Vince Penders

PIRATES, PEACHES AND P-VALUES

Production and distribution: Mosae Verbo, Maastricht
www.boekenplan.nl

Cover and illustrations: Chelsy Penders and Luca Britti

ISBN 978 90 8666 406 1
NUR 123

TABLE OF CONTENTS

THOSE WHO CREATED SALUTE YOU

Welcome, dear reader! A brand-new handbook is sparkling in your hands, and boy, are we proud of the result. As its makers, we'd like to introduce ourselves briefly. Is that all right?

I'm **VINCE PENDERS**… feel free to call me **Vincenzo**. The brain behind this teaching method – that's me. Creative to the bone, a fan of cooking, baking, music, and Japanese video games. Just about anything interests me, especially if it's about society and humans. I used to teach at Maastricht University for a living; nowadays I teach students independently, still with formidable pleasure, and I write novels. These novels are not about statistics: they're thrilling tales with a touch of science fiction! My debut *Zwaluwhart* (let's call it *A Swallow's Heart* in a free translation) was published by Macc in May 2015. It's a great way to relax after passing your statistics course... if you can read Dutch, that is.

My sister **CHELSY PENDERS** is at least equally creative, but in her own way: she sings, dances, aspires a career as an interior decorator, and is darn good at drawing. All the illustrations for this book have flown out of her pencils, as well as the graphic design, with the fruity colours and slick tables. Without her insight, the book and its accompanying website would have been a lot more boring, amateurish, and sometimes a bit painful to the eye. She enlisted Luca Britti's Photoshop skillz to apply the finishing touch.

And this is **WARD SCHOONBROOD**, my good friend. He's an entrepreneur and independent advisor, focusing on creative industry in the region. A family man who cares about his hometown Maastricht. More or less single-handedly, Ward joined the crew when I presented my idea for a new teaching method to him: he largely designed our multi-media approach, and came up with the project name on the side – Pirates, Peaches and P-values. He maintains the website and is responsible for all our video material. "So actually he's got nothing to do with the handbook?" Stupid question, reader. Ward is simply a part of us. ☺

What remains are a few honourable mentions as ways to express my gratitude. The first one goes out to René Bouman, the publisher who helped us realise this book. The second one is for Maastricht University's department of Methodology and Statistics. *Pirates, Peaches and P-values* is a fully independent project, but without the department's teachers with their patience and competence, I would never have learnt all that I can now pass on to you, dear reader. Finally, a big hug to all the students I've been allowed to assist for the past years, who have inspired me, kept my passion burning, and convinced me to launch this project. It's shooting for the skies as we speak!

READERS' RESPONSES

"Are you struggling with statistics? Well, this is your chance – buy this terrific book! Not only does Vince explain everything quite clearly, but he also uses funny examples you'll never forget. During the exam I still thought of Luigi, Mario, the pirates and the cupcakes. It made me laugh for a second, but it really did help! It gives a nice twist to statistics, which makes you understand all those different t-tests and ANOVAs better after all. The book has helped me quite a lot to get a better overview and understanding of statistics, hence, highly recommended!"

~ *Joëlle*

"This book gives a 'twist' to the usually dry data of statistics, by using imaginative data. It's quite informative and clearly structured, and offers a transparent, step-by-step elaboration of formulas and methods. Very effective to pass statistics exams, but very suitable as a reference book as well."

~ *Rosita*

"Superb book! Before I had it, I had a limited oversight of statistics, but this book was my absolute saviour! Very clear everyday examples are given and thanks to the summarising tables, everything is really easy to track down. Whether you have trouble with statistics or not, PPP is truly absolutely recommended!"

~ *Jisse*

"Learning in a fun and interesting manner – you won't regret buying this one."

~ *Laura*

"This book is splendid! I think it's a great addition to anyone, how good you are at statistics doesn't matter. I've never had a lot of trouble with statistics myself, but still there are always moments when you forget something or can't get to a certain step on your own. This book describes everything so clearly – not only is everything explained clearly to really everyone, it also makes sure that each piece of dry statistics becomes an adventure or a story. I can recommend Pirates, Peaches and P-values to anyone, the book is quite complete and of excellent quality. Using this book has helped me a lot, you can find things quite fast and the explanations are often quite clear and easy to understand. I don't think I'll dispose of it later either; it's always handy to keep at hand."

