women
Yes	26	25
No	4	5
Significance	χ^2=0.07; p=.79 (n.s.)	

In this table, the odds are not fifty-fifty as before, but they're still nearly identical. Consequently, the test finds that the difference between the odds is really small, and the

probability that the difference is due to chance is very large (79%).

Numerical example 3: Random difference

The next example shows how a table looks when the difference is quite exactly the size that is expected du to randomness alone:

Table: Mock data to illustrate how the Chi-square test works with different numbers

Answer	Men	Women
Yes	11	18
No	19	12
Significance	χ^2=0.98; p=.32 (n.s.)	

To put it in different words: This is the sort of table you can expect if you throw coins and note where they fall. Or if you just generate random noise and put it into a table. That's what „absolutely nothing" looks like.

Here, the test value is near 1, meaning about as big as what you'd expect by chance alone. Consequently, the probability that it's due to chance is fairly large (32%).

Numerical example 4: Noticeable difference

Now let's look at a table that has a noticeable difference in the odds. Can the following difference be pure coincidence?

Table: Mock data to illustrate how the Chi-square test works with different numbers

Answer	Men	Women
Yes	10	20
No	20	10
Significance	χ^2=2.25; p=0.13 (n.s.)	

According to the test, it can (there's a 13% chance that you get such differences by chance alone). Note that the test value is 2.25, and because chi square test values are squared,

this means that the difference is roughly 1.5 times as big as what you'd expect by chance (1.5 is the square root of 2.25).

So: According to the test, the difference is unusual and slightly bigger than expected, but it's not a huge coincidence to get data like this by chance alone. So you get no statistical significance.

Numerical example 5: Statistically significant difference

To reach statistical significance, the difference must be at least twice as big (roughly speaking) than chance, so the test value must be around 4 or higher, as in the following example:

Table: Mock data to illustrate how the Chi-square test works with different numbers

Answer	Men	Women
Yes	22	8
No	8	22
Significance	χ^2=5.69; p=.017	

Here, I first use my pocket calculator to compute the square root of 5.69: it's 2.39. So the difference is 2.39 times as big as what I'd expect by chance alone. The probability that such a difference happens randomly is 0.017, or 1.7%. That's below the traditional limit of 5%, so we can assume that the difference is not due to chance. Men and women do give different answers to this question.

Numerical example 6: Statistically significant difference, too

Differences can be large without being as symmetrical as in the above examples:

Table: Mock data to illustrate how the Chi-square test works with different numbers

Answer	Men	Women
Yes	20	28
No	10	2
Significance	χ^2=5.14; p=.023	

Here, almost everyone answers „yes“ in both groups, the women somewhat more enthusiastically than the men. This difference is also large enough so that the test says it can‘t be coincidence. Or more precisely: The test says it‘s 2.27 (the square root of 5.14) times as large as what you‘d expect by chance, and that there‘s a 2.3% chance that it‘s chance alone.

Note that some people say that the Chi-square test is not reliable with numbers lower than 5 in any cell. That is a fairly rough rule of thumb, but if you can, use Fisher‘s exact test (see below) in these situations (or in all situations where it‘s available).

Numerical example 7: Larger sample size

You may have noticed that the Chi-square test requires fairly large differences to reach statistical significance. For this reason, here‘s an example with a larger sample size (100 instead of 60 as before). The example just barely reaches statistical significance:

Table: Mock data to illustrate how the Chi-square test works with different numbers

Answer	Men	Women
Yes	36	14
No	14	36
Significance	χ^2=4.76; p=.029	

As you can see, even if you use a large sample, you still need big differences to reach statistical significance.

Fisher's exact test

While the Chi-square test is very useful and widely used, it's technically just an approximation that gives precise results only for very large samples. Especially if the sample is small or contains cells with few observations, it can be off.

For this reason, Ronald Aylmer Fisher has devised an improved test that does the same thing, but calculates much more precisely (and also uses a much more complex formula).

This makes decisions very easy:

- If your statistics package offers you a choice between Pearson's Chi-square and Fisher's exact test, ***use Fisher's exact test.***
- Especially if your sample size is small or if you have cells in your table with less than 5 observations.

Large frequency tables

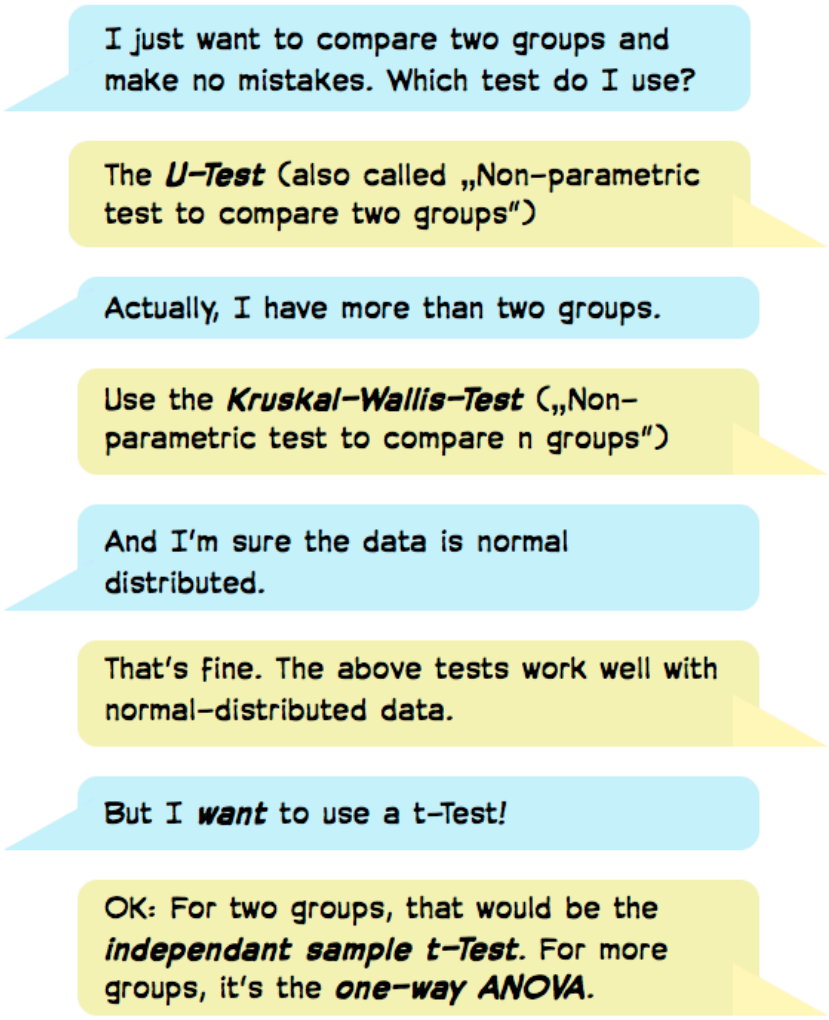
Up until now, we have looked only at 2x2 tables (two rows, two columns). The reason is that these are easy to interpret, so you should ***use 2x2 tables if at all possible***.

Sometimes, you just happen to have data that creates large tables, for example if you examine different levels of education.

9.2. Comparing groups

Comparing groups needs no introduction. It's in our genes, to the point where it's frightening. If there are two groups, then people will have an idea on how these groups are different. Especially people who belong to the better of the two groups. Which always turns out to be everyone.

First, here's how to select the test you need, and then there's one example for each test as it is applied in actual research.



Above: *Follow this chat until you have a satisfying answer which test you need to use.*

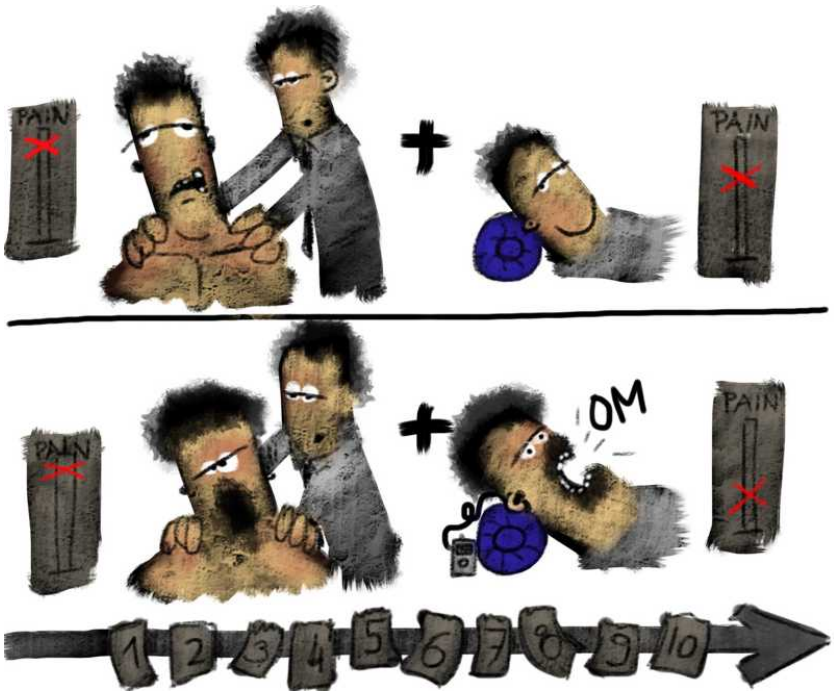
Two groups, any distribution: The U-Test

The U-Test compares whether two groups have similar scores on a variable. It works for any data (that is: data does not have to be normal distributed, but the test works well with normal distributed data, too).

Let's look at an example from Bali Yogitha and John Ebnezar's article about using Yoga relaxation techniques to treat neck pain.

Many people suffer from neck pain, some so much that they get medical treatment. Yogitha Bali and John Ebnezar tested whether Yoga relaxation techniques helped the patients.

Yogitha and John examined 60 subjects who received physiotherapy against neck pain. They split the subjects randomly into a treatment group and a control group. The control group would lie on their back for 20 minutes after physiotherapy. The treatment group would do the same thing and additionally listen to a Yoga relaxation tape. Yogitha and John took several measures of pain and spinal flexibility, both before and after ten days of treatment. For pain measurement, subjects had to place a dot on a 10 cm line that signified their subjective level of pain, from „none at all" to „worst imaginable".



Above: *In a classical treatment study, one group received physiotherapy plus rest, the other physiotherapy plus yoga relaxation. Pain was measured before and after 10 days of treatment. Then, the researchers compared if pain had decreased more in the yoga group.*

As expected, subjects in both groups improved over time (all received physiotherapy, and also most pain gets better over time). What Yogitha and John were most interested in was the following:

Does pain decrease more in the Yoga group than in the control group?

Or, alternatively put, they tried to discard the following Null Hypothesis:

Null Hypothesis: The improvement in the Yoga group is the same as in the control group.

To test the hypothesis, Yogitha and John used the U-Test. Unfortunately, they don't give us many details about the test itself (because readers don't care, besides knowing that it reached statistical significance):

> There were significant (P<0.05) differences between groups on all these variables studied [they were: pain, spinal flexibility, extension, neck movement], with higher percentage changes in yoga than control group. – *Bali & Ebnezar, 2012*

In other words: Subjects who received physiotherapy and Yoga had better pain reduction and flexibility improvement than those who received only physiotherapy.

What is great about the study is that it compares an established therapy with a new approach. Most studies compare a treatment with a placebo, so they essentially find that the treatment is better than *nothing*. Not necessarily better than the existing, old, much cheaper treatments. Or better than the household cures your grandmother would use.

Multiple groups, any distribution: Kruskal-Wallis

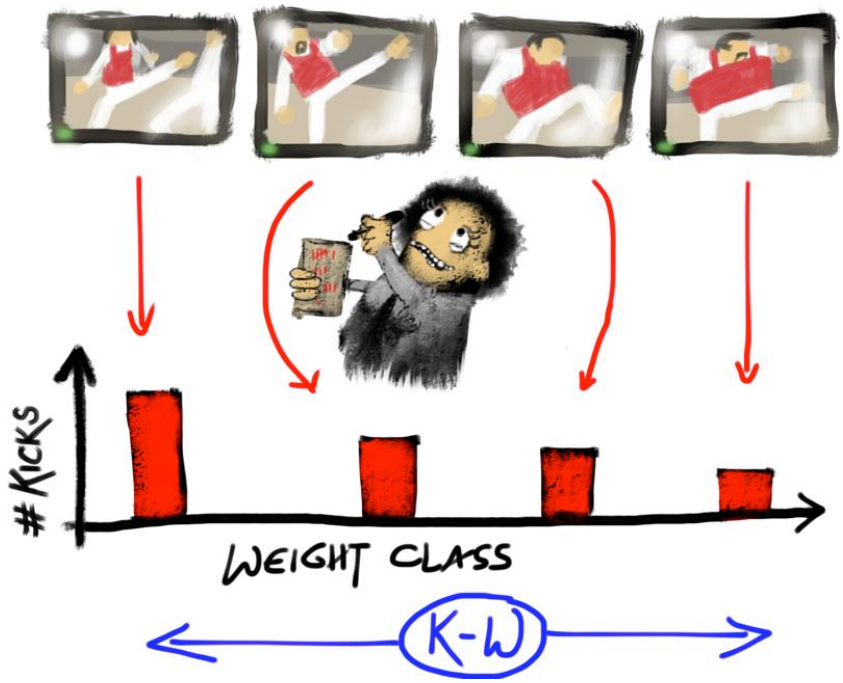
Taekwondo is a South Korean self defense system and the world's most popular martial art. Competitions emphasize ultra-fast spin kicks, so they are also spectacular to watch, especially on championship level.

Coral Falco and her team examined which techniques Taekwondo practitioners in different weight classes used in a match. Because there are four weight classes, she used the Kruskal Wallis test to find differences. She measured a number of different techniques, such as linear and circular kicks used, and made one test for each technique. For this section, we'll only look at linear kicks used by men. For this, the research question is:

Do men in different weight classes use the linear kick more often?

Or, differently stated, she tried to disprove the following Null Hypothesis:

Null Hypothesis: The linear kick is used with the same frequency in all weight classes.



Above: *Falco and friends watched recordings of Taekwondo championships and counted the frequency of certain techniques, such as linear kicks. They compared these frequencies between weight classes using the Kruskal Wallis test.*

The Test value of the Kruskal Wallis test is also a Chi-Square value. This does not have anything to do with the Chi-Square test for frequency tables, except that it uses some of the same Math, though for a very different purpose. In the words of the authors:

> The results of this study in the four weight categories revealed significant differences in males for linear kicks [$\chi 2$ (3) = 8.57, p = 0.04]

Again, chi-square values are squared, so the first step to understanding the result is to take the square root of 8.57, which is 2.93. So the effect is 3 times as big as what you'd expect by chance alone. This translates to a 4% chance (0.04 is 4%) that the effect is due to randomness, so it is statistically significant.

Note that to find out between which groups the difference exists, you have to do Mann-Whitney-U-Tests between any two categories (which they did).

