4. Correlação
Como a Netflix sabe quais filmes eu gosto?

A NETFLIX INSISTE QUE vou gostar do filme *Bhutto*, um documentário que oferece uma "visão em profundidade e às vezes incendiária da vida e da trágica morte da ex-primeira-ministra paquistanesa Benazir Bhutto". Provavelmente vou gostar do filme. (Eu o adicionei ao "Minha lista".) As recomendações da Netflix às quais assisti no passado foram incríveis. E quando eles recomendam um filme a que já assisti, costuma ser um de que eu realmente gostei.

Como a Netflix faz isso? Será que existe alguma gigantesca equipe de estagiários na sede da corporação que usou uma combinação do Google e entrevistas com a minha família e amigos para determinar que eu poderia gostar de um documentário sobre uma ex-primeira-ministra paquistanesa? É claro que não. A Netflix simplesmente domina algumas estatísticas sofisticadas. *A Netflix não me conhece*. Mas conhece os filmes dos quais gostei no passado (porque eu os avaliei). Usando essa informação, junto com as avaliações de outros clientes e um computador potente, a Netflix pode fazer previsões incrivelmente acuradas sobre as minhas preferências.

Adiante voltarei ao algoritmo específico da Netflix para fazer essas escolhas; por enquanto, o ponto importante é que tudo está baseado em correlação. A Netflix recomenda filmes que são semelhantes a outros filmes de que gostei; e também recomenda filmes que foram muito bem avaliados por outros clientes cujas avaliações são similares às minhas. *Bhutto* foi recomendado por causa das cinco estrelas com que avaliei dois outros documentários, *Enron: os mais espertos da sala* e *Sob a névoa da guerra*.

A correlação mede o grau em que dois fenômenos estão relacionados entre si. Por exemplo, existe uma correlação entre temperaturas de verão e venda de sorvete. Quando uma sobe, a outra sobe também. Duas variáveis têm correlação positiva se uma variação numa delas é associada a uma variação da outra no mesmo sentido, tal como a relação entre altura e peso. Pessoas mais altas pesam mais (em média); pessoas mais baixas pesam menos. Uma correlação é negativa se uma variação positiva numa das variáveis está associada a uma variação negativa na outra, tal como a relação entre exercício e peso.

O aspecto traiçoeiro nesses tipos de associações é que nem toda observação se encaixa no padrão. Às vezes pessoas mais baixas pesam mais

que pessoas mais altas. Às vezes pessoas que não se exercitam são mais magras que pessoas que se exercitam o tempo todo. Ainda assim, existe uma relação significativa entre altura e peso, bem como entre exercício e peso.

Se fôssemos colocar num gráfico de dispersão as alturas e pesos de uma amostra aleatória de americanos adultos, seria de esperar ver algo do seguinte tipo:

Se fôssemos criar um gráfico de dispersão entre exercício (medido em minutos por semana de exercício intensivo) e peso, seria de esperar uma correlação negativa, com os que se exercitam mais tendendo a pesar menos. Mas um padrão consistindo em pontos dispersos numa página é uma ferramenta um tanto tosca. (Se a Netflix tentasse me fazer recomendações de filmes com um gráfico das avaliações de milhares de filmes por milhões de clientes, os resultados soterrariam a sede debaixo de gráficos de dispersão.) Em vez disso, o poder da correlação como ferramenta estatística é que podemos encapsular uma associação entre duas variáveis numa única estatística descritiva: o coeficiente de correlação.

O coeficiente de correlação tem duas características fabulosamente atraentes. A primeira, por razões matemáticas que foram relegadas ao apêndice, trata-se de um número único que varia de –1 a 1. Uma correlação de 1, muitas vezes descrita como correlação perfeita, significa que qualquer alteração em uma das variáveis está associada com uma alteração equivalente na outra variável no mesmo sentido.

Uma correlação de –1, ou correlação negativa perfeita, significa que toda alteração em uma variável está associada a uma alteração equivalente na outra variável em sentido oposto.

Quanto mais perto de 1 ou –1 estiver a correlação, mais forte a associação. Uma correlação de 0 (ou próxima a 0) significa que as variáveis não têm associação significativa entre si, como a relação entre o número do sapato e os resultados em exames escolares.

A segunda característica atraente do coeficiente de correlação é que ele não está ligado a nenhuma unidade. Podemos calcular a correlação entre altura e peso – mesmo que a altura seja medida em centímetros e o peso em quilogramas. Podemos até calcular a correlação entre a quantidade de televisores que alunos do ensino médio têm em suas casas e seus resultados em exames escolares, e eu lhes asseguro que será positiva. (Falarei mais sobre essa relação daqui a pouco.) O coeficiente de correlação faz uma coisa aparentemente milagrosa: reduz uma complexa bagunça de dados medidos em unidades diferentes (como o nosso gráfico de dispersão de altura e peso) numa única e elegante estatística descritiva.

Como?

