

Chapter 1

Introduction to Correlation

Suppose that you woke up one morning and discovered that you had been given the gift of being able to predict the future. Suddenly, you found yourself able to predict, with some degree of accuracy, things like:

- How long someone will live
- Whether the stock market will go up or down
- Whether or not someone will become a criminal
- Whether or not a surgery will prolong a cancer patient's life
- Whether or not a depressed person will commit suicide
- Whether or not a person will make a productive employee
- Whether or not a football team will make a first down on the next play
- Whether or not somebody's marriage will survive or end in divorce

How would that make you feel? Would you feel lucky? Would you head for Las Vegas? Would you use your gift *for good and not for evil*? Or would you say, “No thanks! Not interested.”

Well, in fact you are about to be given such a gift and I am going to give it to you! It is in the form of the Correlation Coefficient. The correlation coefficient is an almost magical statistic that allows you to do things like I have listed above and much more!

Developed by Karl Pearson over a hundred years ago, it is officially called the *Pearson Product Moment Correlation Coefficient*. However, it is usually referred to simply as “Pearson's ‘r’” or just the “Correlation Coefficient” is an extremely powerful and useful statistical tool that allows you to evaluate whether or not a relationship exists between two things. These “things” are referred to as “variables”.

Perhaps more importantly, not only does the correlation coefficient tell you whether or not a real relationship exists, it allows you to tell just how strong that relationship is. Finally—and here is where predicting the future comes in—if a real relationship exists between two variables, then the correlation coefficient allows you to make predictions about one variable based solely on knowledge of the other variable.

In other words, if a relationship exists between a child watching violent television programs and his or her aggressive behavior in school, then the correlation coefficient will allow you to predict how violent a child would be in school based solely on how much violent television the child watched at home!

Pretty cool, huh? Think about the kind of power that gives you!

Insurance companies use the correlation coefficient to predict how likely a person is to get into an accident. This enables them to set insurance rates so that high risk drivers pay more. Mutual fund managers use correlation coefficients to decide what a stock's price is likely to do and use that information to decide whether to buy or sell stocks. Healthcare providers use correlation coefficients to decide whether or not someone will likely die if they do not get a certain surgery and so they make decisions accordingly. Psychologists use correlation coefficients to predict whether or not a sexually violent predator (a person repeatedly convicted of sex crimes) is likely to

commit another crime if they are released from prison. Graduate schools use correlation coefficients to make predictions about whether or not a prospective student is likely to succeed in graduate school. And the list goes on and on!

Decisions based on information provided by correlation coefficients can help businesses save millions—even billions—of dollars. The information can enhance public safety and make the world a better place to live in. Think about it! We owe Karl Pearson a lot for the gift he has given us!

How Does Correlation Work?

The best way to give you an idea of how correlation works is to use a graphical tool you learned about in the chapter on descriptive statistics—the scatterplot.

Suppose that you are interested in answering the question: “Is a person’s starting salary (the amount they are paid when they start their first job) related to how many years of education they have successfully completed?” In other words, you are interested in finding out if more school will translate into more money when you get a job. If so, then you can use this information to decide whether or not it is worth getting a college degree (assuming that you are only in it for money and you are not interested in getting a degree just for the joy of learning!)

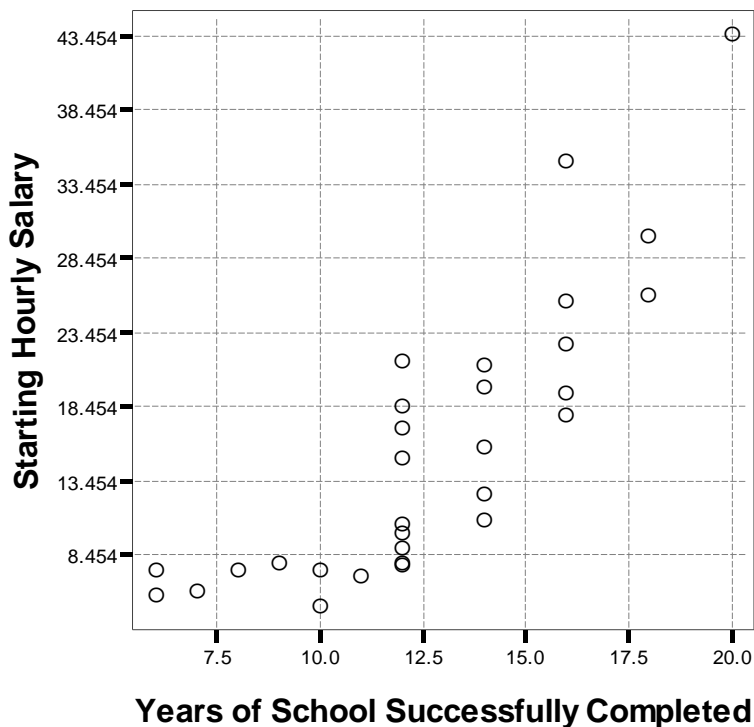
Here is one approach you could take to answer this question:

- **Step 1** – Make sure you have clearly stated your question. This is important because if you ask a confusing question then any answers you get will be just as confusing and un-interpretable.
- **Step 2** – Identify the two variables you will need in order to answer your question. *For standard correlation studies, you will always use two variables.* For example, to answer the question “Is a person’s starting salary related to how many years of education they have successfully completed?” you will need to collect two pieces of information:
 1. How many years of education did the person successfully complete?
 2. What was the hourly salary for their first job when they finished school?
- **Step 3** – Select a random sample of people from whom you will try to get the information you need to answer your question. Let me emphasize that *you need a fair and unbiased sample of people.* If you have a biased sample (say, you only talk to historians) you will get a biased answer to your question. Trust me, unless you are trying to deceive yourself or others, you do not want a biased sample.
- **Step 4** – Ask each person your two questions (the ones I identified in Step 2, above).
- **Step 5** – Graph the answers that people give you in a scatterplot.

Now let's actually follow the steps listed above to see how we could discover whether or not a relationship exists between the amount of education a person gets and their starting salary.

- **Step 1** – Clearly state our question: Is a person's starting salary (the amount they are paid when they start their first job) related to how many years of education they have successfully completed?
- **Step 2** – Clearly identify the two variables we need to answer this question:
 1. How many years of education did you successfully complete?
 2. What hourly salary did you receive in the first job you got after completing your education?
- **Step 3** – Suppose we select 30 people—at random—to whom we will ask our two questions.
- **Step 4** – We ask each of the 30 people our two questions.
- **Step 5** – We graph their responses in a scatterplot (see Figure 11-1, below)

FIGURE 11-1
Scatterplot of number of years of education by starting hourly salary of first job obtained after completing school (Based on fictitious, or made-up, data)



As you look at Figure 11-1, does anything jump out at you? Can you say anything about the relationship between a person's starting salary and the number of years of education that he or she successfully completes? Does it seem like they are related at all? Who would you say tended to have higher starting salaries, people with less education or more education?

A careful review of Figure 11-1 clearly indicates that people who had less education (less than a high school diploma) tended to also have lower starting salaries. On the other hand, as the amount of education a person got went up, so did their starting salary. People with the most education also tended to have the highest starting salaries. This is an example of a "positive relationship".

In answer to our question about whether or not there is a relationship between the amount of education one receives and their starting salary, the evidence we have presented would suggest that "Yes, there is a positive relationship between amount of education and a person's starting salary." As result, you might conclude that you should pursue as much education as you can because it is likely to mean that you will receive a higher salary!

That is the power of using data to establish whether relationships exist between two things. On the one hand, you could just make a "gut decision" about whether there is a relationship—but many times such "intuitive" guesses tend to be wrong. However using data to establish relationships allows you to not only see if the relationship really exists, but it allows you to get an idea about how strong the relationship is! Looking at Figure 11-1, it is pretty obvious (based on this fictitious data) that there is a fairly *strong* relationship between years of education and starting salary.

Okay, now you know what a positive relationship is; it is a relationship where as values on one variable (e.g., years of education) go up the values on the other variable (e.g., starting salary) also goes up. There is another kind of relationship besides a positive relationship—a negative relationship.

Suppose you were interested in using data to find out if there is really a relationship between a person's exercise patterns and their risk of heart attacks. What you are really asking is "Is there a relationship between the average number of minutes of exercise a person gets each week and that person's risk of coronary hearth disease?"

What do you think would be the answer to this question? Just based on what you hear on television and read in newspapers and magazines, you probably would conclude that there is a relationship between how much a person exercises and his or her risk of heart disease, right?