By the way: The reason why such studies are important is obviously not because it's so fascinating to know how many kicks of each type an athlete uses. It's because humans are extremely bad at giving advice based on experience. Especially martial artists. If you train in different martial arts (which I did), then every one of them has a completely different theory on what works in a fight and how you best learn it. Tai Chi has you repeat the same movements very slowly hundreds of times, traditional Karate quickly hundreds of times, Muay Thai emphasizes sparring (that is, free fighting with as much force as you'll use on your buddies), and so on. So: Without empirical analysis, each coach would probably have his or her own private idea of what is best for the athletes, most of them contrary to actual facts. In this example, one could easily assume that lighter athletes would go for complex kicks and heavier athletes for simpler kicks. They don't. Which is good to know when you train for a championship.

Two groups, normal distribution: t-Test

For many students of statistics, the t-Test is the first test they learn, and for many researchers, it's the prototype of a statistical test. That's because it has an elegant formula (not that any of us care), it ties in beautifully with the more complicated tests, and it compares two groups, which is what you'll do most often. It even has a simple name.

That said, the t-test has one drawback: It works only if the data is normal distributed. Along with this one drawback come further complications:

- There is no good way to test whether data is normal distributed.
- If data is not normal distributed, the U-test is several times more powerful than the t-Test in finding significant results. Also, the t-Test occasionally sees an effect that is not really there.
- Even when the data is perfectly normal distributed, the t-Test is only a tiny bit more powerful than the U-Test.

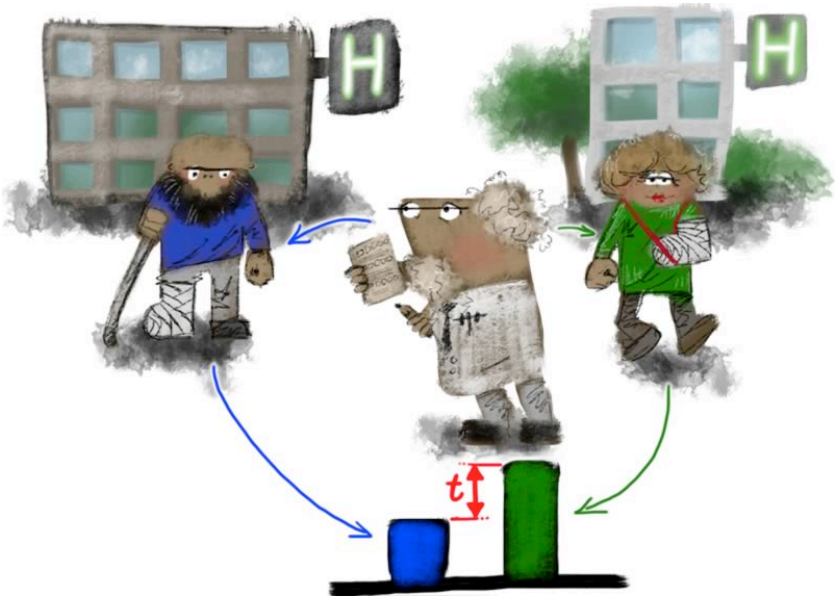
Anyway: The t-Test is still popular, which is probably why Syed Zain-ul-Abideen Shah and colleagues used it in their study of patient satisfaction in Pakistani hospitals. They were interested in the differences between public and private hospitals, so their research question can be put as follows:

Are patients in private hospitals more satisfied than patients in public hospitals?

Or, alternately, one could say that they tried to disprove the following Null Hypothesis:

Null Hypothesis: Patients have the same satisfaction level in private and in public hospitals.

To test it, the team gave a patient satisfaction questionnaire to 100 patients from private hospitals and to 100 patients from public sector hospitals.



Above: *Patients in public hospitals (left) and private hospitals (right) were asked about their satisfaction. The t-test was then used to compare the satisfaction levels.*

They found the following (their words, except that I've inserted a t-value using an online t-Test calculator):

> Mean patient satisfaction score in private sector hospitals was 121.94 ± 20.84 which was significantly higher than that of public sector hospitals, which was 104.97 ± 18.51 ($t = 6.09$; $p < .001$).

First, the t-Value. This is the test value, so we see that the difference is about six times as big as what can be expected by chance. That is: Six times what randomness produces on average. That's pretty big, and test values above 2 are usually significant.

The p-Value tells us, then, that the chance that the difference is just due to randomness is smaller than 0.001, that is, smaller than 0.1%. That's a statistical significance. So: Patients really are more satisfied in private hospitals.

Multiple groups, normal distribution: One-way ANOVA

From Brazil comes a study about surfing, or more precisely, about judging surfing competitions. Rosemeri Peirão and Saray Giovana dos Santos have checked how judges rated different surfers depending on how well they performed certain aspects.

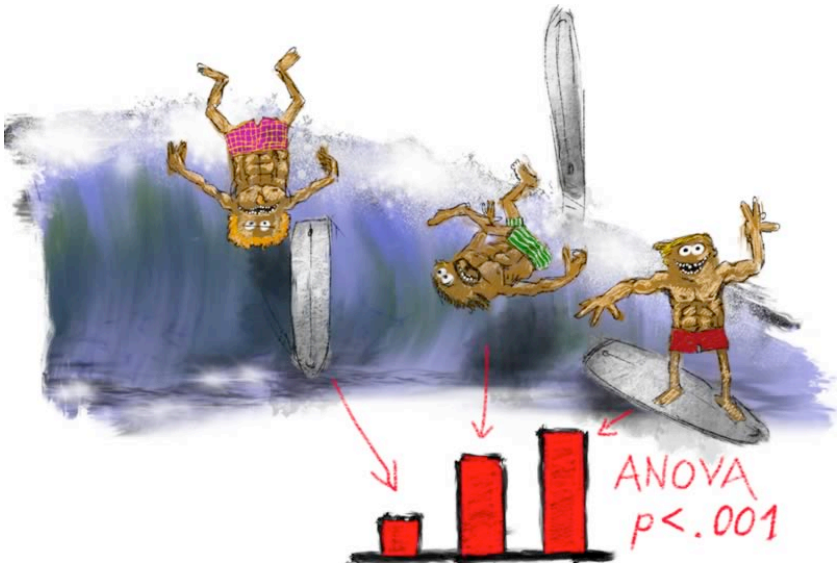
We'll focus on one very small aspect of the study, namely the following research questions:

Do surfers who fall receive lower scores?

Or, posed as a Null Hypothesis, can we disprove the following statement:

Null Hypothesis: Surfers receive the same score from the judges, whether they fall during the main section of the wave, fall after, or do not fall at all.

To test it, Rosemeri and Saray compared the scores for each group (those who fell in the main section, those who fell after the main section, those who did not fall) using a One-way ANOVA. „One-way" is yet another badly chosen statistical term, it just means that there is only one influence variable (how good the takeoff was).



Above: *To see whether judges' scores had any validity at all, Rosemeri Peirão and Saray Giovana dos Santos compared the scores of surfers who fell in the main section of the wave with those who fell after the main section and those who did not fall. An ANOVA found that not surprisingly, not falling in (early) leads to better scores.*

The result is not surprising: On average, surfers who do not fall (early) get better scores. The result could be stated as follows (my wording, using additional test statistics from an online ANOVA calculator):

> Different quality of takeoff led to different scores (F=9.88, p<.001).

Now, F-values are squared (like chi-square values), so the first step to understanding the result is to draw the square root of 9.88. This time, I'll do it in my head, it's three point something. So the difference found is about three times what we can expect by chance alone.

Next, the p-Value. It's below 0.001, so the probability that the result is due to chance is smaller than 0.1%. Which is statistically significant. So surfers get different scores for different entries.

Now, the problem with the One-way ANOVA is that at this point, we don't know whether there is a difference between all groups or only some of them. To do this, we need to do additional tests between two groups. The authors have done this, using one of the usually several options that are

automatically offered when you compute an ANOVA. They have selected Tukey's post-hoc test, and found that indeed all groups are different from each other.

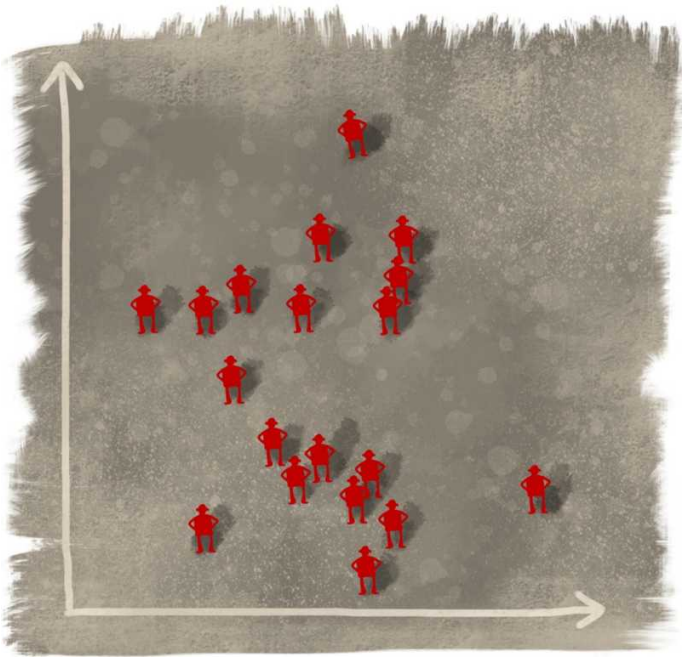
9.3. Correlations

A correlation states how strongly two variables are associated. Typical applications are the correlation between IQ and income, between the number of years people went to school and their scores in a dementia screening, and so on.

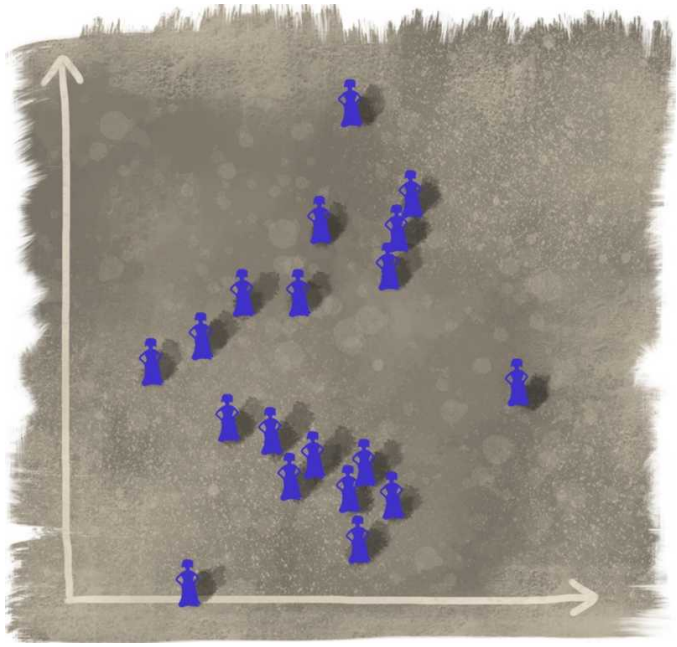
A correlation can have values between -1 and 1, as follows:

- 0: There is no correlation between the two variables.
- 1: There is a perfect correlation: The higher the value of one variable, the higher the value of the other. If you know the value of one variable, you can perfectly predict the value of the other.
- -1: There is a perfect negative correlation: The higher the value of one variable, the lower the value of the other.

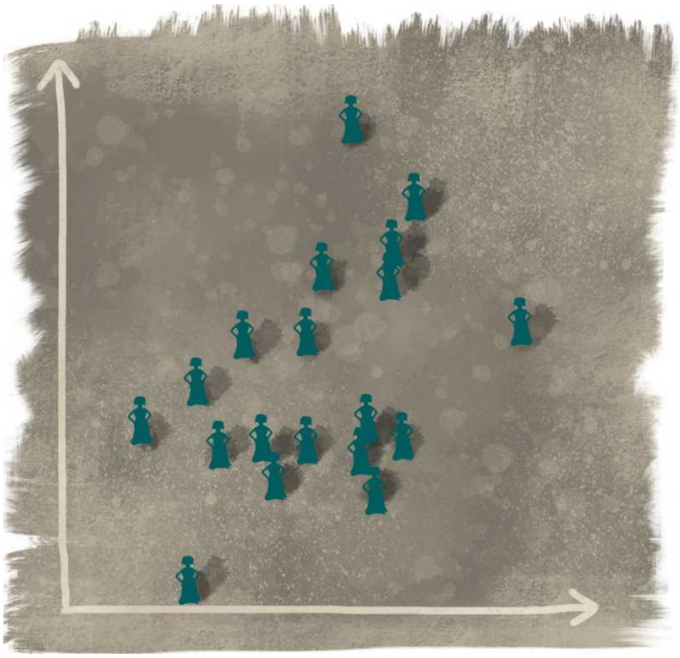
Values in between exist, so for example 0.7 is a moderate correlation. Let's look at a few examples:



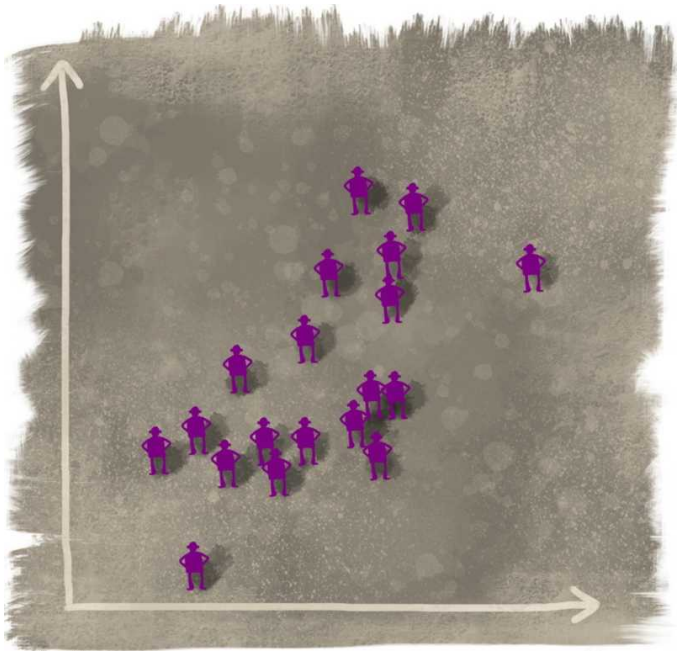
Above: *A correlation of r=0 means no relation between the variables at all. People with high values in one variable can have absolutely any value in the other variable.*