~ *Kobus*

"For years Vince has been known to us at the university as the hero of statistics. When you're a statistical wreck, his book is the place to be. Its methods have dragged me through my multiple statistics courses time and again."

~ *Niki*

"This book is truly a godsend for anyone who must deal with statistics and needs extra help with it. The book is written such that beginners can easily tag along, but advanced readers are still offered interesting explanations. It provides simple, logical, but also hilarious examples to elaborate theories, which makes understanding and remembering the subject matter a lot easier. Everyone who's forced to deal with statistics in his or her studies will have experienced that a point is soon reached when going any deeper is 'no fun anymore'. The humour and simplicity that PPP flaunts prevents this point at which you'd normally fling the book into a corner. Being a University student originally from MAVO level (lower Dutch secondary education, ed.), I was really completely inexperienced in the field of statistics, and each time it was a (terrible) stumbling block. The official literature was simply too abstract. This book has managed not only to clarify the basics for me, but the more advanced topics as well. It's absolutely recommended for anyone who is interested in statistics and would like to experience it more lightly for a change, but certainly also for those people for whom it may be the only solution to make some sense of all the Latin and Greek 'misery', and gain insight in the importance of the dreaded subject called statistics…"

~ *Dennis*

"Absolutely unrivalled."

~ *Angie*

"Pirates, Peaches and P-values is a tiptop book. I can only say positive things about it! The structure is transparent. The application of theory to practice is super clear. The exercises are particularly illustrative and described with a love for detail. Moreover, corresponding SPSS output is explained. I have always struggled with mathematics, but this book has helped me pass all my statistics exams with high grades!"

~ *Stella*

"At last, a clear statistics book! Transparent, consise diagrams at the front of each chapter display the outline of the story well, and the examples are fun to read and help you understand the subject matter. Where other books are often boring and complicated, Vince has found a way to make statistics accessible to everyone."

~ *Emma*

WHY WOULD YOU USE THIS BOOK?

DEAR STUDENT,

You probably came here in pursuit of the rumours: rumours about a handbook that does things differently. A teaching method that understands you, that takes you seriously, and that can truly improve your insight. Statistics is a subject you've never asked for; you study psychology, medicine or a social science, and are not an expert at mathematics. But since statistics is so important, also for you, you can't get around it. Will it ever become more exciting and tangible than a game of Xs and Ys?
As the responses show: yes. Anyone – and I'm not exaggerating – can see through statistics this way. Before you lies my approach of clear, crazy and complete explanations, translated to paper, accompanied by vibrant exercises that immediately try out your newfound skills. You will test rollercoasters, count bananas, measure moustaches and much more. It will make you smarter than you'd deemed possible until now.
In short: try it, and be convinced. Help build the revolution, along with the hundreds of students that came before you. Still think you're a peach in statistics? You're about to become a pirate!

Greetings,
Vincenzo

DEAR TEACHER,

You may open this handbook with a touch of scepticism. What is such a young whippersnapper doing with his own teaching method? Should he not have left that to seasoned professors, whose years of teaching and research have allowed them to grow into experts? Really, what has gotten into him, to tempt students into putting aside the recommended literature and change to his unsolicited alternative?
The claim that I possess less expertise than a scientist with a PhD in statistics needs no long debate: you are undoubtedly right. I have a lot to offer, but I am limited. Perhaps part of my strength arises from this fact. I can empathise with the student, who sometimes has to start from scratch, and often has no sense of mathematics or has not developed it yet. You can guess that I chose unorthodox examples and a light-hearted tone for this very reason. Chances are you will think that both are simply fineries, or even – in the most unfavourable case – that they distract the student from the things he or she should be learning. My view is different. Please allow me to provide an explanation.
First, I suspect that *structure* and *oversight* are two of the key pillars a student needs. Apart from my examples and style, I have done my best to offer these. Most people experience the organisation of new knowledge as exceptionally challenging, especially if the field does not suit them initially. By conducting some preliminary work, I prevent a lot of frustration. Each chapter begins with sharp tabulated summaries of the theory that is discussed, which show a clear continuity throughout the method: research designs, mathematical formulas, assumptions of statistical tests, and action plans for an analysis are mentioned point by point. The general *Action Plan for Test Analysis*, found in the Appendix module, offers further support in developing a bird's-eye view, and can be used at later