Mantendo o hábito, pus a fórmula mais comum para se calcular o coeficiente de correlação no apêndice ao final do capítulo. Essa não é uma estatística que você vai calcular à mão. (Depois de você inserir os dados, um programa básico como o Microsoft Excel calcula a correlação entre as duas variáveis.) Ainda assim, intuitivamente não é tão difícil. A fórmula para calcular o coeficiente de correlação faz o seguinte:

1. Calcula a média e o desvio padrão para ambas as variáveis. Se nos ativermos ao exemplo de altura e peso, saberíamos então a altura média das pessoas na amostra, o peso médio das pessoas na amostra e o desvio padrão tanto para a altura como para o peso.
2. Converte todos os dados de modo que cada observação seja representada por sua distância da média (seu desvio padrão). Acompanhe meu raciocínio; não é tão complicado. Suponha que a altura média na amostra seja de 170 centímetros (com um desvio padrão de dez centímetros); e que o peso médio seja de 75 quilos (com um desvio padrão de cinco quilos). Agora suponha que você tenha 182 centímetros de altura e pese 71 quilos. Podemos dizer também que sua altura é 1,2 desvios padrões acima da média em altura $[^{(180-165)}/_{10}]$, e seu peso 0,8 desvios padrões abaixo da média, ou –0,8 para fins de fórmula $[^{(71-75)}/_{5}]$. *Sim, é incomum alguém estar acima da média em altura e abaixo da média em peso, mas já que você pagou um bom dinheiro pelo livro, achei que deveria pelo menos fazer você alto e magro.* Note que a sua altura e peso, anteriormente em centímetros e quilos, foram reduzidos a 1,2 e −0,8. É isso que faz as unidades sumirem.
3. Aqui eu libero minhas mãos e deixo o computador fazer o serviço. A fórmula calcula então a relação entre altura e peso de todos os indivíduos da amostra, medidos pelas unidades-padrão. Quando os indivíduos da amostra são altos, digamos 1,5 ou dois desvios padrões acima da média, o que tende a acontecer com seus pesos *medidos em desvios padrões da média para o peso*? E quando os indivíduos estão perto da média em

termos de altura, quais são seus pesos, medidos em unidades de desvio padrão?

Se a distância de uma variável em relação à média tende a ser amplamente consistente com a distância da outra variável em relação à média (por exemplo, pessoas distantes da média em termos de altura, em qualquer um dos dois sentidos, também tendem a estar distantes da média no mesmo sentido em termos de peso), então seria de esperar uma forte correlação positiva.

Se a distância em relação à média de uma das variáveis tende a corresponder a uma distância similar em relação à média da segunda variável *no sentido oposto* (por exemplo, pessoas bem acima da média em termos de exercício tendem a estar bem abaixo da média em termos de peso), então devemos esperar uma forte correlação negativa.

Se duas variáveis não tendem a se desviar da média segundo nenhum padrão significativo (por exemplo, número do sapato e exercício), então devemos esperar uma correlação pequena ou nula.

Você sofreu intensamente nesta seção; voltaremos já, já para o aluguel de filmes. Antes de retornarmos à Netflix, porém, vamos refletir sobre outro aspecto da vida em que a correlação é relevante: o Teste de Raciocínio SAT. Conhecido antigamente nos Estados Unidos como Teste de Aptidão Acadêmica (SAT, na sigla em inglês), trata-se de um exame padronizado composto de três partes – matemática, leitura crítica e redação – cujo objetivo é mensurar a capacidade acadêmica e predizer o desempenho universitário. É claro que há motivo razoável para se perguntar (especialmente aqueles que não gostam de testes padronizados): não é para isso que serve o ensino médio? Por que um exame de quatro horas é tão importante quando os funcionários encarregados da admissão universitária têm acesso a *quatro anos* de notas tiradas no ensino médio?

A resposta para essas perguntas encontra-se camuflada nos Capítulos 1 e 2. Notas do ensino médio são uma estatística descritiva imperfeita. Um aluno que tira notas medíocres enquanto enfrenta uma programação difícil com aulas de matemática e ciências pode ter maior capacidade e potencial acadêmico do que um aluno no mesmo colégio com notas melhores em matérias menos desafiadoras. Obviamente há discrepâncias potenciais ainda maiores de uma escola para outra. Segundo o College Board, que produz e administra o SAT, o teste foi criado para “democratizar o acesso ao ensino superior para todos os estudantes”. Muito justo. O SAT fornece uma medida padronizada de capacidade que pode ser facilmente comparada entre todos os alunos que se candidatam ao ensino superior. *Mas será que é uma boa medida de capacidade?* Se queremos um critério que possa ser comparado facilmente entre estudantes, poderíamos também mandar os alunos de último ano

correrem um tiro de cem metros, que é mais barato e mais fácil do que administrar o SAT. O problema, obviamente, é que a performance num tiro de cem metros não tem correlação com desempenho acadêmico. Obter os dados é fácil; só que eles simplesmente não nos revelam nada de significativo.

Então, qual é a qualidade da informação obtida pelo SAT? Infelizmente para futuras gerações de alunos do ensino médio, o SAT faz um trabalho razoavelmente bom em predizer as notas de primeiro ano de faculdade. O College Board publica as correlações relevantes. Numa escala de 0 (absolutamente nenhuma correlação) a 1 (correlação perfeita), a correlação entre a média de notas no ensino médio e a média de notas no primeiro ano da faculdade é 0,56. (Para dar alguma perspectiva a esse número, a correlação entre altura e peso para homens adultos nos Estados Unidos é aproximadamente 0,4.) A correlação entre o placar composto do SAT (leitura crítica, matemática e redação) e a média das notas do primeiro ano universitário também é 0,56.[1] Esse resultado parece argumentar a favor da eliminação do SAT, pois o teste parece não dar resultados melhores na previsão do desempenho universitário do que as notas do ensino médio. Na verdade, o melhor previsor de todos é uma combinação do SAT e da média do ensino médio, que tem uma correlação de 0,64 com as notas do primeiro ano universitário. Sinto muito por ter que dizer isso.