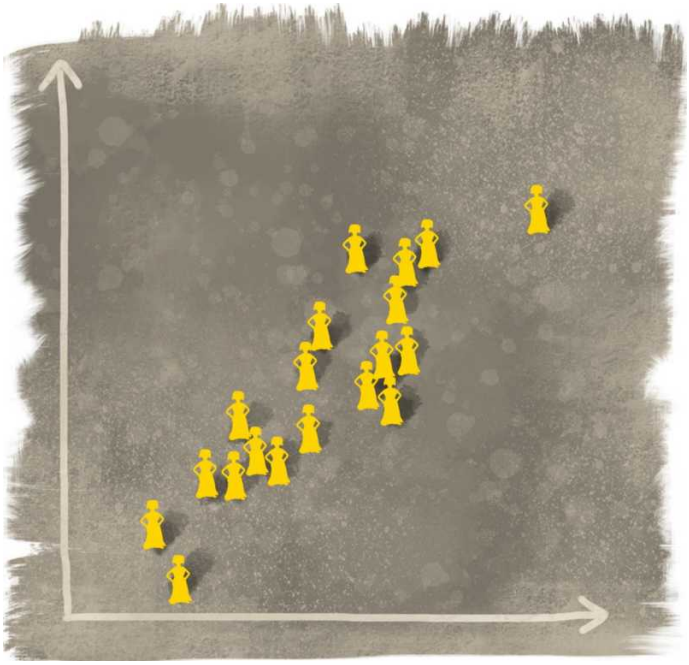
Above: *A correlation of r=0.25 is just about noticeable. The people on the left tend to have some of the lower scores, the ones on the right higher ones. Don't let the single outlier on the right confuse you too much.*



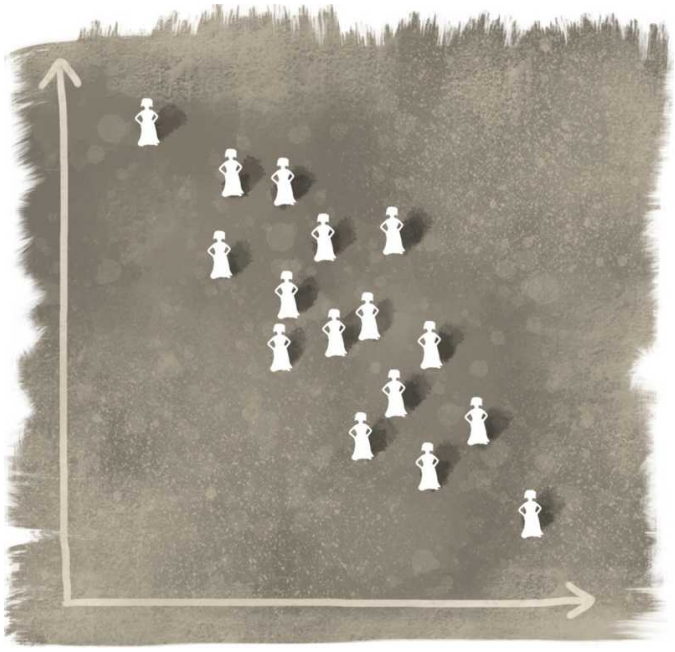
Above: *A correlation of r=0.50 is substantial. It explains a quarter of the variance between two variables. Still, it looks more like an oval than a line.*



Above: *A correlation of r=0.7 is usually considered „high". It explains roughly half the variance of the two variables. At this point, the distribution starts to look more like a line and less like an oval.*



Above: *This is what a very large correlation looks like, r=0.9. It explains 81% of the total variance. Don't expect to find correlations like these unless you measure the same thing twice.*

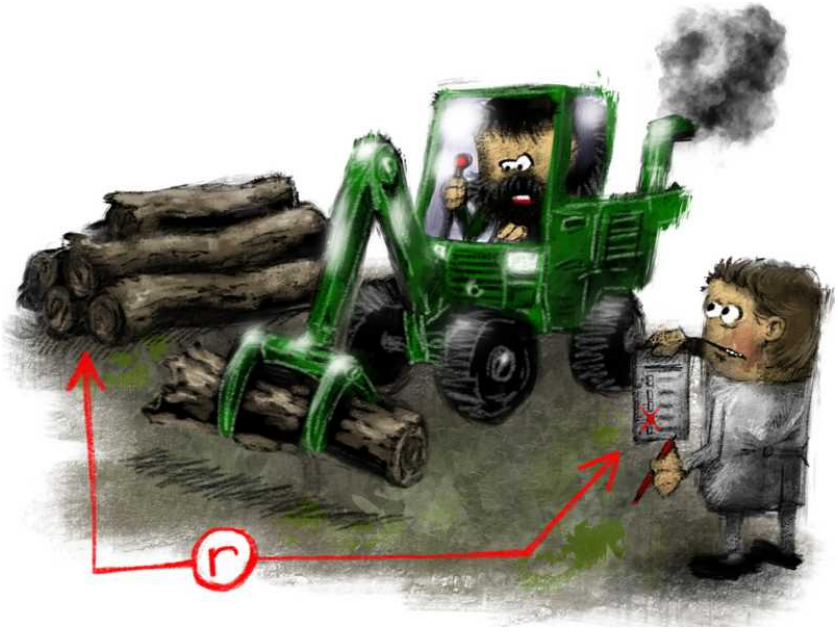


Above: *Negative correlations exist too – they go downward. That is: The higher the values on one variable, the lower they are on the other. That said, a correlation of -0.8 has the exact same strength as a correlation of 0.8.*

Any data: Spearman Correlation

Spearman's correlation coefficient tests whether two variables are correlated, and it works for all variables.

Thomas Purfürst and Ola Lindroos examined performance evaluations in operators of harvesters. They wanted to know if experts could watch an operator for a short time and then accurately say how good that operator was. This is important if you want to train operators who need it (or replace those who are doing a bad job).



Above: *The correlation shows whether high expert ratings go along with high actual performance. The symbol for the correlation coefficient is* ***r****.*

To investigate, Thomas and Ola asked experts to rate the operators after they had watched them for a few hours. Also, they collected data on the operator's performance over the past two months.

Thus, the research question was:

Do operators who receive high ratings by experts actually perform better on the job?

And the corresponding Null Hypothesis that has to be disproved:

Null Hypothesis: The expert rating is independent of the operator's actual performance.

Thomas and Ola correlated the scores from the experts with the actual output and found a strong negative correlation. It's negative because low expert scores indicate better ratings, so the meaning is still that better expert rating goes with better performance. They found (in my words):

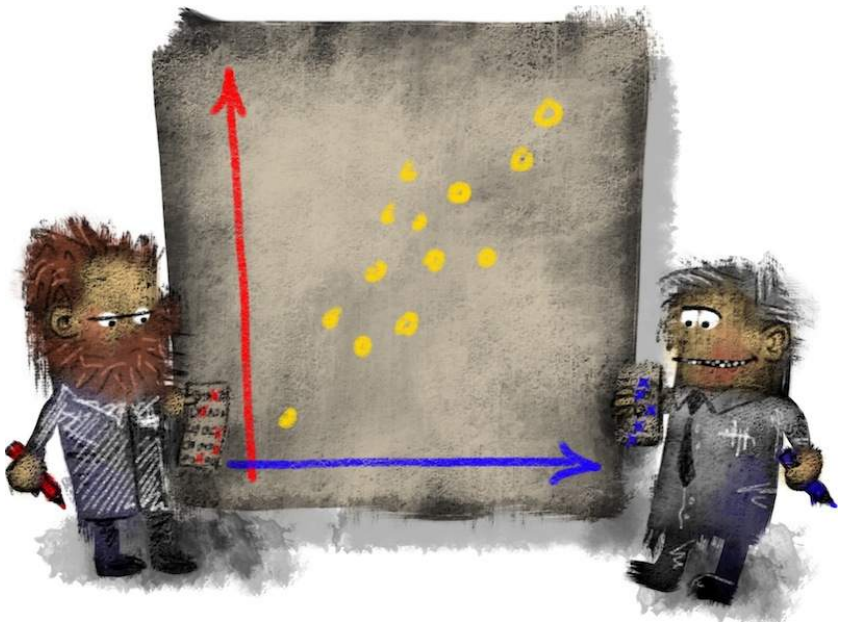
> Expert rating and operator performance were very tightly associated ($r = -0.944, p < 0.001$). Experts very accurately predicted which operators had good or bad performance.

To understand the statistics, first look at the correlation coefficient, r. It's negative because low expert scores are best, so it means that when experts felt that an operator was good, they gave him a low score, which was associated with a high performance. The value is 0.944, which means that experts almost perfectly rated the performance. Basically, knowing the expert rating tells you almost as much as knowing the performance itself. That's really spectacular, so don't expect to see scores like that in a lot of research. Finally, the p value tells us if this can still be a coincidence (that is, a series of lucky guesses). A value of .001 means that there is less than a 0.1% chance that this is random, so we conclude that experts can predict the performance.

In a second step, Thomas and Ola examined the relationship between the scores of their experts. Both of their experts evaluated all of the operators, so we can now check if they agree. And this is what they found (in their words):

> There was a significant positive relationship (r = 0.831, p < 0.001) between the two raters' grading of operators. – *Thomas Purfürst & Ola Lindroos, 2011*

This means that when one rater gave a high rating, so did the other, and vice versa. 0.831 is a high correlation, so there was a high level of agreement and only very few substantial differences.



Above: *One application of correlation coefficients is to check whether two raters agree. In this case, we expect very high correlation coefficients. Note that the correlation does not tell us whether the absolute ratings agree, only that a person who received a relatively high rating by one rater also received a relatively high one by the second one.*

Again, a p value of 0.001 means that there is less than a 0.1% chance that any two raters could get scores this similar just randomly. So this, too, is a statistically significant result, one where we can be confident that it is not just some weird coincidence (and if it is, it's at least a one-in-a-thousand type of weird coincidence).

One thing about correlations: A high correlation does not mean that experts can *exactly predict* the performance. It simply means that they can predict which operators are better and which are worse.

Now, if you think that this is one more of those common sense results, think again. For one thing, performance in most jobs is notoriously difficult to rate. Observers rate people on looks, clothes, whether they're in early in the morning and a lot of factors that have no effect on work performance. As a result, they may fire capable people and promote incapable ones. Both of which costs large amounts of money. So it's a really good idea to occasionally check whether you can judge how well people work or not, and if you find that you can't, then use objective data instead.

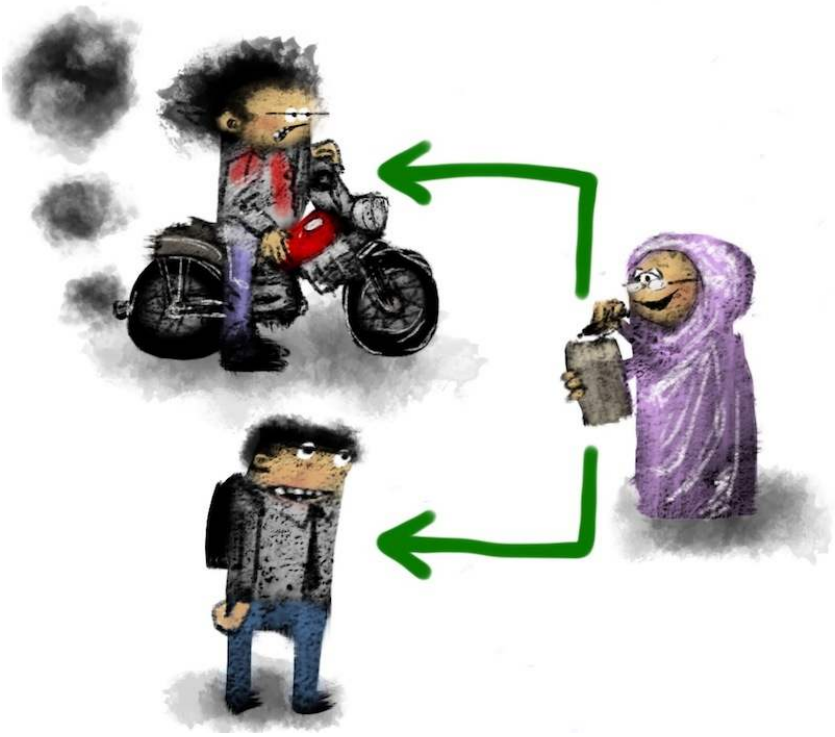
Note that in this research, like in all statistical research, we can simply ignore any effects that are purely random. For example, when measuring the operator's performance over the past two months, there are a number of variables that have an influence on individual performance: Where the operator worked, whether they were ill, whether their harvester worked smoothly all the time, and so on. However, all of these are random and can hit one operator as well as the next, so they cancel out during statistical computations, which look only at effects that are above random influences.

Normal distribution, no outliers: Pearson Correlation

Perason's correlation coefficient does the same thing as Speamans: It tells you whether there is a linear correlation between two variables. However, there is a catch (or two):

- The correlation coefficient is not robust if there are outliers (single points of data that are relatively far away from the bulk of observations).
- Significance testing works only if both variables are normal distributed.

Dr. Wan Shahrazad Wan Sulaiman and colleagues examined personality traits in 68 illegal motorbike racers in Malaysia. It turns out that illegal racers are not the movie type heroes you know from „The Fast and the Furious", but have below average self esteem , leadership and resilience (that is: ability to cope with stress).



Above: *Wan Sulaiman asked both motorbike racers and normal adolescents a series of questions about their behavior. From this, she gained personality profiles of each participant, and could see whether different personality traits are correlated.*

In a further analysis, Wan Sulaiman checked how personality traits were connected, first among 33 normal adolescents she examined for reference. It turned out that in normal young people, self esteem goes hand in hand with a lot of other psychological variables, as follows:

> Results (...) showed that [among normal adolescents,] self-esteem correlated significantly and positively with all the dimensions of resilience: self-assurance, $r = 0.57$, $p < 0.01$; personal vision, $r = 0.56$, $p < 0.01$; flexible, $r = 0.51$, $p < 0.01$; organized, $r = 0.53$, $p < 0.01$; problem solver, $r = 0.43$, $p < 0.01$; interpersonal competence, $r = 0.33$, $p < 0.01$; socially connected, $r = 0.47$, $p < 0.01$; and active, $r = 0.49$, $p < 0.01$. – *Wan Shahrazad Wan Sulaiman et al., 2012*

Now, interpreting correlations is slightly different from other tests, because we are given the correlation coefficient r instead of a test value. For correlations, it is usually best to look at the p-value first: If it is below 5% (that is, below 0.05), then the correlation is statistically significant, which means: We have statistical proof that it is not zero. If the p-value is

not statistically significant, you can discard it, because *the correlation may as well be zero*. In the above example, all correlations are statistically significant, so we can look at the coefficients (the r values). Most are around 0.50, which is a medium-size correlation, so self-esteem is moderately connected to all of the above variables.

In other words: Normal adolescents with high self esteem have generally better personal and social skills than normal adolescents with low self esteem.

As always, we don't know which causes which, and there are three possibilities (that can all be true at the same time, to a degree):

- High self esteem leads to better personal and social skills (for example, because it leads people to be more active and have more learning experiences).
- People with better social and personal skills are more successful, and thus have higher self esteem as a consequence.
- There is an external variable, such as the interaction with parents, teachers and friends, which boosts both self esteem and skills.