stages as well by students and young researchers to properly steer the statistics of their research. You may hold the opinion that I am leaving too little work in the students' own hands. In my own experience, however, they learn most effectively if they are presented with this structure at least once. In contexts that allow for interaction, such as a lecture or lesson, a middle road can be walked which is the optimum in my view: together with the teacher, the students will then fill in the summaries themselves. I also advise readers of this handbook to do the same thing in their own notes (see the following page).

Second, I am convinced that a teaching method for statistics should be both accessible and complete. Accessible in the sense that anyone should understand it; complete in the sense that it skips as few steps as possible. Roughly, statistical education knows two excesses. One excess is the 'just take this formula and fill it in' approach; you spare the student the details, but he or she still has no idea what it is actually about and remains deprived of true insights. At the other end we find formal mathematical definitions, where abstract language, matrices, proofs and subscripts build a heavily fortified castle without a visitor's entrance – and never forget: the student is a visitor in the statistical universe, an immigrant who still has to find his way. Neither of these extremes please me. This book is a roadmap which seeks to stop by a great many locations, using paths without pitfalls, and with tips for the connoisseur. When a formula presents itself, I will tell you how it was constructed; if a statistical test makes a certain assumption, I will explain why. Should a problem truly carry too far for a normal course in statistics, I will offer the solution in footnotes and bonus paragraphs for the reader who still wants to go to the last. Many students (especially from university) think it is hideous to just assume things; they prefer to go through fire and water and root their knowledge into solid ground – so afterward, they will retain that knowledge better, guaranteed.

Third, I will also ask a critical question whenever I come up with a new case: will this example help or will it distract? Is it a bizarre attempt to be interesting, or can a student really learn from it? Of course some will furrow their brows when asked to test washing machines, to taste cupcakes, and to observe orcs. Why does this method not cut to the chase? But soon these same students discover how the data come to life inside their heads. Relationships and effects take shape in their imagination, and at this point, I give those relationships and effects a statistical face. The student starts to see what this mathematical gibberish is supposed to say, what significance and confounding and interaction mean, and succeeds in the end to turn the case around and make the most difficult translation of all: from the result of the analysis to the conclusion of the study at hand. The irony? The student does not learn in spite of, but *thanks to* the cheerful examples, which, to him or her, are infinitely more tangible than a variable X and a variable Y. And when he or she later tries to remember how multiple regression went again, which memory might surface faster? 'Oh, that thing with the Xs!'… or rather: 'Oh, that thing with the pirates!'?

In short: *Pirates, Peaches and P-values* wants to build bridges. A bridge between exhaustiveness and accessibility, one between substance and mathematics, and one between comedy and knowledge. And let us not forget: a bridge between teachers. As I frankly admitted earlier, your expertise in statistics is larger than mine. Thus, this book has been designated for mutual cooperation from its first paragraph. I think that you and I can complement each other from our own specialty, to finally rob the subject of statistics from its persistent status of 'difficult' and 'boring'. Should you see conditions arise on which this teaching method would be a valuable addition to your curriculum, please do not hesitate and contact us at www.pppwaarden.nl. I am looking forward to our first conversation.