UM PONTO CRUCIAL nesta discussão geral é que correlação não implica causalidade; uma associação positiva ou negativa entre duas variáveis não significa necessariamente que uma variação numa delas esteja causando a variação na outra. Por exemplo, anteriormente aludi a uma provável correlação positiva entre os resultados do SAT de um aluno e a quantidade de televisores que sua família possui. Isso não significa que pais superansiosos possam aumentar o placar dos testes de seus filhos comprando cinco aparelhos de televisão adicionais para a casa. E provavelmente tampouco significa que assistir muito à televisão seja bom para o desempenho acadêmico.

A explicação mais lógica para tal correlação seria que pais com elevado nível de educação podem se dar ao luxo de ter uma porção de aparelhos de televisão e tendem a ter filhos cujos resultados nos testes estão acima da média. Tanto televisores como resultados de testes são provavelmente causados por uma terceira variável, que é a educação dos pais. Não posso provar a correlação entre esses aparelhos na casa e resultados do SAT. (O College Board não fornece esses dados.) No entanto, posso provar que alunos de famílias mais ricas têm em média escores no SAT mais altos do que alunos de famílias menos ricas. Segundo o College Board, alunos com renda familiar acima de US$200 mil têm um placar médio no SAT de matemática de 586, em comparação com um placar médio de 460 para alunos com renda familiar

de US$20 mil ou menos.[2] Ao mesmo tempo, também é provável que famílias com renda superior a US$200 mil tenham mais televisores em suas (múltiplas) casas do que famílias com renda de US$20 mil ou menos.

COMECEI A ESCREVER este capítulo muitos dias atrás. Desde então, tive a oportunidade de assistir ao documentário *Bhutto*, um filme excepcional sobre uma família excepcional. As sequências originais, que começam com a partilha da Índia e do Paquistão em 1947 e vão até o assassinato de Benazir Bhutto em 2007, são extraordinárias. A voz de Bhutto é muito bem intercalada ao longo do filme na forma de discursos e entrevistas. Em todo caso, dei cinco estrelas ao filme, que é praticamente o que a Netflix previu.

No nível mais básico, a Netflix está explorando o conceito de correlação. Primeiro, eu avalio um conjunto de filmes. A Netflix compara minhas avaliações com as de outros clientes para identificar aqueles cujas avaliações estejam altamente correlacionadas com as minhas. Esses clientes tendem a gostar dos filmes que eu gosto. Uma vez estabelecido isso, a Netflix pode recomendar filmes que receberam alta avaliação de clientes de mentalidade semelhante à minha, mas que eu ainda não assisti.

Esse é o "quadro geral". A metodologia real é muito mais complexa. Na verdade, a Netflix lançou em 2006 um concurso no qual membros do público foram convidados a projetar um mecanismo que melhorasse as recomendações existentes da empresa em pelo menos 10% (o que significa que o sistema ficaria 10% mais acurado em predizer como um cliente avaliaria um filme depois de assistir). O vencedor ganharia US$1 milhão.

Todo indivíduo ou equipe que se inscreveu para o concurso recebeu "dados de treinamento" consistindo em mais de 100 milhões de avaliações de 18 mil filmes por 480 mil clientes Netflix. Um conjunto separado de 2,8 milhões de avaliações foi "retido", o que significa que a Netflix sabia como os clientes tinham avaliado esses filmes, mas os participantes do concurso não. Os competidores foram julgados com base na acurácia com que seus algoritmos previam as avaliações reais dos clientes para esses filmes retidos. Durante três anos, milhares de equipes de mais de 180 países submeteram propostas. Havia duas exigências para participar: primeira, o vencedor deveria licenciar o algoritmo para a Netflix; segunda, o vencedor tinha de "descrever ao mundo como você fez e por que funciona".[3]

Em 2009, a Netflix anunciou o vencedor: uma equipe de sete pessoas composta de estatísticos e cientistas da computação dos Estados Unidos, Áustria, Canadá e Israel. Sinto muito, não posso descrever o sistema ganhador, nem mesmo no apêndice. O artigo explicando o sistema tem 92 páginas.[a] Eu fico impressionado com a qualidade das recomendações da Netflix. Ainda assim, o sistema é apenas uma supervariação rebuscada do que as pessoas vêm fazendo desde a aurora do cinema: achar alguém com gosto

semelhante e pedir uma recomendação. Você tende a gostar do que eu gosto, e não gostar do que eu não gosto, então, o que acha do novo filme do George Clooney?

Essa é a essência da correlação.

APÊNDICE AO CAPÍTULO 4

Para calcular o coeficiente de correlação entre dois conjuntos de números, você executaria os seguintes passos, cada um deles ilustrado com o uso dos dados de alturas e pesos para quinze estudantes hipotéticos na tabela a seguir.