The correlation matrix

Because you can correlate any two scale variables, and because most research involves several scale variables, there are usually a large number of correlations you can compute with any data set.

This is where the correlation matrix comes in handy: You can select a number of variables and your statistics program makes a table that lists all the correlations between all of them. This can look as follows:

Now you have to understand a few things about a correlation matrix:

- All the correlations are normal correlations between two variables. They are just conveniently arranged.

- All the correlations in the diagonals are 1. That's what you get when you correlate a variable with itself. Some programs print this, others don't.
- The correlations above and below the diagonal are the same, because you get the same value from correlating A with B as you get from correlating B with A. Again, some programs print both values, some omit those above the diagonal.

Oh, and one more (and very important) thing: When you do a correlation matrix, you do a lot of tests (potentially hundreds of them), each with a small possibility of randomly succeeding even if there is no correlation in reality. So expect large correlation matrixes to have a few randomly significant correlations. Interpret these only if they have a very high level of statistical significance or if they create a pattern than cannot be coincidence.

Table 1. Correlation matrix between self-esteem, leadership and resilience dimensions among normal adolescents

	1	2	3	4	5	6	7	8	9
Self-esteem(1)	-								
Leadership(2)	0.45*	-							
Self assurance(3)	0.57*	0.63*	-						
Personal vision(4)	0.56*	0.49*	0.84*	-					
Flexible(5)	0.51*	0.58*	0.83*	0.81*	-				
Organized(6)	0.53*	0.57*	0.74*	0.68*	0.76*	-			
Problem solver(7)	0.43*	0.61*	0.85*	0.69*	0.80*	0.76*	-		
Interpersonal(8)	0.33*	0.46*	0.76*	0.67*	0.72*	0.61*	0.69*	-	
Socially connected(9)	0.47*	0.54*	0.76*	0.71*	0.76*	0.62*	0.68*	0.72*	-
Active(10)	0.49*	0.54*	0.90*	0.82*	0.82*	0.76*	0.83*	0.78*	0.75*

*p < 0.01

Above: *A correlation matrix.*

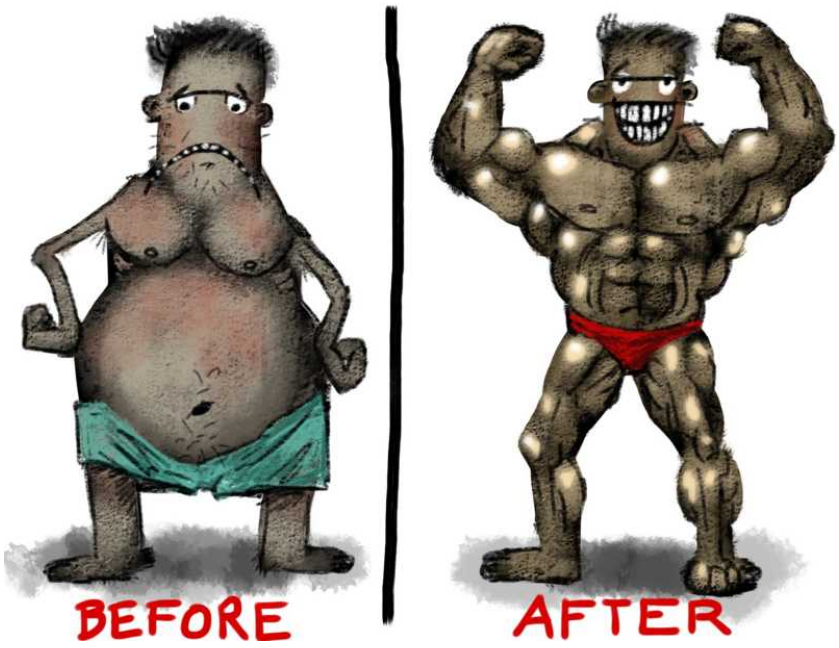
9.4. Repeated measures

Repeated measures are what you know from before-and-after advertising. You measure the same thing twice and check for differences.

Repeated measures are great because they are very precise. Because you compare different scores from the same person, the individual differences do not enter the model, and you get significant results really fast.

Please note that repeated measures are not restricted to before-and-after scenarios. All you need is two measures from the same person on the same scale. As such, you can also use repeated measures for questions like:

- Do people prefer to listen to French or German? (because the preference scores have the same scale, you can compare them with repeated measures)
- Are people better at remembering names of fruits or furnitures? (you could count the percentage of correct recollections, and then compare them with repeated measures)
- Which type of corporate uniform do people judge as the most friendly? (you show pictures of several types of uniforms and let subjects rate the friendliness)
- Which of two products tastes better?



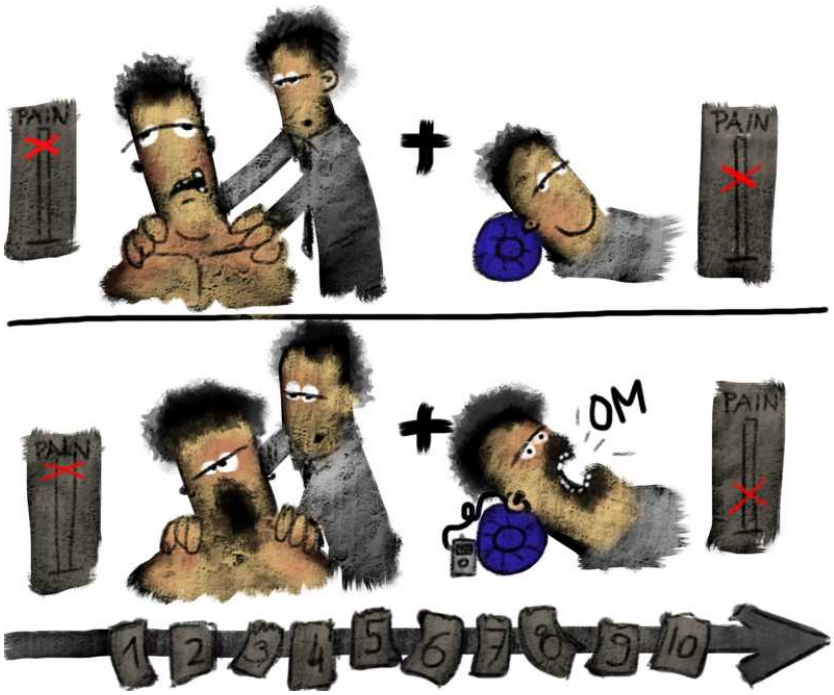
Above: *The most frequent application of a repeated measures analysis is a before-and-after study.*

Any distribution, two measures: Wilcoxon

The Wilcoxon test measures whether people have the same scores in two different measures. Typically, the two measures are the before and after measurement of the same variable.

In their study about the effectiveness of Yoga agains neck pain, Yogitha Bali and John Ebnezar have measured neck pain before and after therapy, and then compared whether the difference was bigger in the Yoga treatment group or in the control group (as we have discussed here).

As part of their analysis, they have also checked whether neck pain was better after the 10 days of treatment. Patients had to give a subjective score of their pain, once before treatment started and once after 10 days of treatment. That results in two measurements, that is, two data points per person. The Wilcoxon test now check if the two data points are the same (plus random variation), or if subjects answered differently the second time.



Above: *Over 10 days of treatment, subjects receives either physiotherapy + rest OR physiotherapy + Yoga. Yogitha Bali and John Ebnezar measured pain and movement before and after.*

The research question for this part of the analysis is:

Does neck pain improve over time?

The corresponding Null Hypothesis, which is to disprove, is:

Null Hypothesis: Neck pain remains constant over time.

(The more interesting question whether pain improves more in the Yoga group, was examined with a different test, discussed here).

Yogitha and John give an extremely short account of the results:

> Non-parametric Wilcoxon's test shows a significant improvement in both the groups in pain ($P < 0.01$).

They omit any test value, which is okay in the case of the Wilcoxon test, because you cannot interpret it anyway. The p value, of course, is the same as for all tests. So $p<0.01$ means that there is less than a 1% chance that the difference is due to randomness alone. So overall, neck pain has improved.

Any distribution, multiple measures: Friedman

The Friedman test is an extension of the Wilcoxon test so that you can compare three and more measures as well.

Safiek Mokhlis, Hayatul Safrah Salleh and Nik Hazimah Nik Mat from Malaysia have examined what young intellectuals look for when they choose a bank. As part of this, they gave subjects a list of criteria and asked them to rate the importance of each of them. As a result of this, they have received a list of scores from each subject. Because they have measured all the scores in the same way (that is, as numbers from 1 to 5), they can compare them using a repeated measures test, such as Friedman's. The corresponding research question is:

Do subjects assign different importance to different factors?

The Null Hypothesis to disprove would be:

Null Hypothesis: Subjects give equal importance ratings for all factors.

Above: *The team asked young intellectuals to rate different criteria for choosing a bank. Thus, they received a number of measures that all use the same scale*

(1-5 points). Using the Friedman test, they checked if subjects rated the criteria differently.

Their result was (their words, slightly modified):

> The Friedman Test indicated highly significant differences in the importance of bank selection criteria ($\chi 2 = 1643.263$, $p < 0.001$).

Like many tests, Friedman's also results in a $\chi 2$ test value, so we first take the square root of 1643, which is 40 (roughly computed). So the differences are 40 times as big as can be expected by chance. As in: Really big.

The p-value of 0.001 means that there is less then a 0.1% chance that the result are due to randomness alone. This is statistically significant, so we can assume that subjects overall prefer some criteria over others.

Normal distribution, two measures: paired samples t-Test

The paired t-Test compares two measures and checks if their means are equal. For example, you could check whether pain level is the same before and after taking a painkiller.

As with all t-Tests, the paired samples t-Test is very slightly more powerful than the corresponding non-parametric test (here: the Wilcoxon test) if all requirements are met, and is much worse otherwise. Also, the test exists in two flavors, one that assumes that both measures have equal variance and one that works with any variance. Because there is no good way to prove that variances are equal, I recommend that you always use the test that works also if variance is not equal (if your statistics program offers you this choice).

Example

Zahra Kordjazi (Teheran) examined how well her subjects were at English grammar.

Now, Zahra's study does not follow all the rules concerning how to write about research, but it's a great example of a

repeated measures experiment, and an easy read. Also because she copy-pasted the output of the SPSS computations into the paper (which you should not do, but which makes it great for you to see what she computed).

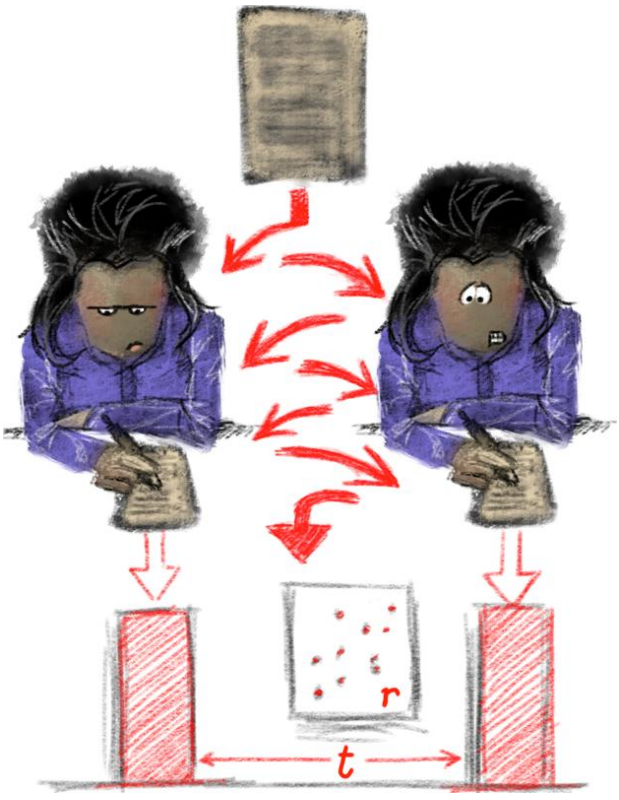
Zahra has studied whether second-language skills are brought down when subjects have to deal with rare words in a sentence. Subjects had to read a sentence and then decide which of the underlined expressions was wrong, for example:

9. The most of what Alice told me was not true and believable.

Now, half the subjects received the above sentence, and half received one in which a word was exchanged for a very rare one:

9. The most of what Alice told me was not true and veracious.

Zahra was interested to see if her subjects performed worse in sentences with difficult words.



***Above:** Zahra gave participants a list of sentences where they had to find the mistake. Half the sentences were regular English sentences (left), half had an extremely rare word in them (right). Thus, Zahra received two scores from subjects: Performance on the sentences with and without rare words, respectively.*

Because these are two measures from the same subject, using the same scale, she can compare them with the paired sample t-Test.

Now let's look at what it was she was doing:

- It's an **experiment** because she randomly assigns subjects to experimental conditions. Half the subjects get the sentence with the easy word, half with the difficult one.
- It's **repeated measures** because she measures the same variable (performance in a grammar test) under two conditions (presence vs. absence of difficult words).
- For the measurement, she uses a **psychometric test**, which measures **behavior** (how well people perform).

For the analysis, Zahra used a paired-samples t-test (that is, the flavor of the t-test you use for repeated measures). For the difference between the group, she got the following result (the wording is mine):

> There was no difference in performance on the two lists (t = 0.962; p = .35).

In other words: The difference she found between the group was almost exactly the size you'd expect by chance alone (t=1 would indicate a result of the same size as chance). There's a 35% chance that it's pure coincidence, which is a long way off the 5% required for statistical significance.

In all repeated measures studies, you can (and should) compute correlations between the measurements, which she did (again, with my wording):

> Performance on both lists was significantly correlated (r=0.68, p<.001).

The positive correlation coefficient means that the better subjects were on one list, the better they were on the other. This correlation reached a strong statistical significance. Note that the output indicates *.000*, which just means that the value is smaller than anything the program can display, or (to put it in correct mathematical terms) <.001.

Where I am sitting now, it's getting late, and I have this one very complicated paper that contains a repeated measures ANOVA. I think it's easier for all of us if you imagine a repeated measures ANOVA simply as a paired samples t-Test with more than two measures.

Okay? – More in the next edition. If there is one.

10. Writing

So after you have collected and analyzed the data, there is one more thing to do: Writing about them. Some people find this harder than others, and nobody finds this entirely easy. I'm a professional writer and I'm still way behind in handing in papers for my university studies. That said, I'm pretty efficient at writing books and stuff. It's crazy how many new talents you discover when you should be doing something else.

When you approach writing, the first question you need to ask yourself is:

What type of paper am I writing?

In my experience, there are only two types, and they pretend that they are the same, but they're not. Understanding the differences will save you a lot of frustration.