Well, you could use the very same procedure we looked at above to answer the question. Suppose we asked 30 people the following questions:

1. How many minutes, on average, do you exercise each day?
2. What is your cholesterol level?

Before I go any farther, let me point out that to make this study work, you would probably need to test each person's cholesterol because most people would not be able to accurately tell you what their cholesterol level is.

Once you have gotten these two pieces of information (minutes of exercise and cholesterol level) you are ready to put your real question to the test. You could plot each

Summary of Types of Relationships

When you are trying to see whether or not relationships exist between two variables, you can find one of three possible things: 1) A positive relationship exists between them, 2) a negative relationship exists between them or 3) no relationship at all exists between them.

Positive Relationships

In a positive relationship, as one variable increases (goes up) the other variable also increases (goes up). Similarly, as one variable decreases (goes down) the other variable goes down. *Both variables tend to move in the same direction.*

If there is a positive relationship between people's amount of education and his or her starting hourly salary that means:

- As amount of education goes up, starting hourly salary tends to go up.
- As amount of education goes down, starting hourly salary tends to go down.

Negative Relationships

In a negative relationship, as one variable increases (goes up) the other variable decreases (goes down). Similarly, as one variable decrease (goes down) the other variable increases (goes up). In other words they go in opposite directions from each other. *The variables tend to move in the opposite direction.*

If there is a negative relationship between average number of minutes of exercise each day and a person's cholesterol level that means:

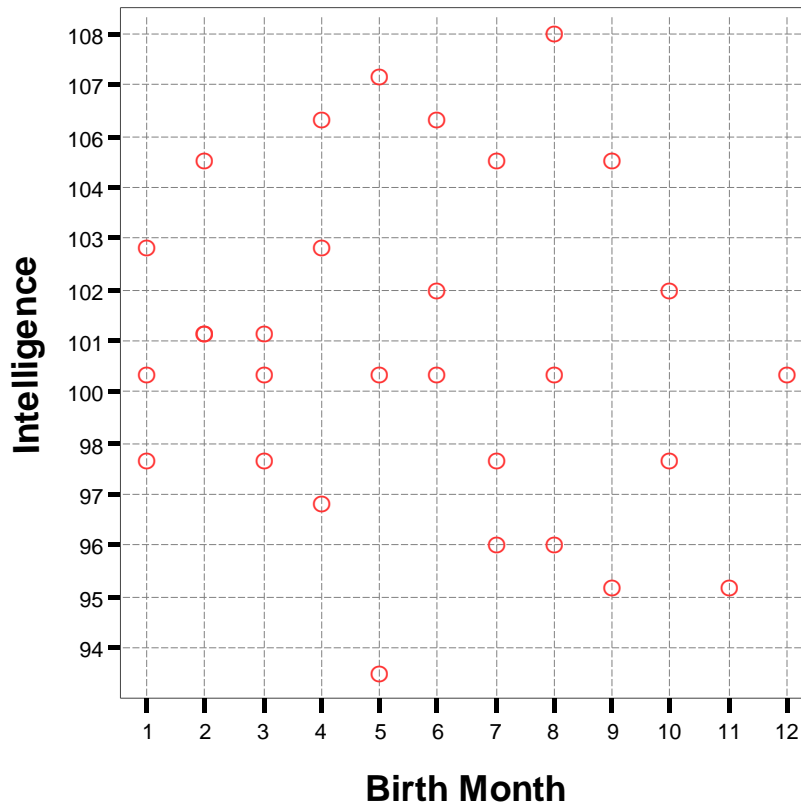
- As minutes of daily exercise goes up, cholesterol level tends to go down.
- As minutes of daily exercise goes down, cholesterol level tends to go up.

No Relationship

When there is no relationship between two variables that means that you really can't say anything about one variable based on values on the other variable. In other words, as values on one variable increases (goes up) there is no predictable change in the other variable. Sometimes the other variable goes up. Sometimes the other variable goes down. Sometimes, it just stays the same.

Is there a relationship between the month in which a person is born and how intelligent they are? If you were to ask 30 people their birth month and then test them to find out what their Intelligence Quotient (IQ) is, what you should find is the month in which a person is born has nothing to do with their IQ. There should be no relationship between these two variables. In cases where there is no relationship between the two variables, a scatterplot should look like a bunch of random dots—kind of like a target that has been shot with a shotgun. See Figure 11-3 for an example of no relationship between two variables.