Warm regards,
Vince Penders

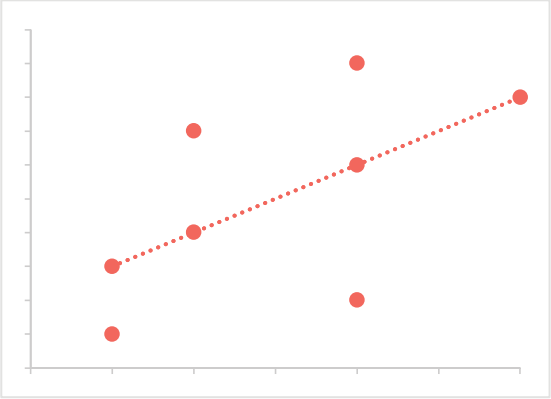
Chapter 3

QUANTITATIVE RELATIONSHIPS

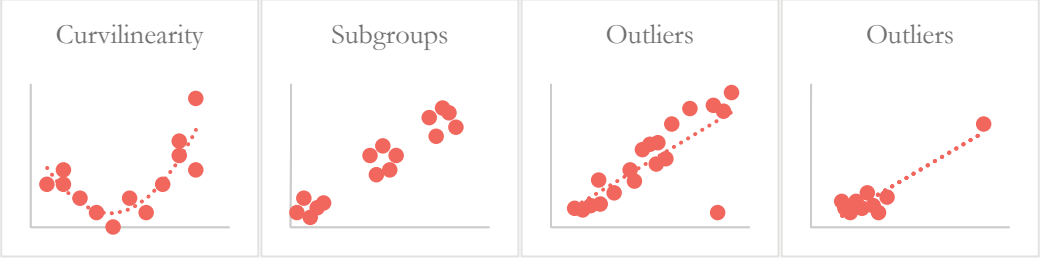
3.1 OVERVIEW

LEVEL	PREREQUISITES
Cupcake	Chapter 1: Mapping Data

SCATTERPLOT



Be wary of:



If the risks presented above don't occur, we may describe a linear trend. Make sure to differentiate two things: the strength of the relationship, and the appearance of the relationship. Often we cannot derive them from each other!

HOW STRONG IS THE LINEAR RELATIONSHIP?

To establish that we calculate the **correlation coefficient**, r_{XY}. This quantity is always between -1 (perfect negative relationship) and 1 (perfect positive relationship); 0 means no relationship.

$$r_{XY} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})/(N-1)}{s_X s_Y}$$

Squaring gives you R^2, the **proportion of explained variation**:

$$R^2 = \frac{explained\ variation}{total\ variation}$$

WHAT DOES THE LINEAR RELATIONSHIP LOOK LIKE?

To establish that we state the **regression equation**:

$$\hat{Y} = a + bX$$

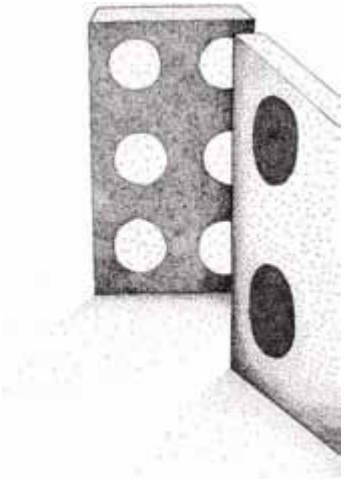
With

- b as the **slope** (1 to the right is b upward).

$$b = r_{XY} * \frac{s_Y}{s_X}$$

- a as the **intercept** (point where the line passes the Y axis).