- The ***Peer Paper*** is where you write about your research to your scientific peers. They want to learn what you have found out.
- The ***Career Paper*** is a paper like a thesis or dissertation, where the readers (usually, there's only one or two of them) want to know whether you're a good researcher. You write this one for your career. If anyone except your professor reads it, it's a bonus.

In short:

The Peer Paper is about the research, the Career Paper is about you.

That said, both have the same overall structure. Of the two, the career paper is probably harder to write, if only because it's not entirely clear what it is. Usually, it *pretends* to be a peer paper too, so we'll look at those first.

10.1. Writing a great Peer Paper

Research papers always have the same structure, and it usually follows the so-called APA style. That's the style the American Psychological Association uses for all of its papers. Psychologists were the first social science to place a heavy focus on research, so by the time other sciences followed, the APA had several research journals running, plus an excellent guide on how you write articles for them. Other publishers and associations either have very similar style requirements, or they just follow APA style.

We'll go into this in a minute. Before, let me point out the master rule:

Readability rules.

So whatever you do, always make your paper as easy to read as possible. For the expert reader, that is. So stick to conventions that the experts expect, but within them, make it easy to understand what you're talking about. Use short sentences. Use simple words. Repeat words and sentence structures. Explain everything (briefly). *Do everything to make yourself understood.*

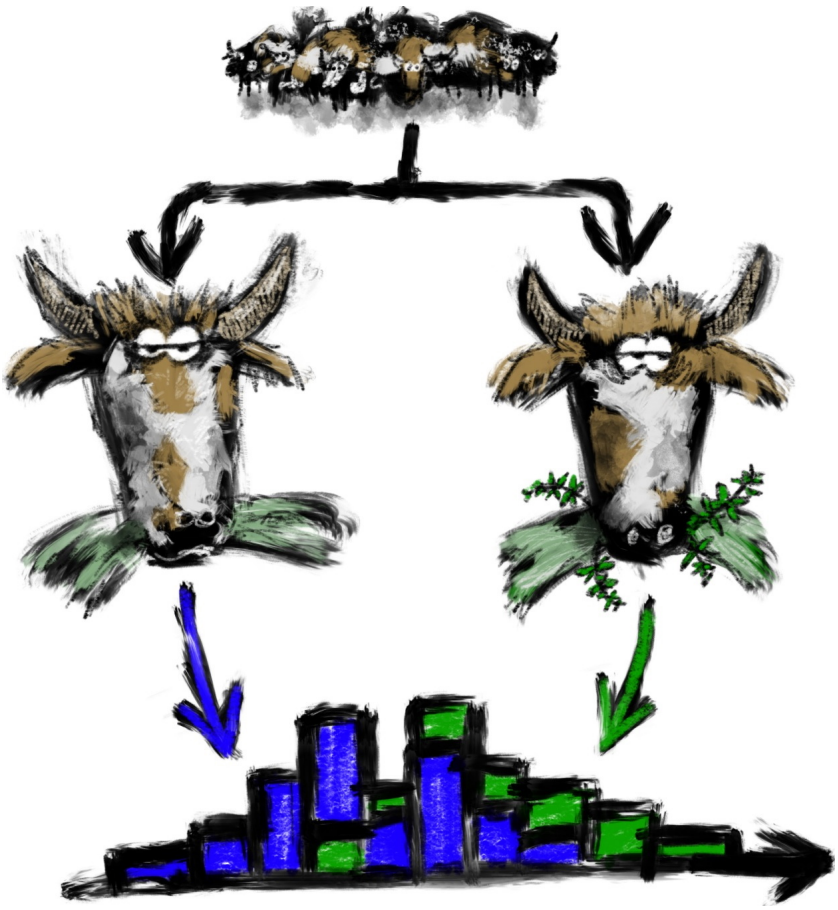
We'll go through each section of a peer paper in turn, using the example from Mehdi Goodarzi and friends: *Evaluation of Effect of Peppermint as a New Preventive Method for Milk Fever and Subclinical Hypocalcaemia in Transition Holstein Cows.*

You are probably aware that calcium is vital for your body, and that children need it for bone growth. Milk ads say this all the time, because other than one of many possible sources of calcium, milk is mostly fat floating in water, with some other stuff mixed in that causes cramps and vomiting in 75% of the world's population. In agricultural societies, people have adapted to drinking cow milk – probably by the simple evolutionary process of everyone dying who could not get used to it. Everywhere else, kids seem to grow up just fine without drinking any of the stuff beyond what they get from their mothers.

Back to calcium: It's important for health and bone growth. As it turns out, cows need it too, especially if they've just given birth, and all the more if they are bred to give off record amounts of calcium-rich milk all year round. So they

can develop „milk fever“, a condition where they become disoriented and can die from lack of calcium.

Iranian scientists Mehdi Goodarzi, Mohammad Sadegh Safaee Firouzabadi, Mohammad Mahdi Zarezadeh Mehrizi and Mohsen Jafarian have researched whether feeding peppermint to the cows can improve calcium levels in the cow‘s blood. And here‘s how they did it.



Above: *Mehdi Goodarzi and friends randomly assigned Holstein cows to one of two groups. The first group received normal food, the second one the normal food plus peppermint. Then they measured calcium levels and compared the groups.*

The Abstract

The abstract contains *as much of the research as possible* in the number of words permitted (usually, about 100). Specifically, tell readers about the results and implications. Most readers will only ever read the abstract, so *make it count*.

Goodarzi's abstract reads as follows:

> ***Abstract:*** Milk fever and subclinical hypocalcaemia are the most important macromineral disorders that affect transition dairy cows. Feeding peppermint to dairy cattle due to physiological or pharmacological functions may effect rumen fermentation and digestibility. Thus, in the present study, peppermint was tested for its effects upon serum calcium and urine pH levels in transition Holstein cows. Twenty dairy cattle (the last 3 weeks prepartum) were divided into 2 groups. Group 1 (control) treated with no peppermint and group 2 (treatment) treated with peppermint. To determine serum calcium and urine pH levels, blood and urine were taken in start of experiment, after 15 days adaptation period with the peppermint being mixed with concentrate and 12 h after calving in second milking. Serum calcium and urine pH levels between control group and treatment group were not significantly different ($p>0.05$) in start of experiment, but their levels after 15 days adaptation period and 12 h after calving were significant ($p<0.05$). In treatment group serum calcium and urine pH difference between before treatment and after treatment in three different times were different significantly ($p<0.05$). According to these results, it can be concluded that peppermint has a great potential as a new preventive method for milk fever and subclinical hypocalcaemia by inducing mild metabolic acidosis or any unknown mechanism. – *Goodarzi et al. (2012)*

This contains almost all you need to know, specifically the method and statistical results plus the authors' conclusions.

Note that this is not a crime novel, so no points for adding suspense. Tell everything up front. Most people will read only this far. Any surprise results you have hidden in the main text will probably stay there. Also, most people who will read your entire paper do so because it says here that it's worth it. If the abstract does not state good and clear results, then people will assume that there aren't any.

The Introduction

The introduction is a short overview of the problem you examine and the existing literature about it. It can be very short, if you do research on a simple, well-documented issue.

The aim of the introduction is to **provide background information to people who are not familiar with the research you do**. As such, it does not contain any of your thought or research, it's just a number of factual statements about other people's research and about the problem you examine.

Also, the introduction contains the **research question**: What you are trying to find out. There is usually no explanation how and why you arrived at this question (beyond telling readers why it's important, which you have already done).

Look at the following introduction:

> ***Introduction:*** Milk fever and subclinical hypocalcaemia are the most important macromineral disorders that affect transition dairy cows. They influence on the transition cow to effect skeletal and smooth muscle contraction and exacerbates the level of immunosuppression experienced by periparturient dairy cattle (Kimura et al., 2006). Normal blood calcium is 8 mg/dL (2.0 mmol/L) in cattle. It has been reported that milk fever cows are up to eight times more likely to develop mastitis in the following lactation, are three times more likely to develop dystocia and two to four times more likely to develop abomasal displacement (Mulligan et al., 2006). On average, 5-10% of dairy cows succumb due to clinical milk fever, suggesting that the incidence rate in individual herds reaches as high as 34% (Houe et al., 2001).
>
> Herbs have been shown to have pharmacodynamical and pharmacokinetical functions. It has been observed that peppermint oil has antifoaming effect on in vitro gastric and intestinal foams (Grigoleit and Grigoleit, 2005a, b). Peppermint has a strong ability to act as a natural manipulator of rumen fermentation (Andoa et al., 2003). In this study we sought to determine if peppermint can alleviate hypocalcaemia and milk fever during the close-up dry period (the last 3 weeks prepartum). – *Goodarzi et al. (2012)*

With this text (and the occasional trip to Wikipedia) you know everything you need to know to understand the research. At the same time, it's just a collection of existing knowledge, up to here, Goodarzi et al. have not made any contribution.

As you may know, most research articles are examined by one or two experts before publication. The authors are not told who they are, but they can often guess. Because of this, they make sure to include those author's publications in the introduction, *if at all possible*. That's one half of rational behavior (after all, research is about collaboration, which includes building on top of other people's work) and one half of bribery.

As a rule of thumb, your research should reference at least one dozen other research articles, and the introduction is usually the best place to do this. If you write a very short article (as the authors have here), you can also reference the articles in the discussion section, where you compare your results with theirs.

The Methods

The method section tells the story of your research.

When readers have read it, they'll know where and when the research took place, how many people (or cows) were involved, and what they did one thing after another until all the measurements were in.

This can be lengthy, but that's fine, as long as it's about the subjects and what they do (or rather, what the researchers do to them):

> ***Materials and Methods:*** The research was conducted in summer 2011 at one of commercial dairy cattle farms of Yazd, Iran. In this study 20 Holstein dairy cattle, 4 years old, were tested. Transition dairy cows were defined as those in the last 3 weeks before calving. Their average body weight was 450±25 kg. The cows were divided into 2 groups. Group 1 (control) treated with no peppermint and group 2 (treatment) treated with peppermint.

They were fed with 1 kg of Alfalfa hay, 3 kg of mashed concentrate mix (10% barley groats, 36% canola meal, 21% wheat bran, 30% rice bran, 2/5% vitamin supplements and 0/5% NaCl) and 12 kg of corn silage at 6:30 followed by 2 kg of wheat straw at 19:30. In treatment group, all dry cattles were fed with 400 g of peppermint (sun-dried imported from central areas of Iran) daily that was mixed with mashed concentrate. Water provided ad libitum.

Macromineral analysis of sodium, potassium, chloride, sulfur and calcium of peppermint were measured using Wet chemistry methods. To calculate Dietary Cation Anion Balance (DCAB) in milli-equivalents per 100 g of ration of dry matter for a peppermint, the following formula was used: [(% sodium/0.023) + (% potassium/0.039)] − [(% chloride/ 0.0355) + (% sulfur/0.016)] (Moore et al., 1997).

Ten mL of blood was taken from jugular vein of each cattle in three different times: in start of experiment (Ca1), after 15 days adaptation period with the peppermint being mixed with concentrate (Ca2) and 12 h after calving (Ca3) in second milking. Serum was separated and the level of calcium was determined spectrophotometrically at 550 nm (Jadhav et al., 2010).

The urine pH of all cows determined at three different times similar to blood sampling. Urine pH was measured by digital pH-meter from fresh urine in middle part of urination. – *Goodarzi et al. (2012)*

Often, the method section includes a note on the statistical measures used for analysis. This is just a matter-of-fact statement, at most with a short statement why one method was selected over another one. There is never a discussion of why one method is better suited, If at all, there is a reference to an article or book where this is discussed in some more detail. More often than not, that referenced text does not actually contain in-depth mathematical analysis.

The Results

This is the beef of the article. It's what you found out, with figures, tables, means, standard deviations and p-values.

The main focus is still on the research question and on the subjects you examine, though, and not about the statistics. Here, it's basically a text about calcium levels. Statistical figures are tucked away at the end of the sentences, in brackets, as are group and measure names.

Note that this text aims to be as clear as possible, and not as neutral, as scientific, or as orderly as possible:

> **Results (not including tables):** Although in start of the study the calcium levels (Ca1) between control group and treatment group was not significantly different ($p>0.05$), blood calcium levels in treatment group after 15 days adaptation period (Ca2) and 12 h after calving (Ca3) was significantly higher than that of the control group ($p<0.05$). Also, in control and treatment groups the differences between the level of calcium before and after treatment in three different times (Ca1, Ca2 and Ca3) were significant ($p<0.05$).
>
> pH levels between control group and treatment group in start of experiment (pH1) was not significantly different ($p>0.05$), urine pH levels in treatment group after 15 days adaptation period (pH2) and 12 h after calving (pH3) were significantly lower than in the control group ($p<0.05$) at similar times. In treatment group the difference between before treatment and after treatment in three different times (pH1, pH2 and pH3) were different significantly ($p<0.05$). In control group the average of urine pH levels in all of times (pH1, pH2 and pH3) was not different significantly ($p>0.05$).
> – *Goodarzi et al. (2012)*

If you write longer texts, do provide some explanatory text around the results you present. Repeat the research question (or the part you're currently examining) so that readers know which part of your research is presented. Then, explain what the results mean in the context of the research question, so readers don't have to wait until the discussion section to find out.

The Discussion

The discussion states what the results mean for the world. Here, the authors can

- Sum up the results.
- Talk about how the results change what we already know from the literature.
- Talk about how the methods they used have worked out (and what problems they encountered).

In the following text, the authors also add a separate conclusion paragraph, which is a summary of the discussion. This is optional (as it has much of the same function and content as the abstract).

DISCUSSION

Optimum urine pH for close-up dry cows is about 5.5 to 6.5 (Davidson et al., 1995). In this experiment, the average of urine pH levels in peppermint-fed cattles shows peppermint can induce mild metabolic acidosis and normal blood calcium by a negative DCAB or any unknown mechanism. Metabolic acidosis increases tissue response to parathyroid hormone and lead to increase calcium resorption from bone; also parathyroid hormone receptors in bone are less functional at high blood pH (Horst et al., 1997).

The most common strategy employed to achieve this negative DCAB is the addition of anionic salts to the diet of pre-calving cattle (Goff, 2004). Anionic salts are expensive, significantly increasing feed costs per day for the close-up group. They are unpalatable and can reduce dry matter intake. Significant reductions in dry matter intake near parturition can predispose animals to metabolic disorders such as milk fever, displaced abomasum and ketosis (Moore et al., 1997) but Ando et al. (2003) reported that ruminal pH was significantly lower in the peppermint-fed steers than in the control steers and Peppermint feeding had no adverse effects upon ruminal fermentation and nutrient digestibility. Also antispasmodic and antifoaming effects of peppermint oil may play an additional role to induce a good palatability (Grigoleit and Grigoleit, 2005a, b).

Peppermint similar to alfalfa is high in calcium but one of the classical strategies often proposed for milk fever prevention

is the restriction of calcium intake pre- calving. This strategy is not a practical alternative for milk fever prevention on farms using grass or grass silage as a large component of the dry-cow diet (Wilson, 2001).