A	B	C	D	E	F
Aluno	Altura (cm)	Peso (kg)	Altura em unidade-padrão	Peso em unidade-padrão	(Peso em unidades-padrão) × (Altura em unidades-padrão)*
Nick	185	88	1,34	1,05	1,41
Elana	165	60	−0,49	−0,74	0,36
Dinah	170	70	−0,03	−0,09	0,01
Rebecca	172	67	0,15	−0,29	−0,04
Ben	183	80	1,16	0,54	0,63
Charu	175	58	0,43	−0,87	−0,37
Sahar	150	45	−1,86	−1,69	3,14
Maggie	158	58	−1,13	−0,87	0,98
Faisal	168	77	−0,21	0,35	−0,07
Ted	175	83	0,43	0,73	0,31
Narciso	175	81	0,43	0,61	0,26
Katrina	175	54	0,43	−1,12	−0,48
CJ	187	103	1,52	2,01	3,05
Sophia	155	53	−1,41	−1,18	1,67
Will	185	96	1,34	1,56	2,09
Média	170,34	71,53			Total = 12,95
Desvio padrão	10,91	15,66		Coeficiente de correlação = Total/n = $^{12,95}/_{15}$ = 0,86	

1. Converta a altura de cada estudante para unidades-padrão: (altura – média)/desvio padrão.
2. Converta o peso de cada estudante para unidades-padrão: (peso – média)/ desvio padrão.
3. Calcule o produto para cada estudante de (peso em unidades-padrão) × (altura em unidades-padrão). Você deve perceber que esse número será o maior em valor absoluto quando a altura e o peso estiverem ambos

relativamente longe da média.

4. O coeficiente de correlação é a soma dos produtos calculados acima dividida pelo número de observações (15 no caso). A correlação entre altura e peso nesse grupo de estudantes é 0,86. Considerando que o coeficiente de correlação pode variar de –1 a 1, temos aqui um grau relativamente alto de correlação positiva, como seria de esperar com altura e peso.

A fórmula para calcular o coeficiente de correlação requer um pequeno desvio relativo à notação. O símbolo ∑, conhecido como somatória, é um caractere conveniente em estatística. Representa a soma da grandeza que vem logo em seguida. Por exemplo, se há um conjunto de observações x_1, x_2, x_3 e x_4, então $\sum (x_i)$ nos diz que devemos somar as quatro observações: $x_1 + x_2 + x_3 + x_4$. Assim, $\sum (x_i) = x_1 + x_2 + x_3 + x_4$. Nossa fórmula para a média de um conjunto de i observações poderia ser representada da seguinte maneira: média = $\sum (x_i)/n$.

Podemos tornar a fórmula ainda mais adaptável escrevendo $\sum_{i=1}^{n} (x_i)$, que soma a quantidade $x_1 + x_2 + x_3 + \ldots + x_n$, ou, em outras palavras, todos os termos começando por x_1 (porque i = 1) até x_n (porque i = n).

Nossa fórmula para a média de um conjunto de observações pode ser representada da seguinte maneira:

$$\text{média} = \sum_{i=1}^{n} \frac{(x_i)}{n}$$

Dada essa notação geral, a fórmula para calcular o coeficiente de correlação, r, para duas variáveis x e y é a seguinte:

$$r = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sigma_x} \frac{(y_i - \bar{y})}{\sigma_y}$$

em que

n = número de observações;
$\bar{x}$ = é a média da variável x;
$\bar{y}$ = é a média da variável y;
σ_x = é o desvio padrão da variável x;
σ_y = é o desvio padrão da variável y.

Qualquer programa com ferramentas estatísticas também pode calcular o coeficiente de correlação entre duas variáveis. No exemplo de altura e peso de

estudantes, o Microsoft Excel dá a mesma correlação entre altura e peso para os quinze estudantes que o cálculo feito à mão da tabela acima: 0,86.

---

[a] Você pode lê-lo em: http://www.netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf.

[b] Embora o termo “unidade-padrão” tenha uma série de significados diferentes, estamos aqui acompanhando o autor e usando o termo “unidades-padrão” referindo-nos a “unidades de desvio padrão”. (N.T.)

DATA SCIENCE

Introduction to Correlation

Ruslana Dalinina | 01.31.17

Prerequisites

Experience with the specific topic: Novice

Professional experience: No industry experience

To follow this article, the reader should be familiar with Python syntax and have some understanding of basic statistical concepts (e.g. average, standard deviation).

Introduction: What Is Correlation and Why Is It Useful?

Correlation is one of the most widely used – and widely misunderstood – statistical concepts. In this overview, we provide the definitions and intuition behind several types of correlation and illustrate how to calculate correlation using the Python **pandas** library.

The term "correlation" refers to a mutual relationship or association between quantities. In almost any business, it is useful to express one quantity in terms of its relationship with others. For example, sales might increase when the marketing department spends more on TV advertisements, or a customer's average purchase amount on an e-commerce website might depend on a number of factors related to that customer. Often, correlation is the first step to understanding these relationships and subsequently building better business and statistical models.

So, why is correlation a useful metric?

techniques

More formally, correlation is a statistical measure that describes the association between random variables. There are several methods for calculating the correlation coefficient, each measuring different types of strength of association. Below we summarize three of the most widely used methods.