$$a = \bar{Y} - b\bar{X}$$

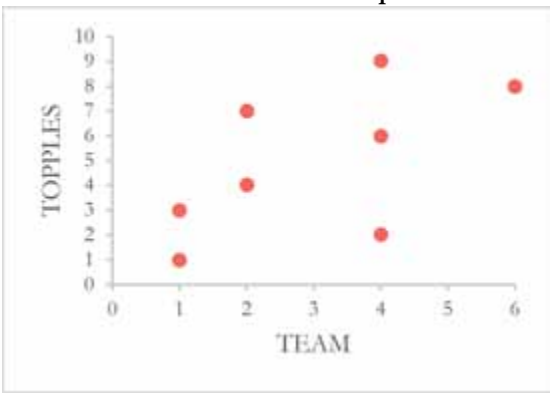


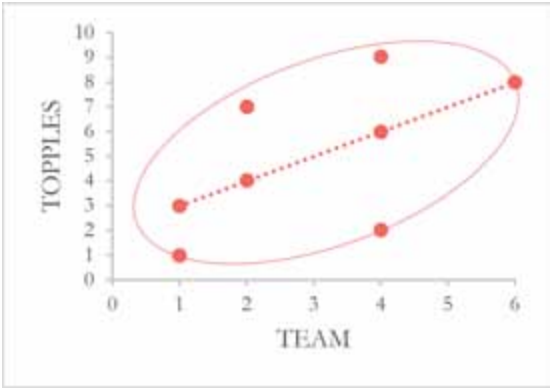
3.2 CORRELATION

Note: in case you skipped chapter 2, please read paragraph 2.2: Introduction to association real quick.

Domino shows: a curious form of entertainment. Some consider it a bit childish, while others train themselves to become absolute building professionals. In all likelihood, dear reader, you built your own show at least once when you were little. Perhaps you did it multiple times, inviting friends to join forces and combine your domino collections to make a mega-long spectacle. Nevertheless, having more people around may also increase the risk of accidents. How often have your dominoes not toppled before the chain was complete?

Is there truly a relationship between the number of domino builders and the number of times the track is toppled prematurely? Let's bring a visit to 8 random projects – that's not a super-representative sample, but it'll make our calculations easier. The size of the TEAM that builds the show, and the number of premature TOPPLES are two quantitative variables. The results can therefore be summarised in the form of a **scatterplot**.





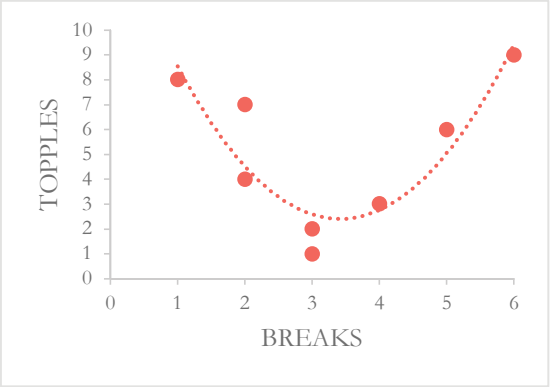
The X axis (horizontal) displays the size of the TEAM; the Y axis (vertical) the number of early TOPPLES. We see for instance that there were two projects made by one lonely builder (1); one of them was quite successful (one single early topple), but the other had a few unlucky incidents (3 topples). Now, how strong is the relationship between team size and performance? To find this out we draw a line that fits best through the points in the scatterplot. The more loyally the points follow this line, the stronger the relationship can be said to be.

Thus the line represents a so-called **linear trend**: this line slopes upward, so we see that bigger teams tend to have more accidents. The number of TOPPLES always increases equally under the influence of a larger group of builders, since the line always runs equally steeply (with linearity that is the case by definition). Try to draw an ellipse around the set of points: the more elongated the ellipse, the stronger the association. Should you get nearly a circle, the relationship is very weak.

Drawing a line of best fit is a good approach, unless one of these situations present themselves:

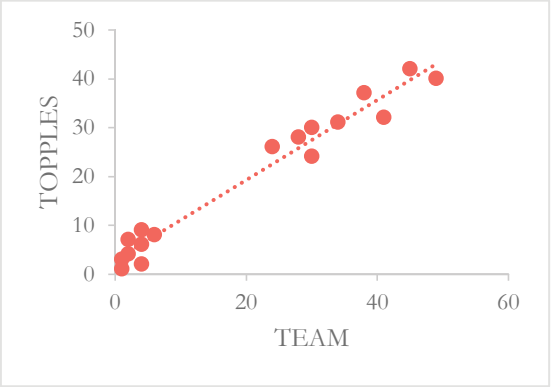
I. A CURVILINEAR TREND

Let's look at the number of breaks that each team takes during the building process, and relate it to the number of accidents. This relationship does not appear linear: it's probably quadratic in this case. Taking no breaks is bad for your concentration, and causes more accidents toward the end. However, taking too many also breaks your concentration, since the builders can't get into the flow of building. There's a sweet spot of taking just the right amount of breaks. In other words, it isn't very useful to draw a straight line through this point cloud. Don't!