CONCLUSION

Peppermint feeding by inducing mild metabolic acidosis and normal blood calcium can be a new preventive method for milk fever and subclinical hypocalcaemia in transition Holstein cows. This method can be better than other preventive strategies such as addition of anionic salts to the diet and calcium restriction; because of tow-reasons: First, peppermint is cheap, significantly decreasing feed costs per day for the close-up group. Second, it's no adverse effects upon ruminal fermentation and nutrient digestibility.

The References

Generally, more references are better:

- They make you look serious.
- They help researchers dive into the subject.
- They lower the chance that the reviewer rejects the paper because you have not considered important work in the area: Namely, his or her own papers. (Few reviewers would say so outright, but of course reviewers are more favorable towards articles that cite them).

Now how to format references is surprisingly complicated. Most publications follow APA style (that is, the American Psychological Association) – if only because they were first, they published a book on the subject, and it's a huge piece of work to cover how to reference every type of publication there is.

You can look up APA style by:

- Buying the APA style guide (you can get it from the APA, from Amazon or from mostly any other bookstore).
- Googling „APA reference style“ or „APA citation stlye“.
- Using one of the large number of tutorials, such as this one.

As a ground rule, the reference includes all of the author‘s names, the year of the publication, the title, and then all the information you need to locate the publication (such as: in which journal and on which page of the journal it appeared).

Andoa, S., T. Nishidab, M. Ishidab, K. Hosodab and E. Bayarub, 2003. Effect of peppermint feeding on the digestibility, ruminal fermentation and protozoa. Livestock Prod. Sci., 82: 245-248.

Davidson, J., L. Rodriguez, T. Pilbeam and D. Beede, 1995. Urine pH check helps avoid milk fever. Hoard's Dairyman, 140(16): 634.

Goff, J.P., 2004. Macromineral disorders of the transition cow. Vet. Clin. North Am. Food Anim. Pract., 20: 471-494.

Grigoleit, H.G. and P. Grigoleit, 2005a. Pharmacology and preclinical pharmacokinetics of peppermint oil. J. Phytomed., 12: 612-616.

Grigoleit, H.G. and P. Grigoleit, 2005b. Gastrointestinal clinical pharmacology of peppermint oil. J. Phytomed., 12: 607-611.

Horst, R.L., J.P. Goff, T.A. Reinhardt and D.R. Buxton, 1997. Strategies for preventing milk fever in dairy cattle. J. Dairy Sci., 80: 1269.

Houe, H., S. Ostergaard, T. Thilsing-Hansen, R.J. Jorgensen, T. Larsen, J.T. Sorensen, J.F. Agger and J.Y. Blom, 2001. Milk fever and subclinical hypocalcaemia-an evaluation of parameters on incidence risk, diagnosis, risk factorsand biological effects as input for a decision support system for disease control. Acta Vet. Scand., 42: 1-29.

Jadhav, S.D., M.S. Bhatia, S.L. Thamake and S.A. Pishawikar, 2010. Spectrophotometric methods for estimation of atorvastatin calcium form tablet dosage forms. Int. J. Pharm. Tech. Res., 2(3): 1948-1953.

Kimura, K., T.A. Reinhardt and J.P. Goff, 2006. Parturition and hypocalcaemia blunts calcium signals and immune cells of dairy cattle. J. Dairy Sci., 89: 2588-2595.

Moore, S.J., M.J. VandeHaar, B.K. Sharma, T.E. Pilbeam, D.K. Beede, H.F. Bucholtz, J.S. Liesman, R.L. Horst and J.P. Goff, 1997. Varying Dietarycation Anion Difference (DCAD) for dairy cattle before calving. J. Dairy Sci., 1: 170.

Mulligan, F., L. O'Grady, D. Rice and M. Doherty, 2006. Production diseases of the transition cow: Milk fever and subclinical hypocalcaemia. Irish V. J., 59: 697-702.

Wilson, G.F., 2001. A novel nutritional strategy to prevent milk fever and stimulate milk production in dairy cows. N.Z. Vet. J., 49(2): 78-80.

10.2. Writing a great Career Paper

The first few papers you will write are typically not for publication, but to prove to a professor or lecturer that you are in command of the scientific method. These papers are very similar, yet very different from actual research papers, which makes it very hard to come up with reasonable standards for them.

The peer paper (which we have discussed in the previous section) has the following message:

Look at these results!

Think about what they mean for the world!

Or, more precisely:

I'm not important! But look at these results and think about what they mean for the world!

Now, career papers have a very different message:

Hey! Look at me! I can do science!

Of course, you should have nice results in your career paper, too – but it's not what it's about. It's about showing that you can do good scientific work. Good results are a bonus.

Which leaves you with a difficult task:

Telling people that you are a good scientist without talking about yourself.

More to the point, the following things are discouraged in scientific writing:

- Making any statement about yourself.
- Talking about anything that is considered „common knowledge“ among scientists (such as basic statistics and how you arrive at valid conclusions, namely any of the stuff you have learned on your way here).

So let‘s go through all the parts of a paper again and discuss the changes that a career paper has. Remember that all of this is 100% pure undiluted compromise: Showing off your scientific skills in any way that is not technically considered showing off.

Abstract

The abstract is pretty much the same as in a peer paper. It‘s usually longer and not as hard-hitting, and it should ooze *effort*. People who read it should realize that a lot of work went into what you‘re presenting here.

Also, because this is your first piece of research, chances are that you do not have stellar results to show. So make sure readers are impressed by the amount of work you put in.

„Introduction“

The introduction is usually a significantly bigger portion of the work than in peer papers. Here is where you show that you have read and understood the literature, and that you can draw conclusions and interpret one work in the light of all the others.

So: Talk about the available literature, the more the better. Don‘t just mention it (as you would in a peer paper), but *work it*. *Squeeze it*. Tear it apart and put it back together. Connect it, draw conclusions, find opposing views and put them against each other. And in all of this, always stay respectful, so you show that even if you disagree, you keep your style and appreciation of the scientific method.

Also, if your reviewers (or their previous students) have published in the area, make sure to include their work.

Because this section is much bigger, it‘s not typically called „introduction“ anymore. Sometimes it‘s split across several chapters. Also, many career papers include a special chapter about hypotheses. This shows that you can extract scientific hypotheses from the literature (and that you know what a Null Hypothesis is).

Methods

In a peer paper, the methods show what you have done, so other scientists can follow it (and try it for themselves if they want to). In a career paper, you have to do this too, and carefully. But there‘s more: In a career paper, you have to reflect the methods you use, at least to some degree. Tell readers why you chose the empirical method you use, which existing approaches you pick up and why you have modified them – nothing says „scientist“ as clearly as a methodic, careful and deliberate approach to your study. Also it‘s very boring to read, so it‘s usually cut short in peer papers.

The one thing where you can be brief is in the choice of the statistical method, because:

- If there is any choice, there is usually no good reason why you use one method and not another.
- Your readers typically don‘t care.

Peer papers are typically limited in their scope: Often they present just one experiment. Career papers often feature large amounts of research and data, so there is more to talk about.

The way in which you describe the results is typically very similar, although there is a tendency to include more detail in a career paper – such as background information that can help understand them, or a brief discussion of the one outlier in figure 2.

The main issue many researchers have here is how to present large amounts of results, also because this is not something they know from other papers they have read. Here are some questions and answers:

Do I have to present all my results?

No. If some of it did not work out, you don't have to include it. You are free to just focus on the interesting stuff.

Do I have to present the results in the same order as the hyotheses?

No, you can use any order that makes sense to you.

Do I have to cite the hypotheses when discussing the result?

No, the result already includes that information. *If* you choose to present the hypotheses, make sure to label them appropriately, so they're not confused with actual results.

How much context do I include?

Every section of the results should stand on its own. Readers should not have to go back to the methods or forward to the discussion to understand what you are talking about.

It's perfectly okay to include some redundancy. Academic readers are very good at skipping over the stuff they don't care to read.

Discussion

The discussion has essentially the same content as in a peer paper: It says what the results mean for the world in general and for the area in which you did the research.

However, in a career paper, a main focus of the discussion in that you carefully reflect what you did. Which parts worked well, which could be improved the next time around (and how). Don‘t overly degrade your work, though: You‘re still showing off how good a scientist you are. You‘re just adding another layer that shows that you can recognize problems and deal with them. Show your readers that you care, that you are passionate and that you want to (and did) improve.

Because the career paper is about you, the discussion should contain ample text covering what *you* did, and what you think about it (even if you use other people‘s works to express those thoughts).

Reference

The reference section works the same as in a peer paper: Include lots of references, especially from the people who will judge your work.

A note on the style

Here, I have two pieces of advice for you, which are contradictory, so take your pick:

- Use simple style so your paper is easy to read.
- Use a moderately complex style so it looks like you‘re a professional scientist.

You'll find that a lot of the best peer papers use very simple style, but it can happen that this looks simplistic to some people who will judge you.

There is also a cynical take on this (and I wish I were just joking), and it works like this: In order to fail you, a reviewer needs to have read and understood your work. So if you hand in a large, complicated, impossible-to-understand career paper, this might actually increase your chances of passing. However, I recommend that you do not try this.

When you find nothing, you can write the best paper

So what if you do not find any statistically significant results? – Then you may be about to write your best paper, if you have the guts to do so.

First of all, you're not punished for lack of results. So you can just write the same paper you would have, except that the results are not so captivating and the discussion probably will trudge along, too.

Second, you can also cut all of this short, and shift the focus of your paper to the methods. So instead of asking

What are the results?

you ask

Why did I not get any results?

This involves more reading, usually. Find out what the difference is between your study and everyone else's. Make a fine-grained analysis of what your subjects did, one thing after another, get as much information on what was happening inside their skull, read up on decision making and information processing literature. Ask friends and other researchers (mail them, if you need to – most researchers are happy if people take an interest in what they do).

Then, sit down and write a compelling text about how to do great research the next time around. Tell readers which additional variables need consideration, because they can make or break a research effort. Form hypotheses about what really has an effect and what does not. Make suggestions for

experiments that test this. Talk about which other lines of research feed into what you were doing, which you (and other researchers) had not noticed.

In short: If you do this, you'll have a great platform to show off all your scientific muscle and talent, and to write a compelling text. *Do so!*

10.3. Underestimating the effort

Above: *This is what finishing a research project looks like. You have been warned. It's also what finishing a script looks like (it's shortly past midnight one week before the semester starts as I'm writing this).*

To my knowledge, nobody ever over-estimated a research effort. Research always takes up a lot more time than you thought it would. There is one simple reason for this:

You can solve practically any problem that comes up in planning simply by increasing your research effort.

Judging the effort

As a rule, practical research efforts follow Parkinson's law:

Work expands so as to fill the time available for its completion.

Academic research, in contrast, follows Hofstadter's Law:

It always takes longer than you expect, even when you take into account Hofstadter's Law.

There are, I think, two reasons for this: First, academic research is often bigger and more complex than practical research, and second, deadlines are less strict. So an academic

researcher might hope to hand in the thesis by June, but knows that September will be fine, too.

So how can you make sure that your research effort is somewhere within reasonable bounds? – Actually, you can't, but if you care about your mental health at all, then **do all of the following:**

- **Plan enough time for the writing** (you'll probably spend more than half your research time analyzing data and writing about the results).
- **Stick with the original idea.** If anybody (including, above all, yourself) has additional ideas, then save them for the next round. It's always great to expand on your work later.
- **Do not interpret questions from your supervisor as work assignments.** If your supervisor says „maybe that's different in Australia", that's not an assignment to fly there and check. If you're unsure whether it is, ask them, and point out that this would be a fascinating question *for future research.*

Catching up

If you follow all the above advice, you'll probably still run late, so here's some advice how you can get back on track:

- **Get help**. In most situations, there are perfectly acceptable ways to get help: You can get it from your supervisor, usually from friends and paid professionals as long as you mention it in the paper, or from the company you do the research for if they decide that it's worth the extra cash. There are some student papers you have to do all on your own, for the rest, help is available. Use it.
- **Cut short**. There is no obligation that your report covers everything you set out to research. If some hypotheses have not worked out in any meaningful way, you have the option of skipping them and focus on the stuff that has delivered results. Usually, that is also easier to write about.
- **Plan to work less and slower**. If you're running behind, your instinct tells you that you have to pick up speed. That

won't work: The problem is precisely that you can't work as fast as you thought you would. Over time, this will not get better. So if you adjust your plans, then plan to slow down some more as your batteries run empty and your efficiency to handle your own text decreases. I'm a professional author and I can go over my own text maybe four times before it starts to blur because my brain refuses to focus on stuff it knows perfectly well already.

For your education and entertainment, here are two studies that are really the extremes of efficiency and effort.

The most efficient study ever done

The reward for the most efficient study ever must go to Charles Tart, Professor of Psychology and researcher for extrasensory perception (ESP). Hold your skepticism (and I don't care whether you believe in paranormal effects or not), this is some of the finest and smartest research that anyone ever designed.

Charles Tart was researching whether a person in a sealed-off room could correctly predict the playing card that a person in another room was looking at.

The setup was a basic experimental design: The experimenter sat in a room in front of a dial that showed 10 cards. He focused on one card at random and pushed the button next to it. At this point, the subject was notified, who sat in front of a nearly identical apparatus, and tried to press the button next to the card the experimenter was focusing on. The subject was told whether he or she was correct, and the machine was reset for the next trial.

How is this the most efficient study ever? – Well, according to the Null Hypothesis, performance should be entirely due to chance, and being right or wrong in the first round would not influence the result of the second one at all. This means that every button press is an independent event. Charles Tart's sample size was the number of button presses, not (like everybody else's) the number of subjects on the study.

That's maybe a bit of a stretch, but perfectly okay according to all requirements of scientific measurement. You can count

anything as an observation, so long as it's statistically independent from the other observations you make. Almost always, this boils down to using one person as one observation, because if you perform your experiment twice with the same person, you expect (roughly) the same results. Now Charles had found his little niche where that did not apply, and used it like a Pro.

Charles tested his subjects for 20 runs of 25 trials each, netting a neat sample size of n=500 from every one of his subjects. Now, at that size, almost any slight deviation becomes statistically significant, and the effects he found were anything but slight. His subjects (which were carefully pre-selected for ESP ability) scored as many as 124 correct hits (when 50 were expected by chance). The probability of such a result happening by coincidence is less than one in a billion billions. According to any scientific standard, Charles Tart has proven beyond any doubt, in one day, with one subject, that ESP exists.

Now what do we make of this? – Faced with results like these, most scientists will look really closely at alternative interpretations, such as experimenter errors, failures of the test apparatus or really anything. I know I did (and there aren't any, except assuming that he flat out lies). Looking for alternative explanations is perfectly reasonable, because you don't want to topple most of established science on a whim. It's also extremely unscientific, unfair and lame. Scientists are supposed to look for new discoveries and judge them on the basis of the available evidence. They are not supposed to judge perfectly good evidence on the basis of how well it fits with how they think the world works.