Types of Correlation

Before we go into the details of how correlation is calculated, it is important to introduce the concept of *covariance*. Covariance is a statistical measure of association between two variables X and Y. First, each variable is centered by subtracting its mean. These centered scores are multiplied together to measure whether the increase in one variable associates with the increase in another. Finally, expected value (E) of the product of these centered scores is calculated as a summary of association. Intuitively, the product of centered scores can be thought of as the area of a rectangle with each point's distance from the mean describing a side of the rectangle:

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

If both variables tend to move in the same direction, we expect the "average" rectangle connecting each point (X_i, Y_i) to the means (X_bar, Y_bar) to have a large and positive diagonal vector, corresponding to a larger positive product in the equation above. If both variables tend to move in opposite directions, we expect the average rectangle to have a diagonal vector that is large and negative, corresponding to a larger negative product in the equation above. If the variables are unrelated, then the vectors should, on average, cancel out – and the total diagonal vector should have a magnitude near 0, corresponding to a product near 0 in the equation above.

If you are wondering what "expected value" is, it is another way of saying the average, or mean μ, of a random variable. It is also referred to as "expectation." In other words, we can write the following equation to express the same quantity in a different way:

$$E(Y) = \bar{Y} = \mu_Y$$

The problem with covariance is that it keeps the scale of the variables X and Y, and therefore can take on any value. This makes interpretation difficult and comparing covariances to each other impossible. For example, $Cov(X, Y) = 5.2$ and $Cov(Z, Q) = 3.1$ tell us that these pairs are positively associated, but it is difficult to tell whether the relationship between X and Y is stronger than Z and Q without looking at the means and distributions of these variables. This is where correlation becomes useful – by standardizing covariance by some measure of variability in the data, it produces a quantity that has intuitive interpretations and consistent scale.

Pearson Correlation Coefficient

years ago is still the most widely used today.

In this section, we will introduce several popular formulations and intuitive interpretations for Pearson correlation (referred to as ρ).

The original formula for correlation, developed by Pearson himself, uses raw data and the means of two variables, X and Y:

$$\rho_{X,\,Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

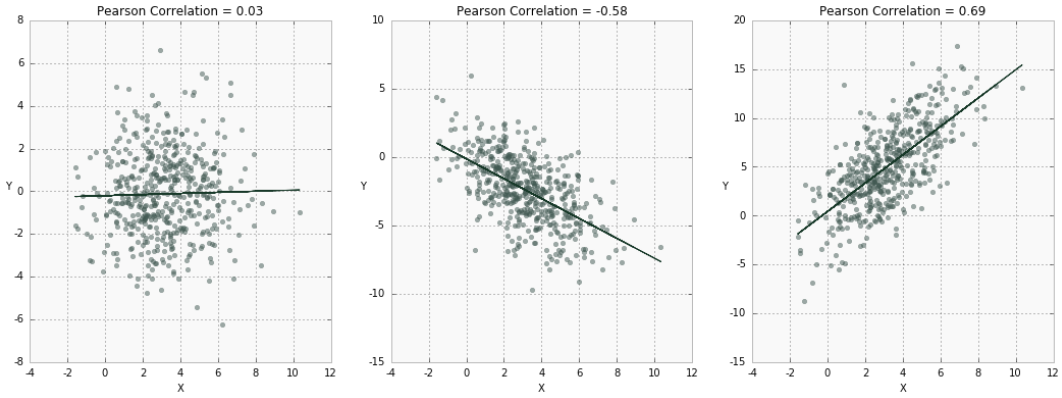
In this formulation, raw observations are centered by subtracting their means and re-scaled by a measure of standard deviations.

A different way to express the same quantity is in terms of expected values, means μ_X, μ_Y and standard deviations σ_X, σ_Y:

$$\rho_{X,\,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Notice that the numerator of this fraction is identical to the above definition of covariance, since mean and expectation can be used interchangeably. Dividing the covariance between two variables by the product of standard deviations ensures that correlation will always fall between -1 and 1. This makes interpreting the correlation coefficient much easier.

The figure below shows three examples of Pearson correlation. The closer ρ is to 1, the more an increase in one variable associates with an increase in the other. On the other hand, the closer ρ is to -1, the increase in one variable would result in decrease in the other. Note that if X and Y are independent, then ρ is close to 0, but not vice versa! In other words, Pearson correlation can be small even if there is a strong relationship between two variables. We will see shortly how this can be the case.



it minimizes the distance of all points to itself. Because of this property, the slope of the regression line of Y and X is mathematically equivalent to correlation between X and Y, standardized by the ratio of their standard deviations:

$$\rho = b\frac{s_X}{s_y}$$

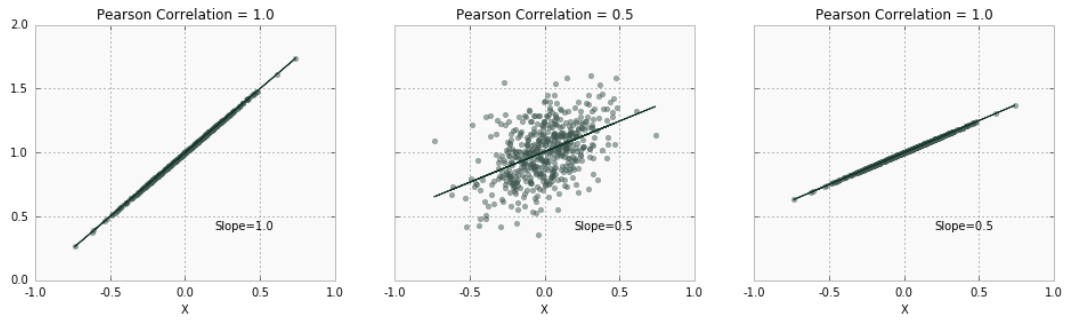
where b is the slope of the regression line of Y from X.