II. SUBGROUPS

Suppose that we extend the sample by also visiting a few professional domino shows that dozens of builders contribute to? Then we obtain this. The relationship between TEAM and TOPPLES suddenly appears really strong. But something is wrong here… do you get it, dear reader?



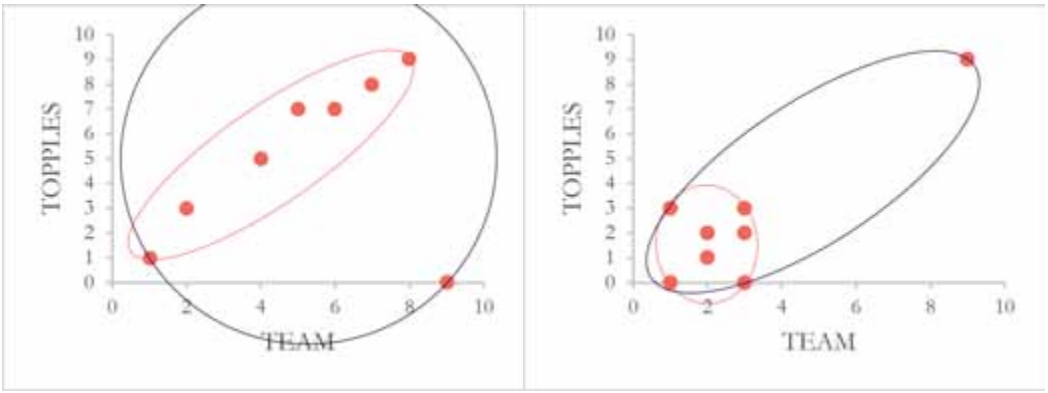
The bigger teams also work on much bigger domino projects. Rather than setting up a line of a few hundred dominoes, we're talking about hundreds of thousands – perhaps even a million. These big teams don't necessarily suffer more accidents because the builders are in each other's way, but rather because things can go wrong more often when you have more dominoes to set up. So we've got two subgroups – amateur and professional builders – and within these subgroups, the relationship between TEAM size and early TOPPLES isn't all that strong. If we ignore the presence of these subgroups and draw a single line through the scatterplot, it suggests that bigger teams will always perform worse – while this is not the case. This phenomenon is called **confounding**. Be careful about it! In our current study of amateur domino shows, I've made sure that every project built an equally long domino track.[19]

III. OUTLIERS

In bivariate distributions, an outlier can be nastily in your way as well: it will distort the image of the association between the two variables. A few different outliers are possible. Take the left example on the next page. The team at the bottom right is very large, yet makes 0 mistakes. In this regard it falls outside the general pattern; the larger teams tend to have more accidents. You can see the consequence of this outlier: the relationship suddenly looks weaker (the dark ellipse), since the points follow a straight line far less well than without the outlier (the light ellipse). The relationship between TEAM and TOPPLES is underestimated as a result.

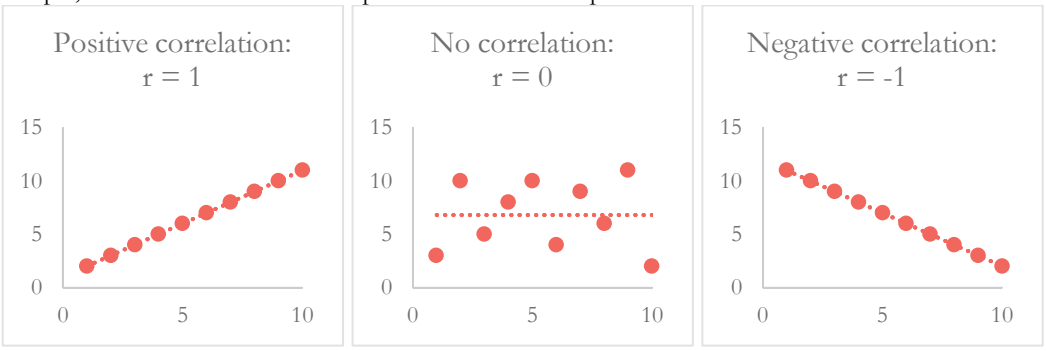
The opposite is possible too. Look at the right graph. We have one big team that caused an amazing mess, which is not very representative for the sample as a whole. This outlier, by contrast, causes the relationship between TEAM and TOPPLES to be overestimated.