FIGURE 11-3
Scatterplot illustrating an example of no relationship between two variables (Based on fictitious data)



Notice that you really can't say anything about a person's intelligence based on his or her birth month. There is no predictable trend or relationship that exists (at least in this fictitious data set. If you really want to find out, go out and collect some data and see what you find!). This is an example of there being no relationship between two variables.

How Strong Is The Relationship?

Now that you know how to see if there is a relationship between two variables, the next question to ask is, "how strong is that relationship?" Why is this an important question? Let's take a look at an example.

Suppose Spring Break is fast approaching and you and some of your friends are thinking about heading south to Cancun, Mexico, for some rest, relaxation and fun. Being the wise consumer/traveler that you are, you ask yourself "Is there a relationship between how expensive a hotel is and how much people enjoy their stay?" Your thinking is that if there is a relationship that indicates that the more people pay, the more fun they tend to have, then that is what you will do. On the other hand, if people who stay at cheaper hotels have as much as or more fun, why pay the extra bucks?

To answer your question, you ask a bunch of people who have been to Cancun over spring break about how much their hotel cost and also ask them to rate how much fun they had on a scale of 0 to 10 (0 = no fun at all and 10 means so much fun that your head exploded). Next, you plot their responses on a scatterplot. Figures 11-4 and 11-5 give a couple of examples of what you might find from your study.

FIGURE 11-4
Example of a weak positive relationship between hotel price and amount of fun reported by people visiting Cancun over spring break (Based on fictitious data)

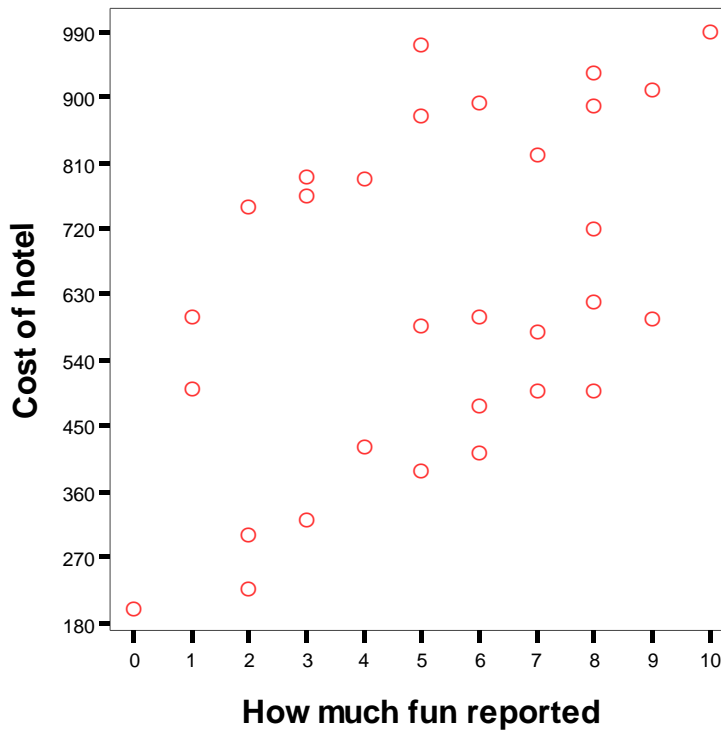


Figure 11-4 provides an example of a “weak” positive relationship. Notice that you can tell that there is a positive relationship between hotel cost and how much fun people who have stayed there reported (as one variable goes up so does the other). Notice however that there is a lot of variability. What I mean by variability is that there are a lot of people who only paid around \$500 who had almost as much fun or more fun than people who paid much more!

As a result, you know that if you spend more you may have more fun than if you go cheap, but it is not at all a sure thing. Some people who paid nearly \$1000 had less fun than people who paid only half as much. So what should you do? Take a look at Figure 11-5, below.