In other words, correlation reflects the association and amount of variability between the two variables.

This relationship with the slope of the line has two important implications:

1. It makes it more clear why Pearson correlation describes linear relationships
2. It also shows why correlation is important and so widely used in predictive modeling

However, note that in the above equation for ρ, correlation **does not equal slope** – rather, it is standardized by a measure of data variability. For example, it is possible to have a very small magnitude of slope but large correlations between variables. In the figure below, the line describing this relationship is relatively flat, but correlation is 1 since variability s_y is very small:



Note that, so far, we have not made any assumptions about the distribution of X and Y. The only restriction is that Pearson ρ assumes a linear relationship between the two variables. Pearson correlation relies on means and standard deviations, which means it is only defined for distributions where those statistics are finite, making the coefficient sensitive to outliers.

Another way to interpret Pearson correlation is to use the coefficient of determination, also knows as R^2. While ρ is unitless, its square is interpreted at the proportion of variance of Y explained by X. In the above example, ρ = -0.65 implies that $(-0.65^2)*100$ = 42% of variation in Y can be explained by X.

Spearman's Correlation

Spearman's rank correlation coefficient can be defined as a special case of Pearson ρ applied to ranked (sorted) variables. Unlike Pearson, Spearman's correlation is not restricted to linear relationships. Instead, it measures **monotonic association** (only strictly increasing or decreasing, but not mixed) between two variables and relies on the rank order of values. In other words, rather than comparing means and variances, Spearman's coefficient looks at the relative order of values for each variable. This makes it appropriate to use with both continuous and discrete data.

The formula for Spearman's coefficient looks very similar to that of Pearson, with the distinction of being computed on ranks instead of raw scores:

$$\rho_{rank_X,\ rank_Y} = \frac{cov(rank_X,\ rank_Y)}{\sigma_{rank_X}\sigma_{rank_Y}}$$

If all ranks are unique (i.e. there are no ties in ranks), you can also use a simplified version:

$$\rho_s = 1 - \frac{6\sum d_i^2}{N(N^2 - 1)}$$

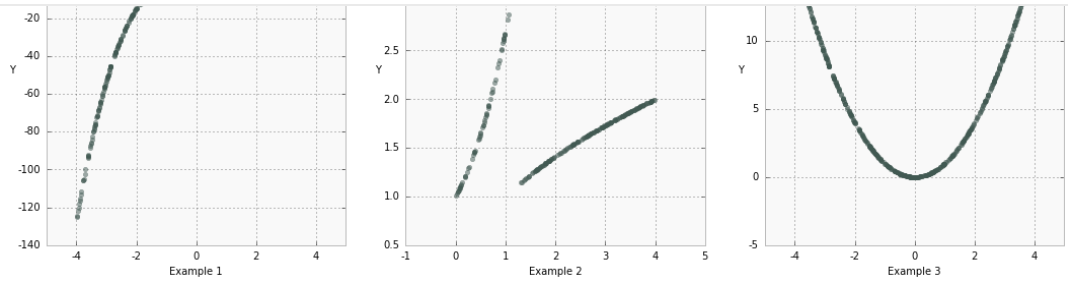
where $d_i = rank(X_i)$ - $rank(Y_i)$ is the difference between the two ranks of each observation and N is the number of observations.

The difference between Spearman and Pearson correlations is best illustrated by example. In the below figure, there are three scenarios with both correlation coefficients shown. In the first example, there is a clear monotonic (always increasing) and non-linear relationship. Since ranks of values perfectly align in this case, the Spearman's coefficient is 1. Pearson correlation is weaker in this case, but it is still showing a very strong association due to the partial linearity of the relationship.

The data in Example 2 shows clear groups in X and a strong, although non-monotonic, association for both groups with Y. In this case, Pearson correlation is almost 0 since the data is very non-linear. Spearman rank correlation shows weak association, since the data is non-monotonic.

Finally, Example 3 shows a nearly perfect quadratic relationship centered around 0. However, both correlation coefficients are almost 0 due to the non-monotonic, non-linear, and symmetric nature of the data.

These hypothetical examples illustrate that correlation is by no means an exhaustive summary of relationships within the data. Weak or no correlation does not imply lack of association, as seen in Example 3, and even a strong correlation coefficient might not fully capture the nature of the relationship. It is always a good idea to use visualization techniques and multiple statistical data summaries to get a better pictures of how your variables relate to each other.



Kendall's Tau

The third correlation coefficient we will discuss is also based on variable ranks. However, unlike Spearman's coefficient, Kendalls' τ does not take into account the difference between ranks – only directional agreement. Therefore, this coefficient is more appropriate for discrete data.