[19] Often subgroups aren't distinguished as clearly as in the current scatterplot. We should therefore keep in mind that an external variable may always distort the image. If we suspsect such an external variable, we can mark the points in the scatterplot on the basis of that variable: the points of the amateurs would then become small triangles and the professionals rectangles, for example. Should the relationship look similar for the triangles and the rectangles, and should they be each other's extension, we won't really have subgroups. In that case it's fine to analyse the sample as a whole.



Now, in our very first scatterplot (2 pages earlier) everything's looking neat: as far as I can see there are no outliers, subgroups or a curving trend. The relationship between TEAM and TOPPLES looks reasonably strong; can we also indicate how strong exactly? Yes we can, by means of the **correlation coefficient**.[20] It is commonly denoted by the letter ***r***.

Before going to the mathematical formula in order to calculate the correlation coefficient for our sample, we'll first look at how it expresses the relationship between two variables:



So: when r_{XY} is about 0, the relationship between the two variables is very weak or even absent. The further r_{XY} deviates from 0, the stronger the relationship. The coefficient can be -1 at least; in that case we speak of a **perfect negative association** (think back to contingency tables for categorical variables, in chapter 2). When the correlation coefficient is 1 (maximum), there's a **perfect positive association**.

r_{XY}	Relationship	Predictions
-1	Perfect negative	Perfect: a bigger team of builders will always topple their dominoes prematurely less often, and always to the same extent. (That might sound counterintuitive for this example.) TEAM is thus the only factor that influences TOPPLES.
Between -1 and 0	Strong to weak negative	Good to lacking: a bigger team sometimes (but not always) causes less accidents, to different extents.
0	Absent	Worthless: some bigger teams have more premature topples whereas some have less. We're utterly unable to predict the number of TOPPLES from a team's size.
Between 0 and 1	Strong to weak positive	Lacking to good: a bigger team sometimes (but not always) causes more accidents, to different extents.
1	Perfect positive	Perfect: a bigger team of builders will always topple their dominoes prematurely more often, and always to the same extent. TEAM is thus the only factor that influences TOPPLES.

[20] Full name: Pearson's product-moment correlation coefficient. "Oh, that one!"

Okay? Now we're going to calculate the linear correlation coefficient.
The most common formula[21] is:

$$r_{XY} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})/(N-1)}{s_X s_Y}$$

Upon seeing this formula you might think: 'Blimey, calculating such a correlation coefficient could be an absurd load of work.' You thought that completely right. To show you what we're doing and why we choose to calculate the correlation in this way, I shall elaborate it once. Once, got it? After that we'll leave it to a computer program for the rest of our lives.
Here you can find for each domino project how many builders contributed and how many times someone was stupid enough to topple a domino too early (I've even ordered them for you):

TEAM (X)	TOPPLES (Y)
1	1
1	3
2	4
2	7
4	2
4	6
4	9
6	8

First of all we shall need the mean and standard deviation of X and Y (we'll use the formulas from chapter 1):

$$\bar{X} = \frac{\sum X_i}{N} = \frac{1+1+2+2+4+4+4+6}{8} = 3$$

$$s_X = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N-1}} = \sqrt{\frac{(1-3)^2 + (1-3)^2 + (2-3)^2 + (2-3)^2 + (4-3)^2 + (4-3)^2 + (4-3)^2 + (6-3)^2}{8-1}} = 1{,}77$$

$$\bar{Y} = \frac{\sum Y_i}{N} = \frac{1+3+4+7+2+6+9+8}{8} = 5$$

$$s_Y = \sqrt{\frac{\sum (Y_i - \bar{Y})^2}{N-1}} = \sqrt{\frac{(1-5)^2 + (3-5)^2 + (4-5)^2 + (7-5)^2 + (2-5)^2 + (6-5)^2 + (9-5)^2 + (8-5)^2}{8-1}} = 2{,}93$$

Good. Now does a relationship exist? Does a team that scores high on X (number of builders) scores high on Y as well (number of topples)? When X changes, how does Y change along? We're going to see how X and Y vary together, or **covary**. And how, in chapter 1, did we indicate the extent of varying scores again? We calculated how far scores deviated from the mean.

[21] You *could* also calculate the correlation coefficient using $r_{XY} = \frac{\sum z_X z_Y}{N-1}$, but if you ask me that would cost even more time and it isn't clearer than the formula that most statisticians handle. This variant may mostly show you that the correlation is a standardised measure, since it's based on z-scores (standard scores). I've uploaded a z-score-based calculation of r to the website version of chapter 3, at www.pppwaarden.nl.

TEAM: $X - \bar{X}$	TOPPLES: $Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
$1 - 3 = -2$	$1 - 5 = -4$	$-2 * -4 =$ **8**
$1 - 3 = -2$	$3 - 5 = -2$	$-2 * -2 =$ **4**
$2 - 3 = -1$	$4 - 5 = -1$	$-1 * -1 =$ **1**
$2 - 3 = -1$	$7 - 5 = 2$	$-1 * 2 =$ **-2**
$4 - 3 = 1$	$2 - 5 = -3$	$1 * -3 =$ **-3**
$4 - 3 = 1$	$6 - 5 = 1$	$1 * 1 =$ **1**
$4 - 3 = 1$	$9 - 5 = 4$	$1 * 4 =$ **4**
$6 - 3 = 3$	$8 - 5 = 3$	$3 * 3 =$ **9**

So of each time we take their deviation from the X mean *and* their deviation from the Y mean. We then multiply the two deviations by each other.
Next we add up the results. Behold the **covariation**:

$$\sum_i (X_i - \bar{X})(Y_i - \bar{Y}) =$$
$$8 + 4 + 1 - 2 - 3 + 1 + 4 + 9 =$$
$$22$$

Now we'd like to have the average covariation per team (usually per person), better known as the **covariance**. We therefore divide the covariation by $N - 1$:

$$\frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{N - 1} = \frac{22}{7} = 3{,}14$$

The covariance is already a nice measure of how TEAM (X) and TOPPLES (Y) vary together, a measure that's used in many statistical analyses.[22] However, it's hard to interpret it. How big is 3,14? A positive relationship, sure, but is it fairly strong or does it mean nothing? How the covariance comes out will strongly depend on the scale we've measured X and Y on.
That's why we shall finally standardise the covariance: we'll turn it into a number that always falls between -1 and 1. To this end we simply need to divide it by the two standard deviations:

$$\boldsymbol{r_{XY}} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})/(N-1)}{s_X s_Y} = \frac{3{,}14}{1{,}77 * 2{,}93} = \mathbf{0{,}61}$$

Which is a strong positive relationship. It would seem that bigger teams make more mistakes.[23]
Now what is the idea behind this calculation? We don't look for nothing at how the projects deviate from the X and Y means, let alone multiply these deviations for naught. Look at the scatterplot again. Now behold the four quadrants that arise when we indicate the means in it, on the next page.

- The three domino projects in the upper right quadrant were set up by a team sized above the average (X) and were also plagued by an above-average number of accidents (Y). Thus their deviation from the mean was positive on both X and Y. If we multiply those deviations – like

[22] A special case: suppose we calculate the covariance of X with itself? Then we obtain $\frac{\sum_i (X_i - \bar{X})(X_i - \bar{X})}{N-1} = \frac{\sum_i (X_i - \bar{X})^2}{N-1}$. Well I'll be… that's the **variance** of X. ☺

[23] Careful now: the presence of a correlation doesn't yet mean that a causal link truly exists. We can never be sure about that in a correlational research design. It's always possible that a variable in the background creates an imaginary relationship. This phenomenon is called confounding.
We can try to correct for confounders, in order to find the *true* relationship between X and Y. The statistical techniques for that require a fair deal of prescience though; all in good time. ☺ The first chapter that will seriously deal with confounding is chapter 12.