FIGURE 11-5

Example of a strong positive relationship between hotel price and amount of fun reported by people visiting Cancun over spring break (Based on fictitious data)

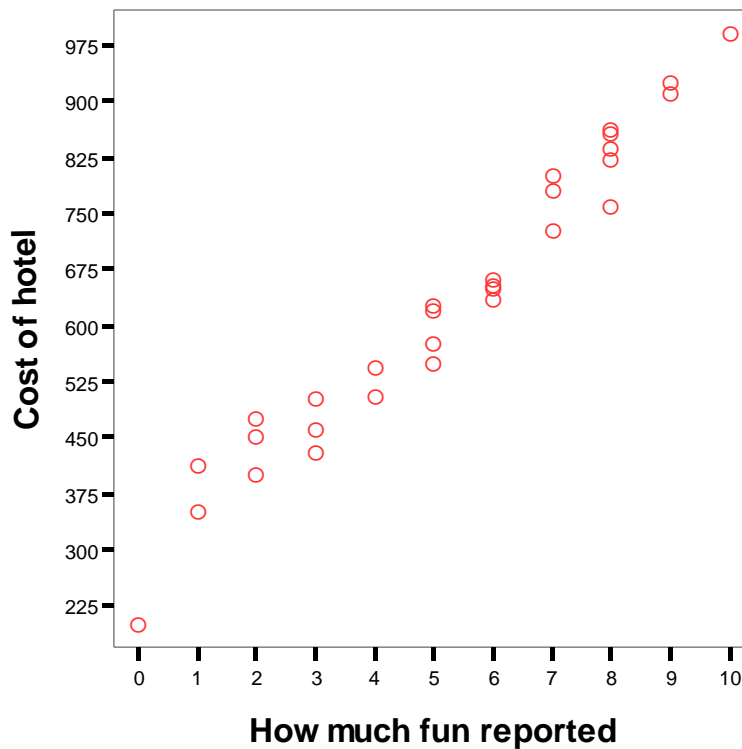


Figure 11-5 is an example of a “strong” positive relationship. In a strong relationship (whether it is positive or negative) the dots tend to group to form more of a straight line. Notice how in Figure 11-5 there is a very clear trend that tells you “as people pay more for their hotel they also tend to report having more fun during their stay”.

Why Is It Important To Know The Strength Of A Relationship?

Now that you have some idea about what it means to say that a relationship is strong or weak, why is it so important? Because it allows you to make predictions!

Let’s get back to our Spring Break example. Suppose that you have \$2,000 to spend and you want to be as certain as possible to get the most for your money. Aside from the cost of your airline tickets, your next largest expense is likely to be your hotel. Should you pay top dollar for your hotel or go on the cheap side?

If you had conducted our little study and got the results found in Figure 11-4, above, you could say “Yes, I need to avoid being too cheap because it seems like people report having more fun in more expensive hotels. However, it is hard to tell if paying more than \$500 would ensure that I would have a better time than going for a \$500 hotel”. In other words, because the relationship is weak, it is difficult to predict with any

accuracy how much fun you would have although it seems pretty certain that cost seems to be related to fun.

On the other hand, take a look at Figure 11-5 which illustrates a strong positive relationship. Notice that with a strong relationship it is very clear that higher price definitely seems related to more fun. Given the information in Figure 11-5, one could clearly say that “People who paid \$900 reported having much more fun than people who paid \$500.” Therefore, you could predict that if you paid top dollar for your lodging, you would have more fun. Note, this does not mean you will have more fun only that it appears likely that you will. You cannot say that paying more will cause you to have more fun, only that since it appears cost and fun are related it seems reasonable to predict that you *may* have more fun.

The main point of this is that a strong relationship allows you to make much more accurate predictions than a weak relationship. This ability to make accurate predictions is critical in a great many professional settings. Psychologists, medical professionals, business executives, stock brokers, military leaders, law enforcement agents are all interested in being able to make predictions. The concept of correlation provides a tool that helps people make predictions and to do so with some amount of accuracy.

The Problem

There are a couple of problems with what we have done so far. Let me review them:

1. Creating scatterplots can be time consuming to construct.
2. While scatterplots can be useful to get an idea about whether or not a relationship exists between two variables, it can be very difficult to differentiate between weak relationships (which are still important) and instances when there are no relationships.

What Is Needed

What we need is a way to take all the information we have collected and somehow reduce it to a single summary number that tells us:

1. Is there a relationship between the two variables?
2. Is the relationship, if it exists, positive or negative?
3. How strong is the relationship?

In addition, it would be great if the same number could be used to actually make predictions about one variable based on information about the other. For example, wouldn't it be great if this number would allow us to decide how much fun we want to have in Cancun and have it tell us how much we would need to spend on our hotel?

As a matter of fact, there is such a number. It is called the Correlation Coefficient and it will be discussed in the next chapter.

Terms to Learn

You should be able to define the following terms based on what you have learned in this chapter.

Correlation
Correlation Coefficient
Negative Relationship
Positive Relationship