Formally, Kendall's τ coefficient is defined as:

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{N(N-1)/2}$$

As an example, consider a simple dataset consisting of five observations. In practice, such a small number of data points would not be sufficient nor reliable to draw any conclusions. But here, we consider it for the sake of the simplicity of calculation:

	X	Y
a	1	7
b	2	5
c	3	1
d	4	6
e	5	9

Concordant pairs $(x_1, y_1), (x_2, y_2)$ are pairs of values in which ranks coincide: $x_1 < x_2$ and $y_1 < y_2$ or $x_1 > x_2$ and $y_1 > y_2$. In our mini example, (4,6) and (5,9) in rows d and e is a concordant pair. A discordant pair would be one that does not satisfy this condition, such as (1, 7) and (2, 5). To calculate the numerator of τ, we compare all possible pairs in the dataset and count number of concordant pairs; 6 in this case:

- (1,7) and (5,9)
- (2,5) and (4,6)
- (2,5) and (5,9)
- (3,1) and (4,6)
- (3,1) and (5,9)
- (4,6) and (5,9)

and discordant pairs:

The denominator of Kendall's τ is just the number of possible combinations of pairs, which ensures that τ varies between 1 and -1. With five data points, there are $5 * 4/2 = 10$ possible combinations, making $\tau = (6\text{-}4) / 10 = 0.2$ in this example. Kendall's correlation is particularly useful for discrete data, where the relative position of data points is more important that difference between them.

```
# fake kendall
k = pd.DataFrame()
k['X'] = np.arange(5)+1
k['Y'] = [7, 5,1, 6, 9]
print k.corr(method='kendall')
```

```
     X    Y
X  1.0  0.2
Y  0.2  1.0
```

Calculating Correlation in Pandas

```
# pandas provide a convenient method for highlighting
# mpg_data.drop(['m
```

Below, we show how to calculate correlation for an example problem using a Python library. We will be using a dataset on vehicle fuel efficiency from University of California, Irvine. Let's say it is of interest to see what vehicle characteristics can help explain fuel consumption (mpg) of a vehicle. We begin by reading the dataset from the UCI online data repository and examining first few rows. Dataset documentation states that a special character is used for missing values (**?**), which can be used as one of the parameters to pandas **`read_csv()`** function:

```
import pandas as pd
path = 'http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/

mpg_data = pd.read_csv(path, delim_whitespace=True, header=None,
            names = ['mpg', 'cylinders', 'displacement','horsepower',
            'weight', 'acceleration', 'model_year', 'origin', 'name'],
            na_values='?')
```

is appropriate for your use case. If that is not the case, there are many existing methods for filling in and handling missing values, such as simple mean imputation.

```
mpg_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 398 entries, 0 to 397
Data columns (total 9 columns):
mpg             398 non-null float64
cylinders       398 non-null int64
displacement    398 non-null float64
horsepower      392 non-null float64
weight          398 non-null float64
acceleration    398 non-null float64
model_year      398 non-null int64
origin          398 non-null int64
name            398 non-null object
dtypes: float64(5), int64(3), object(1)
memory usage: 28.1+ KB
```

pandas provides a convenient one-line method **`corr()`** for calculating correlation between data frame columns. In our fuel efficiency example, we can check whether heavier vehicles tend to have lower **`mpg`** by passing the method to specific columns:

```
mpg_data['mpg'].corr(mpg_data['weight'])
```

```
-0.8317409332443354
```

As expected, there seems to be a strong negative correlation between vehicle **`weight`** and **`mpg`**. But what about **`horsepower`** or **`displacement`**? Conveniently, pandas can quickly calculate correlation between all columns in a dataframe. The user can also specify the correlation method: Spearman, Pearson, or Kendall. If no method is specified, Pearson is used by default. Here, we drop model year and origin variables and calculate Pearson correlation between all remaining columns of the data frame:

In [ ]:

```
# pairwise correlation
mpg_data.drop(['model_year', 'origin'], axis=1).corr(method='spearman')
```

cylinders	-0.821864	1.000000	0.911876	0.816188	0.873314	-0.47
displacement	-0.855692	0.911876	1.000000	0.876171	0.945986	-0.49
horsepower	-0.853616	0.816188	0.876171	1.000000	0.878819	-0.65
weight	-0.874947	0.873314	0.945986	0.878819	1.000000	-0.40
acceleration	0.438677	-0.474189	-0.496512	-0.658142	-0.404550	1.00(

pandas also supports highlighting methods for tables, so it is easier to see high and low correlations. It is important to understand possible correlations in your data, especially when building a regression model. Strongly correlated predictors, phenomenon referred to as multicollinearity, will cause coefficient estimates to be less reliable. Below is an example of calculating Pearson correlation on our data and using a color gradient to format the resulting table:

```
odel_year', 'origin'], axis=1).corr(method='pearson').style.format("{:.2}'
```