That said, I have no idea whether ESP works or not. Probably, we'll never find out, because nobody will dare to do much independent research on it. It's a bit sad that physicists get to research all the cool crazy stuff like time travel and parallel universes, while empirical scientists have to cling to the sane stuff.

The least efficient study ever (almost) done

Okay, it's unfair (and probably wrong) to single out one study, but here we go anyway: The least efficient study to my

knowledge was Switzerland's project to get definitive answers about the causes of mental illnesses.

Mental illness is very well researched, especially the psychological treatments of it: Google Scholar has *nearly two million results* on behavior therapy alone – compare this with the *three unpublished studies* you need to get a new drug approved. What is less well known is where mental illness originates. Almost all patients you find have multiple problems with their mental health (as well as with a lot of other things), so it's hard to know what is the cause and what is the effect.

A valiant, but doomed effort to find this out was the project „SESAM". The idea was to follow people through their lives from infancy, so you'd know once and for all. However, there immediately were many, many problems:

- You need to follow a lot of people over ten or twenty years, which costs buckets of money.
- Also, you need the people. SESAM was stopped because they could not get enough subjects willing to participate. It did not help that they had to ask pregnant mothers and among the first planned measurements was a DNA sample of the infant.
- Most psychological illnesses are (thankfully) very rare: Only around 0.5% of people ever develop schizophrenia in their lives. That is, if you plan to observe a (statistically very modest) 20 schizophrenics, you need to follow 4000 subjects for all their lives, or 8000 for half of it.
- Even major depression, the most frequent mental illnesses, affects only about 10% of people in their entire lives. So even if you're just looking at depression, *90% of the data collected will be in vain* (depressive symptoms are much more common, of course, and bad enough, just not quite as crippling as a major depression).
- Typically, the participants in such studies turn out to be healthier than average: They come from privileged backgrounds (because it's easier to talk educated people into participating in university research) and they will seek help sooner if a problem develops (because they are aware of the issues and spend a lot of time with psychologists, whom they perceive as nice, helpful and knowledgeable).

Not going ahead with SESAM cost around 10 million Swiss Francs (around 10 million US dollars). Going ahead would

have cost around the same amount again, and it's doubtful that there would have been any strong results except for major depression. And even in the face of the results, it might have been difficult to act on them. So in 20 years you will know what caused depression back in the old days? That's great if you're a historian, but everyone else will still have to guess whether any of the results apply to the current generation as well, so it might just not be the huge breakthrough it seemed like.

10.4. Seven mistakes you'll make in your first paper (unless you read this chapter)

Just to be clear on this: Any mistakes you (or the researchers cited here) make do not de-value the results of the research. So long as you try to do good research, the results are valuable. Even the professionals sometimes make glaring mistakes (such as confusing data labels, computing scores wrong, or just using a test that does not work for what they do), and that's just as well because if they find anything extraordinary, somebody will try to confirm it, and no harm is done if they can't.

So when you try to avoid errors, it's because you want to look professional, not because your p-values gain any more weight if they are formatted correctly.

All the research cited here is from open access journals. As such, the articles are peer reviewed, but it seems that the publishers were lazy. Or that they are in areas of research where people just do not care to follow APA style. Still, I do not point out which authors made which mistakes.

Including program output

To you, the output of your statistics program might look very sophisticated. It's not. It's just a bunch of numbers, and you're supposed to pick the one or two that really matter and then put them in the text. „Text" meaning where you tell readers what actually happened in your study.

Here's an example where an author compares performances on grammar questions in two conditions: In the presence of high-frequency words and in the presence of low-frequency words. The author presents the output of the statistics program (SPSS, in this case), with little text around it.

Wrong

The correlation between two variables:

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	LOW & HIGH	28	.676	.000

There seems to be a strong positive correlation.
The results of the Paired Samples T Test:

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	LOW - HIGH	.3571	1.9854	.3752	-.4127	1.1270	.952	27	.350

The T value= .952
Degrees of freedom=27
The significance is .350

Better

There was a strong correlation between the subjects' performance in the low-frequency condition and their performance in the high-frequency condition (r=0.67, p<.001). However, performance was not significantly different in the two conditions (t=0.95, p=.35).

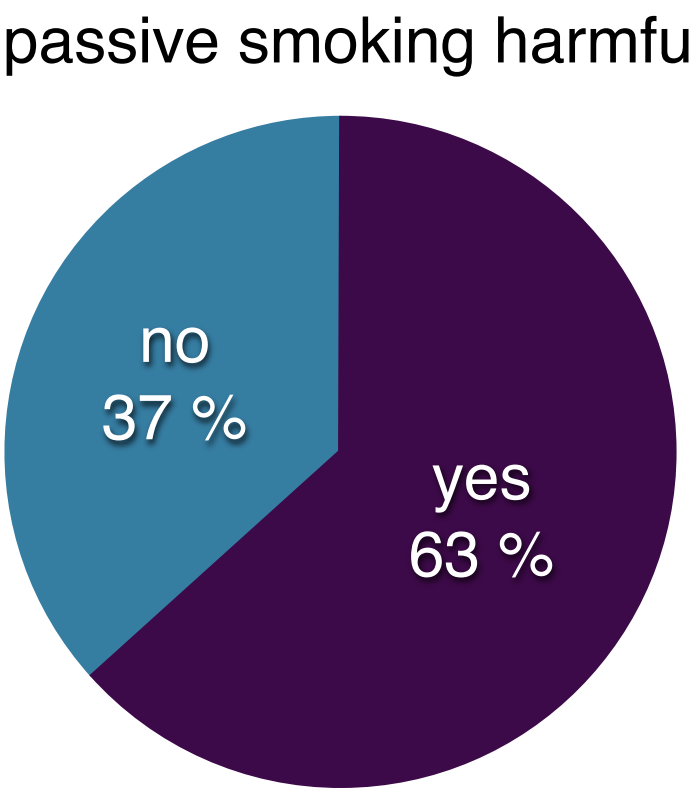
Using pie charts

Most laypeople think that pie charts are synonymous with statistics, yet most experts never use them in their articles. So:

1. Don't use them.
2. If you have the space and think they would work well, ask your boss/mentor whether they think so too or whether they would prefer if you stick to the professional style of presenting your numbers.

Here, I'll use a pie chart from before (I could not locate an article that actually uses one – even if publishers might be lazy in enforcing scientific standards of writing, they're also greedy and don't like to print large images when a few lines of text will do).

Wrong



Better

Overall, *37%* of articles concluded that passive smoking is not harmful. – *Barnes & Bero, 1998*

Writing about statistics, not subjects

When you do research about people, write about people. Statistics is just the filter that tells you which of the findings are genuine and which are not. After you have applied it, *go on talking about the findings*, not the statistics.

Also, don't talk about means, talk about people being better or worse. If you have any difficulty, try starting your sentences with the word „Subjects".

Wrong

Based on the findings, the low frequency word test mean score is a little higher than the mean score of the second variable.

Better

Subjects showed slightly better performance on the low frequency word test than on the high frequency word test.

Writing about hypotheses instead of results

Most articles do not mention the Null Hypothesis (or the Alternative Hypothesis), ever. Hypotheses are just tools to make students reflect what they‘re doing.

Also they‘re boring: You always try to reject the Null Hypothesis, and if you succeed, you can take the Alternative Hypothesis as a scientific fact. There is no need to go explain this while discussing the outcome of your research.

So: If you have to, write the hypotheses in your paper, clearly mark it with „Hypothesis“, so readers don‘t confuse it with an actual result. When talking about the results, don‘t talk about the hypotheses.

Wrong

As the significance value is greater than .05, the null hypothesis is supported. Consequently, there is no relationship between word frequency and answering grammar questions.

Better

There is no relationship between word frequency and answering grammar questions.

Separating results and interpretation

There is a difference between these two statements:

- When judging the performance of harvester operator, the measurements of the two raters correlated with each other and with the actual performance of the operators.
- The raters could correctly judge how good the operators were.

The first statement is a research result, the second one is an interpretation. That is: The first one is fact, it's just a way of expressing the numbers that were measured. The second one tells us what it all means.

Now there is one tradition that says you should separate these two: Results belong into the results chapter, interpretation is part of the discussion. That's wrong, and here's why:

- If the interpretation is obvious, there is nothing wrong with putting it into the results chapter.
- If the interpretation is not obvious, you have to put it into the results chapter, *or the readers won't know what the results mean*.

In other words: If there is any doubt as to what the results mean, it's your job to clear it up. Immediately, not 50 pages later, because your readers can't keep stuff they don't understand floating in their heads for so long.

Of course, a research paper can have a result section that is as short as two sentences, in which case the authors are correct in putting the interpretation into the discussion section immediately afterwards. In all other cases, do tell readers what the results mean, in all the detail required so that they understand what is going on here and how it relates to your research questions.

Getting the table format wrong

Professional researchers look at tables first, because that is where information is densest. If an article contains tables, you can probably understand half of it just by understanding what the tables want to tell you.

Also, table format is easy to do right and it's equally easy to do it wrong, which makes it a good marker of whether any student papers (or dissertations, for that matter) follow scientific code or not. So do yourself a favor and read up on the table style (and follow it rigorously).

The short form is:

Do not use any vertical lines in a table.

Following the wrong examples

When you are writing a paper, you will be unsure about the style you are supposed to follow. So why not take a paper that was well received and follow the example?

Here's why not: Because a number of factors determines scientific fame and recognition, such as the amount of work that went into a paper or whether a famous author wrote it. So when picking an example, proceed as follows:

- **Do not use student papers or dissertations,** even if they received perfect marks. These marks are given based on a number of factors, formal correctness being a lesser one. Also, these papers were typically not returned for revisions, so any errors that the authors made are still in them.
- **Do not use papers by eminent scientists.** Most publishers gladly print anything that a famous person writes for them, even if it's below their usual publishing standards.
- **Do use papers from peer reviewed journals**. Especially from Psychology journals, which have very high formal standards (especially those expensive APA journals). If something is published there, it was read by two

scientists who publish in the same field and by an editor, and anything anyone found objectionable was changed. Sometimes multiple times. So if it's in the journal, it's bound to be formally correct.

11. *Appendix: A history of empirical research*

This was originally the start of this book: It's how modern science started, what it is, and generally a compelling read, especially if you're into history, like I am.

Also, it's not strictly required if you just want to do good research. So have an enjoyable and relaxed read.

11.1. What is science?

Some things are obviously science, such as Math, because it's tough and mathematicians can prove they're right. For other fields, there is no formal definition. Scientists like to think that there are obvious and large differences between what they do and what fortune tellers do, but it's surprisingly hard to put that into words.

To start, I'll follow people like Jacques Barzun and use the following definition:

Science is everything that can convince scientists.

Or, in other words:

Science is the set of beliefs that scientists share.

That said, science is like religion, it just does not have prophets and revelations. To convince a scientist, you cannot point to a sentence in a particular book, and you cannot have a high priest declare what is true and what is not. You need other methods. This section is about those methods, and how to justify them.

Science is not truth

Contrary to what you may think, truth does not have anything to do with science. For example, String Theory is a pinnacle of a scientific theory (it might explain how the universe works), yet scientists cannot prove that String Theory mathematically even adds up, and all the predictions it has made so far cannot be tested with current equipment (although that may change in a matter of years).

Also, it turns out that you can never prove most statements about the world, starting as simple as this:

„All swans are white"

At best, we can say that all swans we have observed so far are white, but even if we observe every swan but one, there is a possibility that the last one is black, and the statement is false.

Or take the following example:

„All swans are either white or some other color"

This statement is certainly true. It is also completely meaningless. One might say it is a bit too true, that is, it is true to the degree that it cannot possibly be false, which seems to be problematic as well.

Science is argument

You can think of science like a giant courtroom, where ideas are tried. Scientists try to convince each other that their idea is true, they bring evidence, discuss it, and come to a verdict (or occasionally overturn an old verdict in the light of new evidence they have collected).

Now, there are different ways on how to convince a court. The old way was simply to be brilliant, like the detectives in the old novels and movies. By thinking hard, coming to conclusions, and finally offering a comprehensive theory that can explain all the evidence you have. Those were the days. A lot of science that was produced in that time reads like poetry: Beautiful, imaginative texts that try to capture the reader.

The one drawback is that if you have two theories that both look sensible (and have their brilliant advocates), you can't really decide which one is the right one. You can argue all day, or all century, and not get anywhere.

So there is the new way: Don't go for the big theory that can explain everything, but for the small fact. Sometimes that fact stands on its own, sometimes it contributes to a larger theory, sometimes it brings one down. This means that the modern courtroom of science is filled not with huge theories, but with the murmur of thousands of small, proven facts.

Oh, and also scientists are not the rock stars of the past (except the occasional literal rock star scientist), but can be fairly boring and uninteresting at times. They need not even be famous. Which is good news for all of us normal people: As long as we follow the methods of science, then we can contribute to it.

This chapter is about those methods and how scientists think that they should be able to convince each other.

11.2. How empirical research works

Above: *Human existence (top left) between physical rules (right) and randomness (bottom left). Note that Newton was never really hit on the head by an apple, but it's always fun to imagine, because he was a jerk.*

For the purpose of this chapter, let's divide all the things that ever happen in the world into three categories:

- **Physics**: Things that follow exact rules, all the time. Such as an apple following gravity and hitting Newton on the head with some force (which did not actually happen).
- **Randomness:** Things that follow no rules, like which number comes up when you throw dice.
- **Life:** Things that follow some rules, most of the time, in surprisingly complex ways.

Of the three, **physics** is the easiest to define as a science. If things happen consistently all the time, then any description of how and why that happens is science.

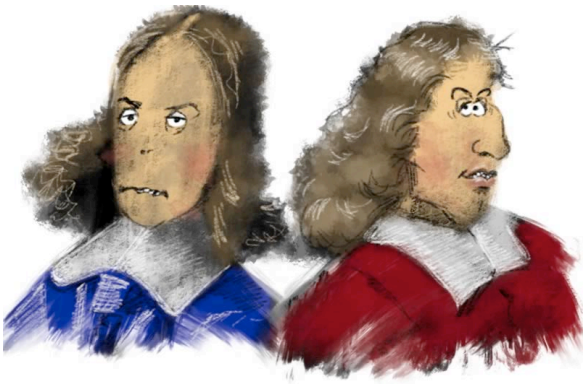
Early on, scientists like Freud (and many others) tried to find similarly consistent rules that apply to humans. Freud thought that there was a constant energy inside humans – much like the laws of thermodynamics in physics. This led to

complex theories that were argued by brilliant minds (who also could write scientific papers that read like literature). However, they could never really predict anything with quite the same certainty as a physical phenomenon.

In parallel to this, a second approach emerged, from the opposite direction. Mathematicians started to explore the concept of **randomness**. As it turns out, randomness follows very strict mathematical rules. Not the type of rules that tells you exactly what will happen, but you can figure which events are probable and which are not.

This, in turn, allowed empirical scientists to re-define how they researched **life**. They would not look for rules that were true all the time, but merely for things that were not random, however small or large. And how they got scientific truth from finding non-random phenomena is the focus of this chapter.

11.3. Pierre Fermat and Blaise Pascal invent probability calculus



Above: *Pierre de Fermat and Blaise Pascal, who created modern probability calculus (among a huge amount of other extremely complicated things)*

In 1654, Pierre Fermat and Blaise Pascal discussed the gambling habits of their friend, the Chevallier de Méré, and came up with the theory of probability in its modern form.

According to their probability calculus, whenever you observe some random events, like a dice roll, then a number of things can happen, and you can compute the probability of these things happening using some mathematical formula (which we won't go into).

Their theory is technically not a theory of science – it's a theory of how you can win at the gambling table. It's a description of how coincidence works, and coincidence was what Fermat and Pascal were interested in. In opposition to philosophers and natural scientists, who were interested in the phenomena that were the same all the time, such as gravity, the foundations of Mathematics, and the existence of god.

However, in creating probability calculus, Fermat and Pascal also opened up the field of empirical science:

Many things happen that are neither completely random not entirely consistent. That's what empirical science is about.

And now we'll see how other people took up probability calculus and ran with it.

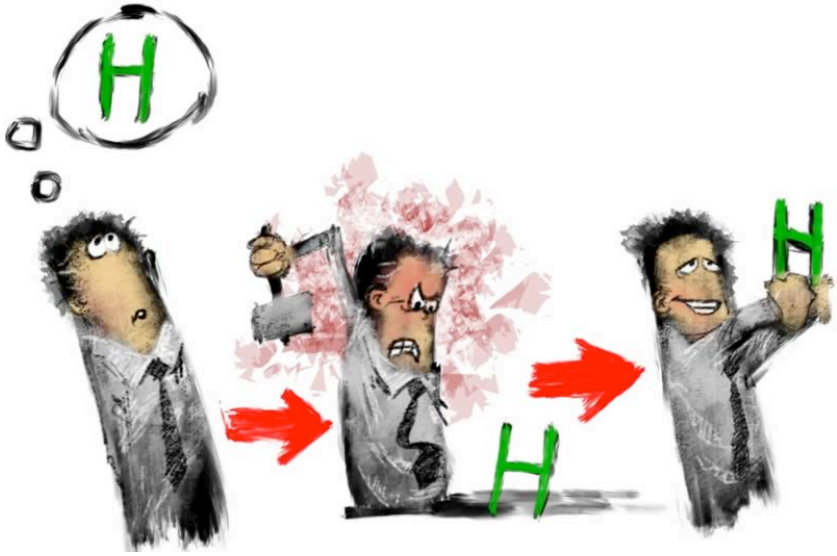
11.4. Karl Popper: Getting truth from probability

Above: *Karl Popper, philosopher and scientist*

Karl Popper developed a very influential theory about how to get scientific knowledge. It goes as follows:

- First, **make a hypothesis that is refutable**. It must be possible to make an experiment or an observation that will lead you to abandon it. A hypothesis like *„All swans are either white or some other color"* is thus unscientific and not worthy of a scientist's professional interest, because you can never possibly observe a swan that refutes it.
- Second, **if you fail hard enough at refuting the hypothesis, then it must be true**. If you have really tried to find a non-white swan for some time and have consistently failed, then you have corroborated the hypothesis that all swans are white.

Here, Popper moves from probability to fact: Any evidence that is less than 5% likely under your hypothesis refutes your hypothesis. (To be fair, Popper does not say „5%", but „whatever scientists think is a good number" – which is usually 5%).



Above: *Research according to Popper: You start with a hypothesis (which must be refutable), and then try to destroy (refute) it using evidence. If it survives, you accept it and present it to the world.*

Popper's theory has a few drawbacks, as far as I can see:

1. How can you trust scientists to try hard enough to destroy their own theories? And how do you define „hard enough"?

2. To destroy a theory, you need evidence that is less than 5% likely. However, you get such evidence by complete coincidence roughly once every 20 observations. That's the definition of „less than 5% likely". So Popper's method eventually would destroy all theories – the wrong ones fast, the correct ones slowly.

3. What if a scientist has a theory and proves it true? – According to Popper, the theory is unscientific because it cannot be falsified, yet it seems weird to categorically exclude anything that we certainly know to be true from science.

11.5. Ronald Fisher: Refuting the Null Hypothesis

Above: *Ronald Fisher: Probably the smartest man in statistics. And biology.*

Sir Ronald Aimler Fisher was (with Jerzy Neyman and Egon Pearson) one of the founders of modern statistics as well as population genetics. Fisher more or less single-handedly created much of today's statistical methodology and testing. Among this is the idea of the Null Hypothesis. It goes as follows:

- First, state the **Null Hypothesis**, which says that there are no differences whatsoever.
- Second, **make observations**.
- Third, **compute how likely your observation is, if the Null Hypothesis is true.** If it's very unlikely, then you can assume that the Null Hypothesis is false. From this, you can assume that the contrary of it is true.

This calls for an example:

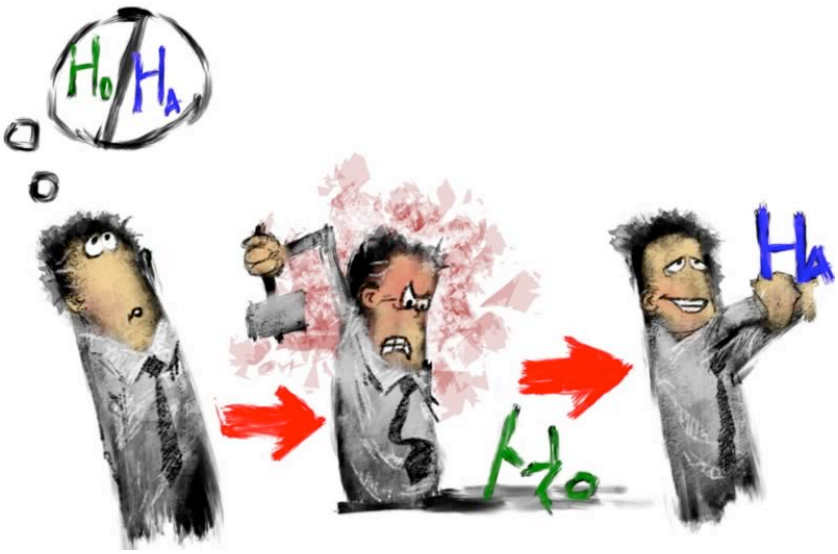
1. The Null Hypothesis is *„There are as many black swans as white swans“*.

2. We spot our first swan, and it‘s white. If the Null Hypothesis were true, there is a 50% chance of this happening, so it‘s no big deal.

3. We spot the second swan, and it‘s white. If the Null Hypothesis were true, there is a 25% chance that the first two swans we observe are white. That still happens.

4. We continue, and somewhat later, we have observed a total of 10 white swans in a row. Under the null hypothesis, there is a 1 in 1000 chance of this happening. Because that is very unlikely, I discard the Null Hypothesis.

5. Having done so, I assume its opposite to be true. In this case, I assume that there is a majority of white swans (using Fisher‘s methods, I can never prove that all of them are white).



Above: *Research according to Fisher: You start with two hypotheses. The null hypothesis says that there is no difference, the alternative hypothesis says that there is one. If you manage to destroy the null hypothesis, then the alternative hypothesis must be true.*

11.6. Popper vs. Fisher

If you have paid attention, then both Popper and Fisher provide a method to arrive at scientific knowledge, and their methods have certain similarities. Let's see:

- Both Popper and Fisher say that you start out with a statement that you can refute.
- Then they try to refute the statement (that is, they try to find evidence that makes it seem very unlikely that the statement is true).
- Popper says, if you can't refute it, then you can accept your statement as scientific fact.
- Fisher says that if you can refute it, then you can accept the opposite of the statement as scientific fact (because if the statement is false, then its opposite must be true).

In other words: We have two eminent and very smart persons who have found a way to separate a heap of statements into two buckets, one containing scientific truth, the other one garbage. The only problem is that they do not agree which is which.

So what do scientists do with that? – As far as I have observed, they ignore it. They mostly follow Fisher and Common Sense, tell students that Popper is important, and use Bayesian reasoning whenever Popper or Fisher would lead them to absurd conclusions, usually without noticing. More about Bayesian reasoning coming up now.

11.7. Thomas Bayes and the conditional probability

Above: *Thomas Bayes: Religious leader and mathematician*

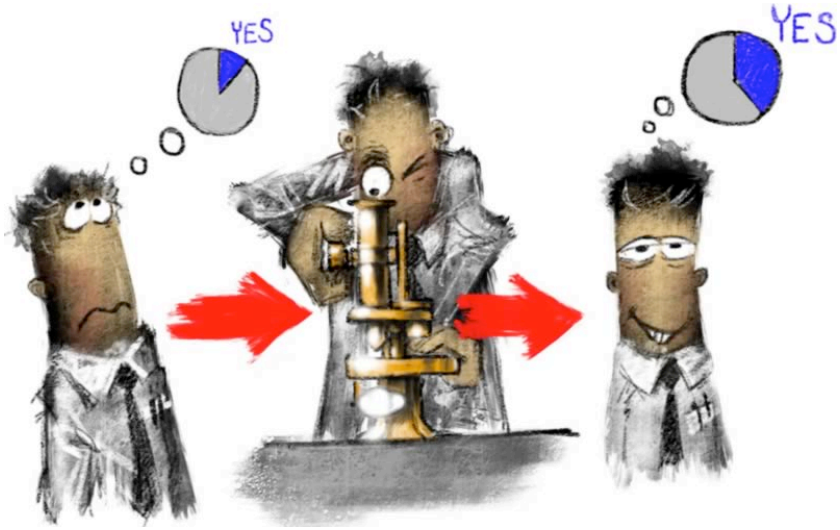
An entirely different approach to scientific truth uses a neat bit of thinking that the reverend Thomas Bayes introduced around 1760, and which has been expanded upon ever since.

Bayes said that you don't have to look at things as either true or false, or even as probable or improbable. What you do is you look at evidence, and at how new evidence changes your existing beliefs. In Bayes' thinking, that works as follows:

- You start with a hypothesis. Your belief in the truth of that hypothesis can be measured as a probability. You could be, say, 50% convinced that the hypothesis is true.
- Then you make observations. Depending on whether the observations are better explained by your hypothesis or by

any other explanations, you change your belief in the hypothesis.

- Bayes actually developed a neat little formula to compute how much you should revise your belief. It is based on how likely the observations are if your belief is correct, and how likely they are if it's wrong.



Research according to Bayes: You have a belief in a theory, then you do research, then you revise that belief depending on how well the research fits with the theory.

If Bayes and his modern-day followers are correct (which both Popper and Fisher vividly deny, by the way), then classical statistics is a fairly crude instruments, that works only under optimal circumstances, and even then not quite reliably.

However, very few people do Bayesian statistics today – it's also a lot more complex than it has to be, because it's mostly done by genius Mathematicians with little interest in getting it to a level that you and me can understand or apply. That said, Bayesian statistics points to a few flaws in classical statistics:

- Bayesians have degrees of belief in a theory. Once you get used to that, it seems strange that classical statistics is so obsessed with truth and falseness of hypotheses (given that we know that there is no absolute truth in empirical science).

- If you have five nearly significant results in a row, can that be coincidence? – Classical statistics can't decide that. Bayesians would take it as a strong indicator that the theory is correct.

I think that classical statistics is so successful because it helps us with the one thing humans are really bad at: Handling probabilities. It even turns them into yes/no answers for us, and that's stuff we're good at. Also, humans are quite reliably to use Bayesian arguments when faced with absurdity. For example, if I drop a ball and time its fall, and the results contradict Newton's theory of gravity, I'll probably assume that my measurements are wrong, and not Newton. That's essentially a Bayesian argument: My belief in Newton is stronger than the result of the experiment.

11.8. Thomas Kuhn and the Paradigm Shift

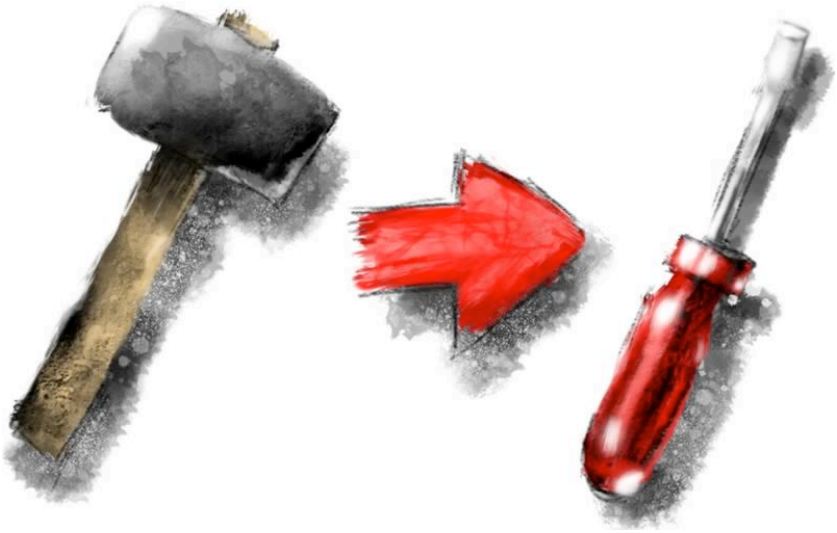
Above: *Thomas Kuhn, one of the most hotly discussed philosophers ever. A Google search for „thomas kuhn idiot" returns nearly five million documents. It seems you can't mention him without offending people.*

Thomas Kuhn was one of the most influential philosophers of science, and he is responsible for the terms *„paradigm shift"* and *„scientific revolution"*. Or people around him are. Who might or might not have understood any of what he said. It's all a bit fuzzy.

What is certain is that people love to quote „paradigm shift" and „scientific revolution", and they love to think that their interpretation of what Kuhn said fundamentally explains how science works. Also they love to call people idiots while they talk about Kuhn. Anyway, the theory goes something like this:

- Scientists create a large body of scientific evidence, following established scientific methods, thus arriving at a mostly coherent understanding of their field of study.

- At one point, they get stuck, and somebody proposes entirely different methods to get completely new results. This new view of science (and the world) is incompatible with the old one: If you follow it, much of the old thinking is wrong, and if you follow the old thinking, the new one is mostly meaningless in reverse.
- Then, a paradigm shift can happen: The scientific community can form a consensus that the new approach is better, and it becomes the standard way of doing things.



Above: *Scientific progress according to Kuhn: People start using different tools and get new results which make no sense in the old ways of thinking.*