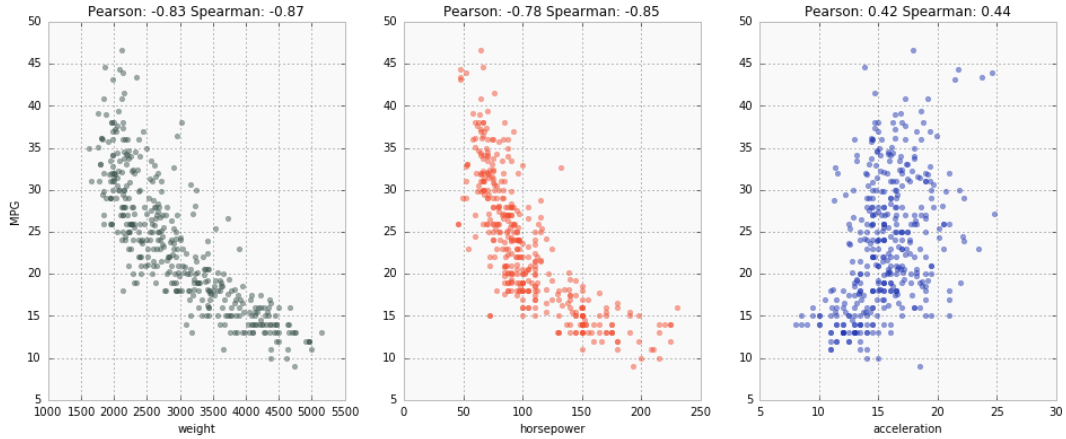
	mpg	cylinders	displacement	horsepower	weight	acceleration
mpg	1.0	-0.78	-0.8	-0.78	-0.83	0.42
cylinders	-0.78	1.0	0.95	0.84	0.9	-0.51
displacement	-0.8	0.95	1.0	0.9	0.93	-0.54
horsepower	-0.78	0.84	0.9	1.0	0.86	-0.69
weight	-0.83	0.9	0.93	0.86	1.0	-0.42
acceleration	0.42	-0.51	-0.54	-0.69	-0.42	1.0

Finally, to visually inspect the relationship between **`mpg`**, **`weight`**, **`horsepower`**, and **`acceleration`**, we can plot these values and calculate Pearson and Spearman coefficients. The dataset at hand consists of less than 400 points, which can be easily displayed on a scatter plot. If you are dealing with much larger datasets, consider taking a sample of your data first to speed up the process and produce more readable plots.

In this case, Spearman's coefficient is higher than Pearson for **`horsepower`** and **`weight`**, since relationship is non-linear. For **`acceleration`**, both coefficients are close since the relationship is not as clearly defined:

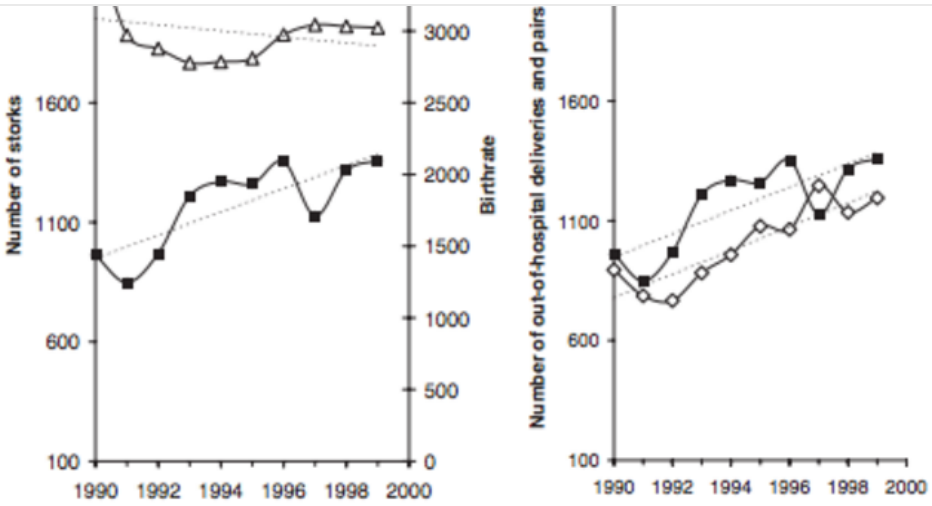
```
# plot correlated values
plt.rcParams['figure.figsize'] = [16, 6]

```

```
colors=[ #415952 ,  #735134 ,  #243AB5 ,  #243AB5 ]
j=0

for i in ax:
    if j==0:
        i.set_ylabel('MPG')
    i.scatter(mpg_data[cols[j]], mpg_data['mpg'],  alpha=0.5, color=colors
    i.set_xlabel(cols[j])
    i.set_title('Pearson: %s'%mpg_data.corr().loc[cols[j]]['mpg'].round(2)
    j+=1

plt.show()
```



Correlation and Causation

The relationships between variables in our fuel efficiency example were very intuitive and explainable through vehicle mechanics. However, things are not always this straightforward. It is a well known fact that correlation does not imply causation, and therefore, any strong correlation should be thought of critically. For example, German researchers used the concept of correlation in this humorous paper to support a theory that babies are delivered by storks. This figure shows correlation between the number of storks and baby deliveries:



The chart on the left shows an increasing trend in the number of storks (black line) and a decreasing trend in the number of clinical deliveries. On the other hand, the chart on the right shows that a number of out-of-hospital deliveries (white square marks) follow the increasing pattern in the number of storks. Looking at the correlation between these series, the authors suggest that the increase in out-of-hospital deliveries paired with the increase in number of storks and simultaneous decrease in hospital deliveries suggest that more and more babies in Germany are being delivered by storks.

Of course, this is a silly example. Nonetheless, it demonstrates an important point: Spurious statistical associations can be found in a multitude of quantities, simply due to chance.

Often, a relationship may appear to be causal through high correlation due to some unobserved variables. For example, the number of grocery stores in a city can be strongly correlated with the number of ice cream creameries. However, there is an obvious hidden variable here – the population size of the city: