3.3-Dados bivariados.

(Correlação linear, regressão linear, tabelas de contingência.)

Introdução.

Aqui pretendemos analisar a relação entre duas variáveis.

Por exemplo, a relação entre o peso e a altura de uma pessoa.

Neste tópico, recolhemos pares de dados, que designamos por **dados bivariados.**

Diagrama de dispersão.

Aqui usamos diagramas de dispersão ou gráficos de correlação.

Exemplo:

Na tabela seguinte temos os dados referentes ao "peso"(x) e ao "número de sapato que calça"(y), para os alunos de uma turma:

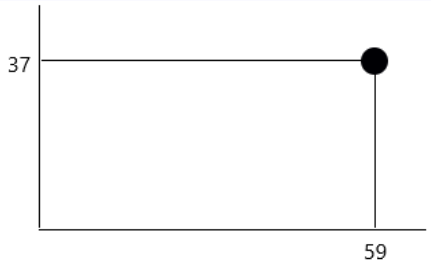
x	59	50	49	50	56	54	52	68	70	47	75	49	51	53	48	73	65	65
y	37	37	38	37	39	37	38	42	41	37	41	36	37	37	36	42	43	42

Observemos a primeira coluna de números.

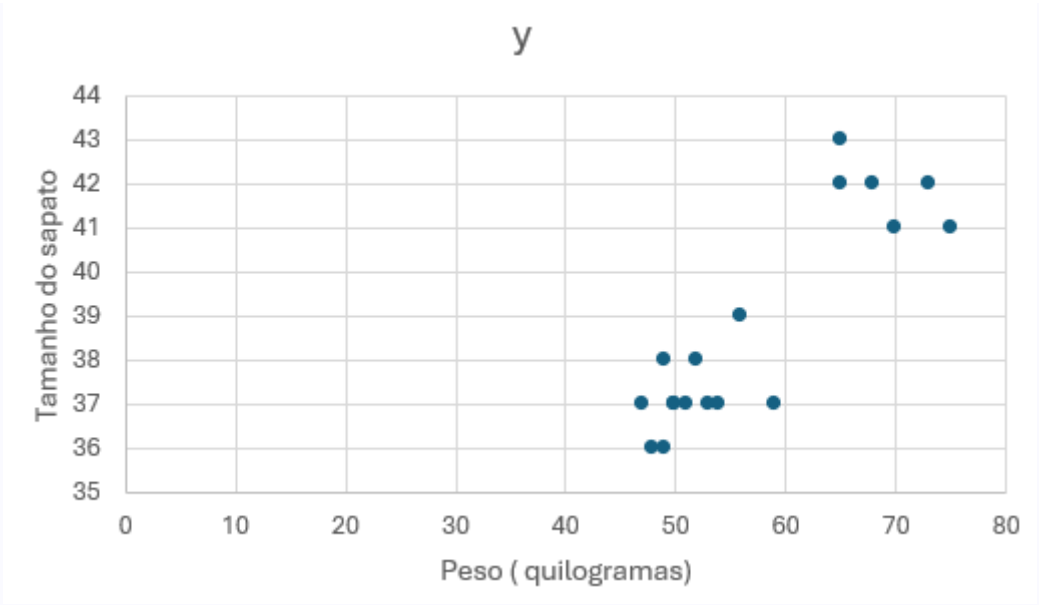
59
37

Interpretamos do seguinte modo: Um aluno pesa 59 quilogramas e calça um sapato com o número 37. Do mesmo modo para os restantes.

Par representar no gráfico, indicamos um ponto (59; 37), isto é, no eixo do xx ou eixo horizontal colocamos o 59 e no eixo dos yy ou eixo vertical colocamos o 37.



Fazendo para todos os valores, ficamos com um gráfico do tipo:



A este tipo de gráfico chamamos **gráfico de correlação** ou **diagrama de dispersão**.

O conjunto dos pontos num gráfico de correlação designa-se por **nuvem de pontos**.

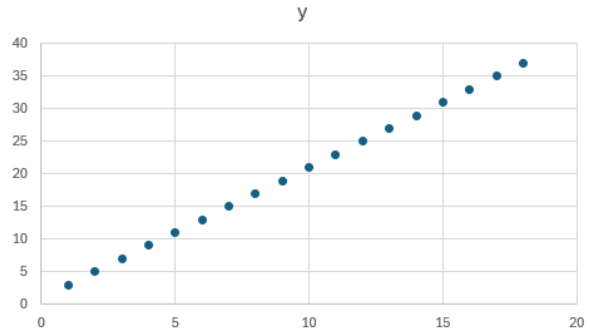
Nota: No exemplo acima, existe correlação entre as variáveis. Se um modo geral, podemos constatar que a tendência é:

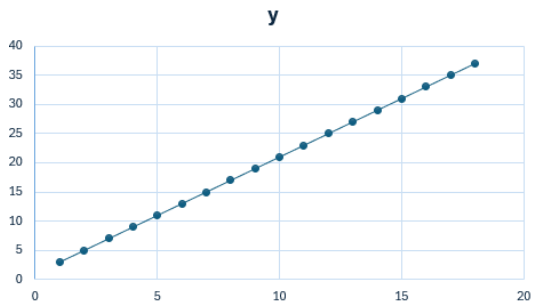
"quanto maior o peso, maior é o tamanho do sapato que calça".

Diagrama de dispersão ou **gráfico de correlação** é um gráfico de pontos em que as coordenadas de cada ponto são os valores das duas variáveis em estudo.

A **correlação** diz-se **linear** se a nuvem de pontos se distribuir ao longo de uma linha reta, a reta de regressão.

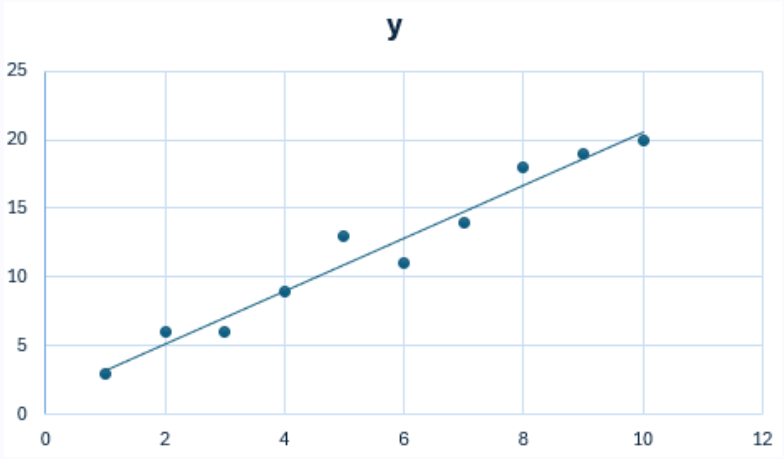
Exemplo: **Correlação linear**





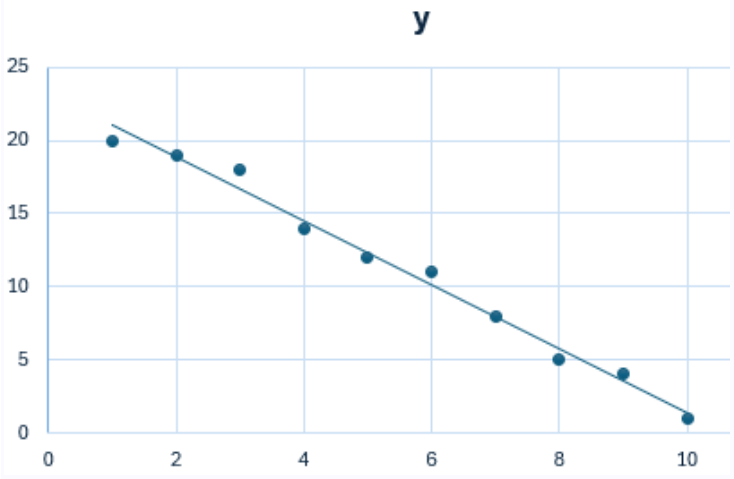
Correlação **linear positiva**- à medida que os valores de uma variável aumentam, os valores correspondentes da outra variável também aumentam.

Exemplo: **Correlação linear positiva.**



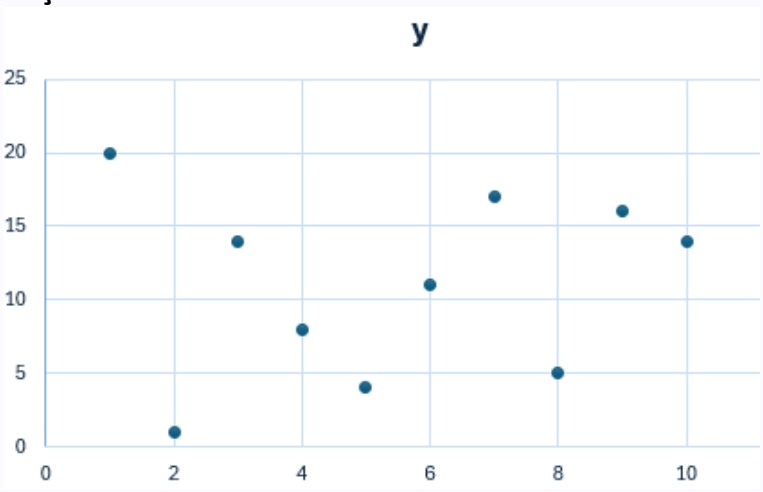
Correlação **linear negativa**- à medida que os valores de uma variável aumentam, os valores correspondentes da outra variável diminuem.

Exemplo: **Correlação linear negativa**

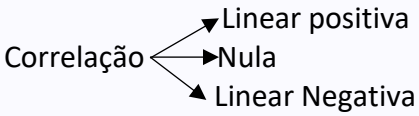


Correlação linear nula: não há qualquer padrão de relação linear. mas pode haver algum outro padrão

Exemplo: **Correlação linear nula**



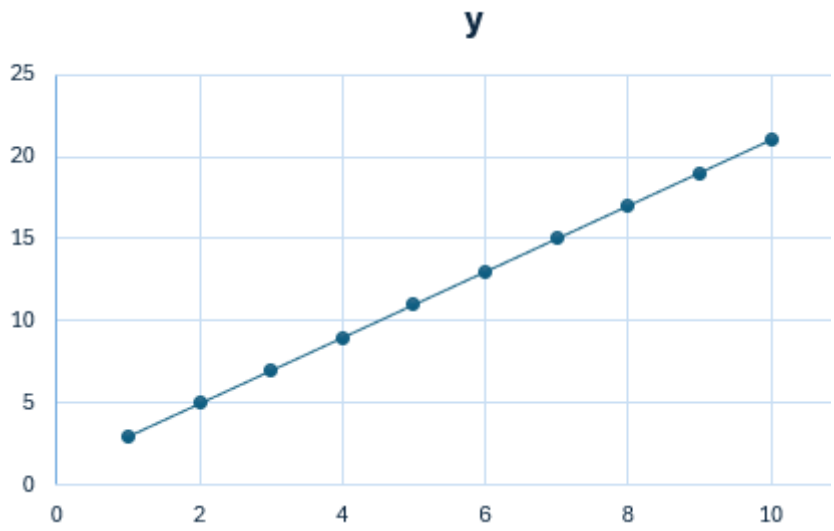
Nota: Na correlação linear nula não há um padrão linear, mas pode haver algum outro padrão. Por exemplo um padrão em forma de uma parábola, ou de uma circunferência, etc…



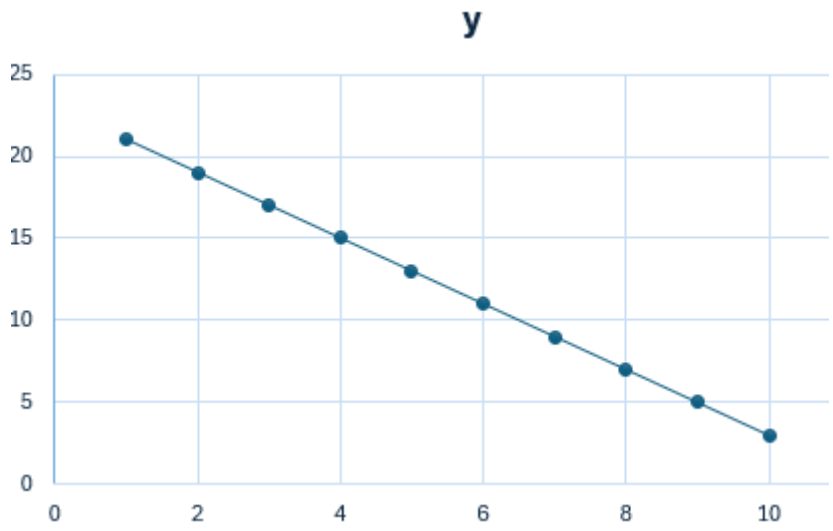
Relação linear perfeita.

Os pontos estão totalmente encaixados em cima da reta.

Exemplo: Correlação linear perfeita positiva:



Exemplo: Correlação linear perfeita negativa:



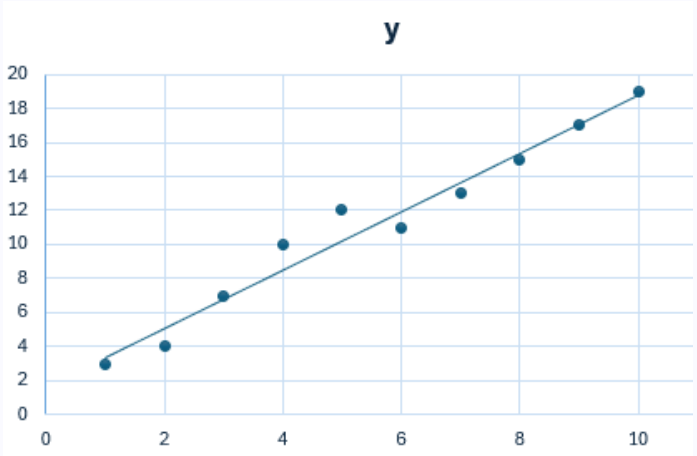
Reta de regressão.

Reta de regressão é a reta que melhor se ajusta aos pontos de um diagrama de dispersão.

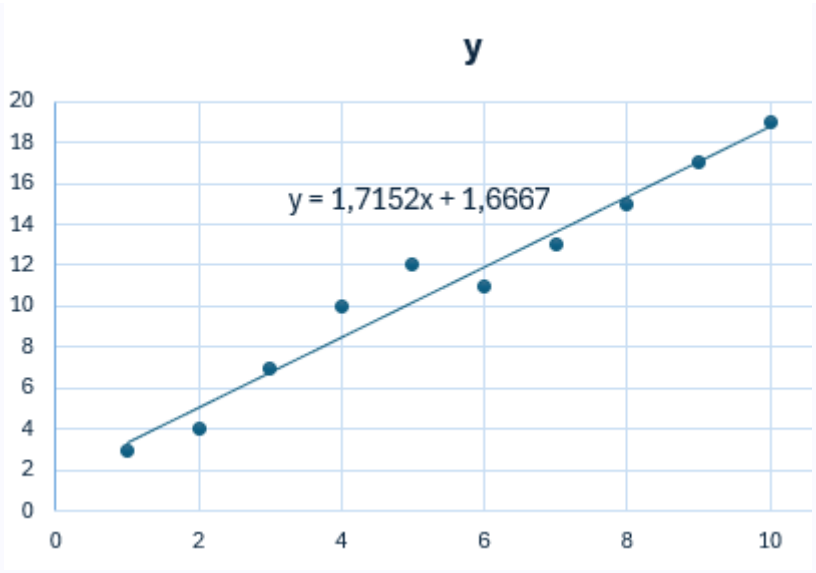
Exemplo:
Para os dados correspondentes aos pares (x,y) da tabela seguinte

x	1	2	3	4	5	6	7	8	9	10
y	3	4	7	10	12	11	13	15	17	19

Temos:



O até podemos indicar a equação da reta:



Nota: Esta reta serve sobretudo para descrever a relação entre as variáveis ou para fazer estimativas para valores desconhecidos.

Exemplo:
Qual é o valor aproximado de y, para o valor x=4.5?

Resposta:

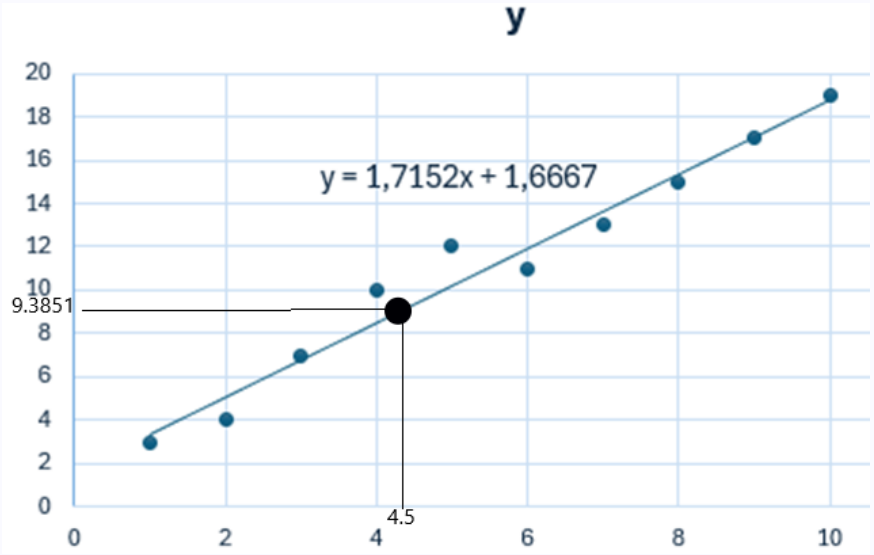
Substituímos na equação y = 1,7152x + 1,6667,

x por 4.5 e calculamos:

$Y(4.5) = 1{,}7152 \times 4.5 + 1{,}6667 \approx 9.3851$

O valor de y é aproximadamente 9.3851

Podemos interpretar no gráfico:



Importante: devemos evitar fazer previsões para valores de x que estejam fora do intervalo dos valores observados. Tais previsões podem nem fazer sentido!...

Nota importante: Na equação Y=ax+b, muitas vezes dão-nos um valor de x e pedem uma estimativa para o valor de y, como fizemos no exemplo acima.
Devemos **evitar** a situação contrária, isto é, estimar o valor de x a partir do valor de y.

Exemplo,
Consideremos uma equação que resultou de uma recolha de dados onde o ***x*** variou entre 3 e 18, tendo obtido a equação da reta de regressão:

$$Y= 0.58x+8.08.$$

Para x=14, qual seria a estimativa para Y=?

Resposta: basta substituir: $Y= 0.58\times14+8.08\approx 16.2$

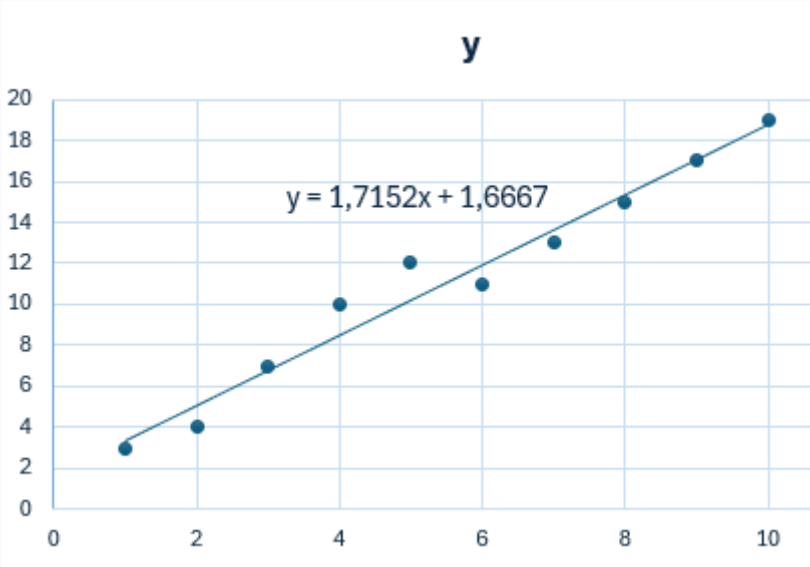
Centro de gravidade

Centro de gravidade (C) da nuvem de pontos é o ponto cujas coordenadas são as médias das distribuições em análise. C($\bar{x}$, $\bar{y}$).

Exemplo: Centro de gravidade.
Consideremos para os dados correspondentes aos pares (x,y) da tabela seguinte

x	1	2	3	4	5	6	7	8	9	10
y	3	4	7	10	12	11	13	15	17	19

Cuja reta de regressão está representada:



O centro de gravidade resulta de fazer: A média dos valores de x: 1,2,3,4,5,6,7,8,9,10.
Obtemos $\bar{x}$=5.5

E a média dos valores de y: 3, 4, 7, 10, 12, 11, 13, 15, 17, 19
Obtemos $\bar{y}$=11.1

Logo **C(5.5; 11.1)**

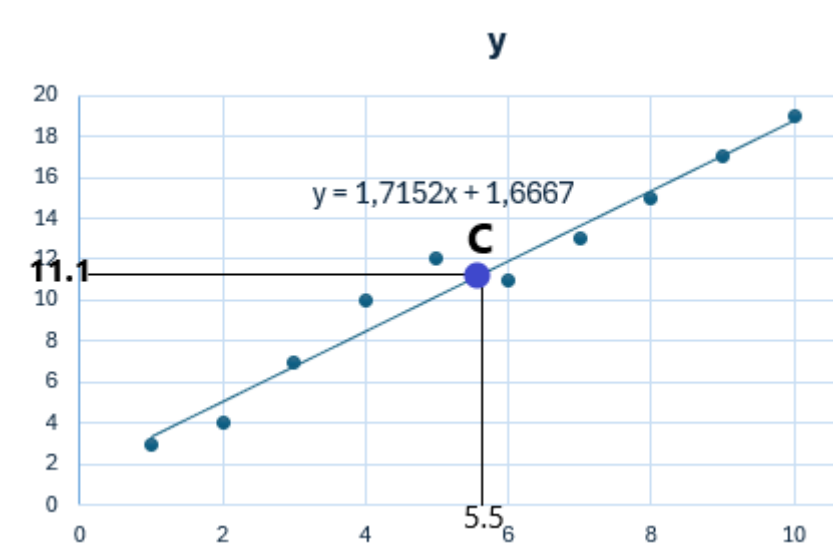
Podemos verificar que este ponto pertence à reta de regressão:

y = 1,7152x + 1,6667.

Bastaria substituir x por 5.5:

y = 1,7152×5.5 + 1,6667≈11.1.

Isto mostra que o ponto C($\bar{x}$, $\bar{y}$) pertence à reta de regressão, como podemos confirmar no gráfico:



Nota: de um modo geral, o centro de gravidade pertence sempre à reta de regressão.
Nota: Para obter o centro de gravidade na calculadora gráfica, pode ser prático recorrer ao calc 2var, onde encontramos as médias de x e de y.

Sugestões: Reta de regressão na calculadora gráfica:

Casio e Texas:

https://pedronoia.net/_private/Calculadoras/calc10EstatBiv2009.pdf

TI-Nspire:

https://pedronoia.net/nspire.pdf

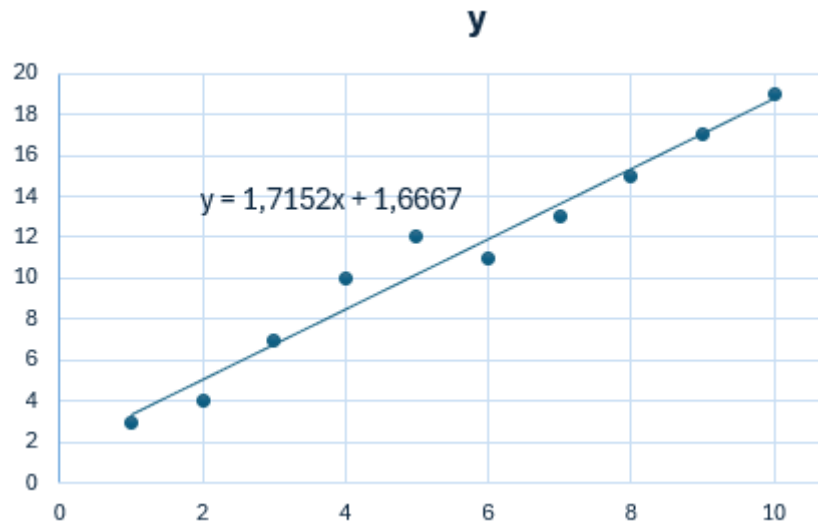
Numworks:

https://www.numworks.com/pt/professores/tutoriais/regressao/

Exemplo: **Efeito de acrescentar um ponto anómalo ao conjunto de dados**.

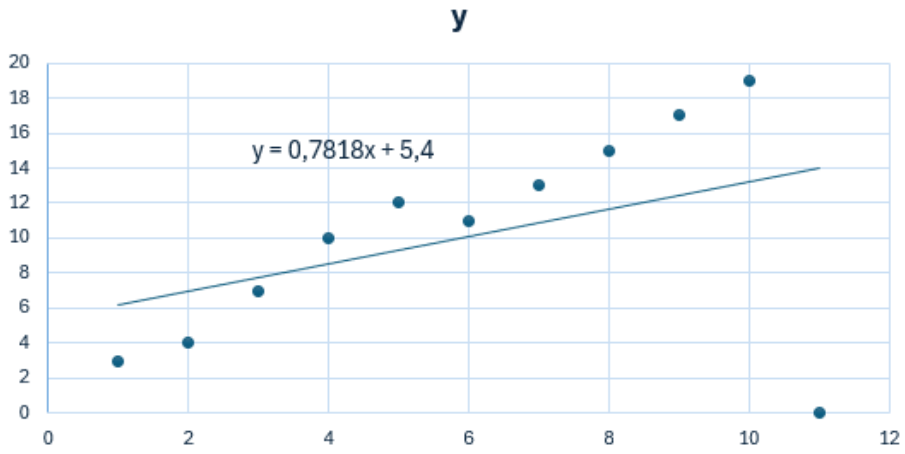
Retomando o exemplo da regressão baseada nos valores da tabela:

x	1	2	3	4	5	6	7	8	9	10
y	3	4	7	10	12	11	13	15	17	19



Se a esse conjunto de dados, fosse acrescentado o ponto (11, 0), o aspeto da reta de regressão mudaria de forma muito e a reta ficaria …

x	1	2	3	4	5	6	7	8	9	10	11
y	3	4	7	10	12	11	13	15	17	19	0



Haveria uma mudança muito grande, quer no ajustamento dos pontos em relação à reta, quer mesmo, na equação da reta.

Poderia ser mais prudente eliminar este último ponto, por ser um ponto discrepante…

Nota: O exemplo anterior mostra que nunca devemos considerar a reta de regressão independente da nuvem de pontos, pois a existência de **pontos anómalos** pode induzir em erro o cálculo de estatísticas e de estimativas. A eliminação destes pontos conduz a estatísticas e estimativas mais realistas.

Coeficiente de correlação linear.

Já vimos em exemplos anteriores que os pontos podem estar mais próximos ou mais afastado da reta de regressão. Para medir o melhor ou pior ajustamento dos pontos à reta de regressão, usamos o "coeficiente de regressão linear".

O **coeficiente de correlação linear** mede o grau de associação linear entre duas variáveis. Representa-se por ***r*** e varia entre –1 e 1.

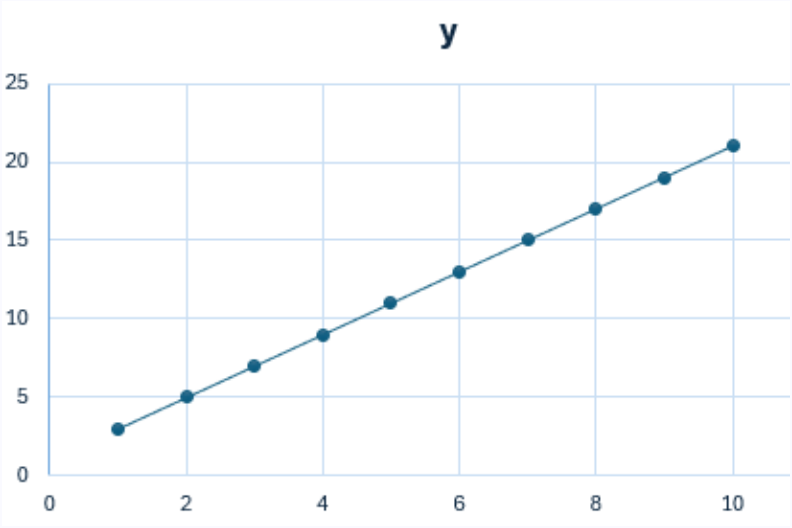
Se r= 1, a correlação é total (ou perfeita) positiva.

Se r= 0, a correlação é nula: não há correlação linear.

Se r= -1, a correlação é total (ou perfeita) negativa.

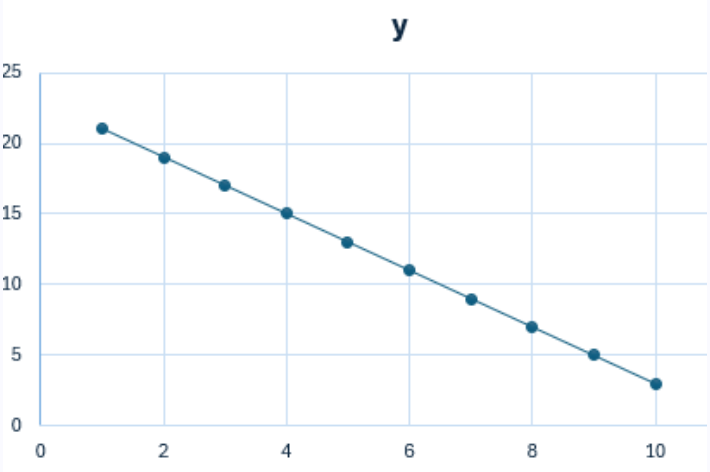
Exemplo 1:

Correlação perfeita positiva r=1.

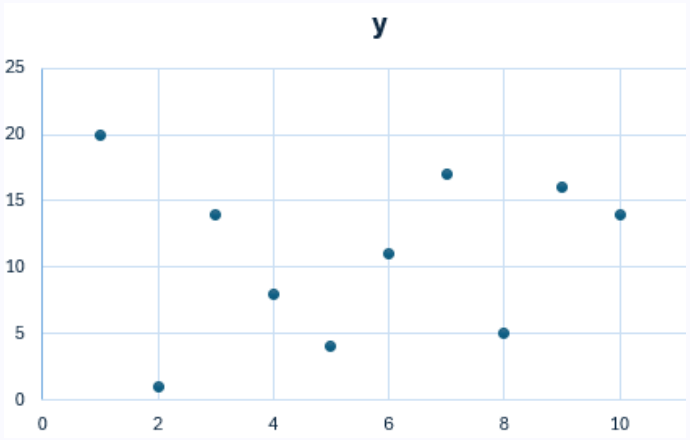


Exemplo 2:

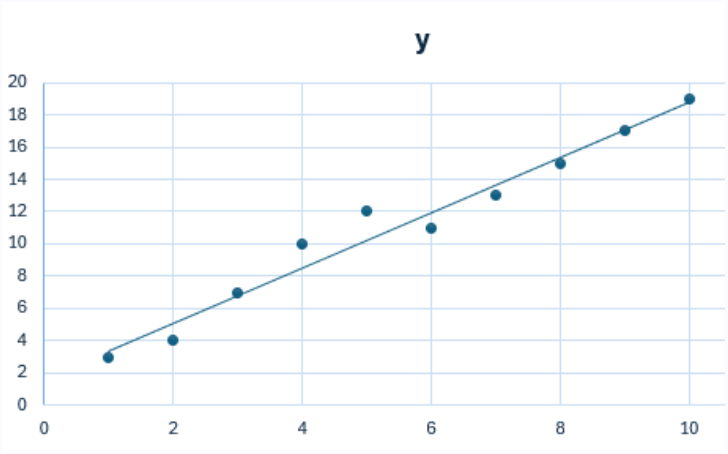
Correlação perfeita negativa r= -1.



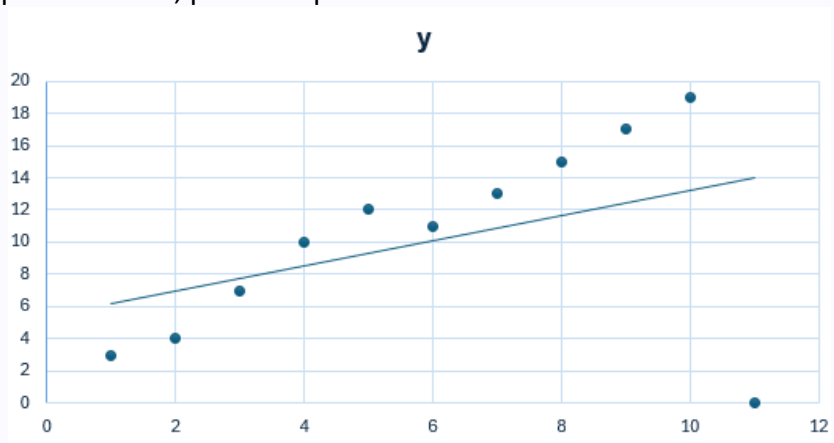
Exemplo 3:
Correlação linear nula r=0



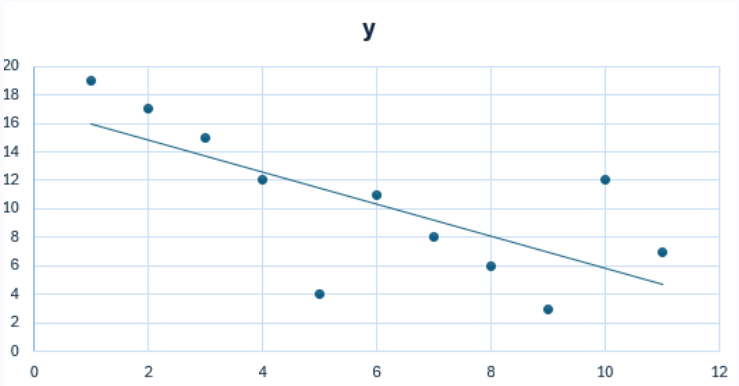
Exemplo 4:
Correlação positiva forte, por exemplo r=0.984



Exemplo 5:
Correlação positiva fraca, por exemplo r= 0.43



Exemplo 6:
Correlação negativa, por exemplo r= - 0.71



Nota: Para os valores intermédios, a correlação é tanto mais forte quanto mais próximo o valor de r se encontrar de 1 ou de –1, enfraquecendo à medida que se aproxima de zero.

Nota: o facto de o coeficiente de correlação ser zero, ou próximo de zero, não significa que as duas variáveis não tenham qualquer relação. Podem ter uma relação não linear, por exemplo, pode ter a forma de uma parábola, ou de uma circunferência.

Nota: o facto de o coeficiente de correlação ser elevado não significa que as duas variáveis tenham uma relação de causa-efeito.

Por ***exemplo***, se a variável x for o número de garrafas de água vendidas mensalmente e y for o número de casamentos em cada mês, naturalmente que, nos meses em que se vendem mais garrafas de água, existem mais casamentos. De desenhássemos um gráfico com estas duas variáveis, haveria uma correlação positiva forte.

Isto nunca poderia significar que, o aumento da venda de garrafas de água provoca um aumento do número de casamentos!... apenas, por coincidência, nos meses de verão é que se vendem mais garrafas de água e também há mais casamentos.

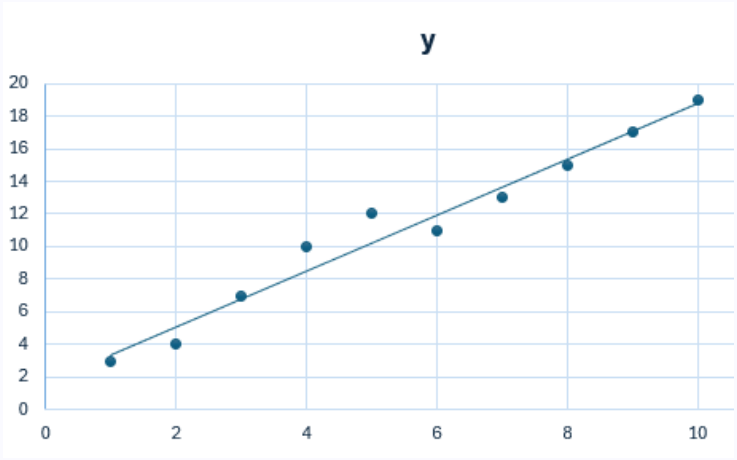
Coeficiente de correlação -Calculadora gráfica.

Os procedimentos são os mesmos que vimos para a reta de regressão. O coeficiente de correlação é “r” e surge juntamente com a reta de regressão.

Apenas nas calculadoras **Texas**, pode ser necessário pedir explicitamente o valor de “r”, fazendo: //Catalog//Diagnostic On// Enter//. Ao pedir a correlação linear, já aparecerá o valor de “r”.

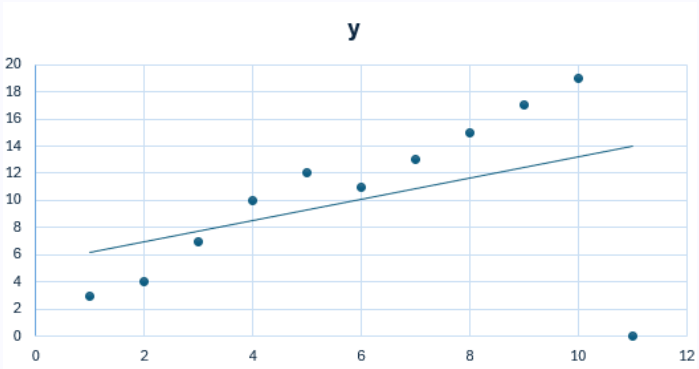
Exemplos: Efeito de acrescentar um ponto anómalo ao conjunto de dados.

r=0.984



Exemplo:

r= 0.43



Nota: Os dois exemplos anteriores mostram que a existência de **pontos anómalos(*outliers*)** pode condicionar o valor do coeficiente de correlação linear. Por vezes, a eliminação destes pontos conduz a uma interpretação mais realista dos dados.

Exemplo

Consideremos os pesos e as alturas de alguns alunos:

X Altura(cm)	145	150	156	160	167	174	171	170	176	182	175	174	162
Y Peso(kg)	55	58	56	61	65	70	69	72	71	78	76	70	64

.1) Obtenha o coeficiente de correlação e a equação da reta de regressão

.2) Com base na equação da reta de regressão obtida, determine:

.2.1) O peso esperado para um aluno com 169 cm.

.2.2) A altura esperada para um aluno com 75 kg.

Resolução:

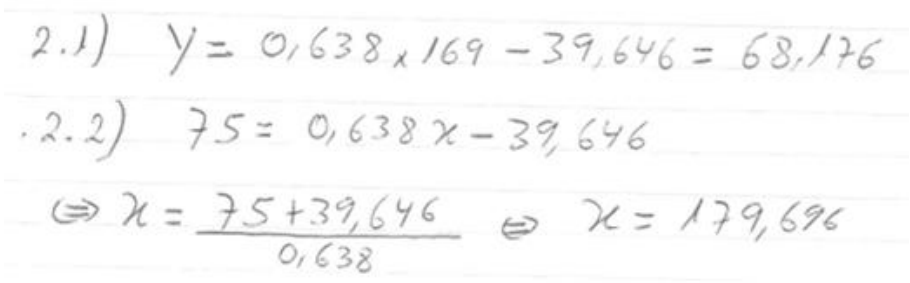
.**1)** Colocamos os dados das tabela na calculadora. Na lista 1 colocamos a altura e na lista 2 colocamos o peso:

Lista 1	145	150	156	160	167	174	171	170	176	182	175	174	162
Lista 2	55	58	56	61	65	70	69	72	71	78	76	70	64

Pedimos o coeficiente de correlação e a reta de regressão.

Obtemos o valor r=0.9499≈0.950 e a equação da reta Y=0.638x-39.646

.**2.1)** Y(169)=....



Tabelas de contingência.

Quando pelo menos uma das variáveis estatísticas em estudo é do tipo qualitativo, recorre-se à representação dos dados em **tabelas de contingência**.

Exemplo 1:

Consideremos um conjunto de homens e mulheres, que foram inquiridos num centro comercial, acerca do hábito de fumar. Uns fumam e outros não fumam como se pode ver na tabela abaixo:

	Homens	**Mulheres**	**Total**
Fumam	100	50	150
Não fumam	200	150	350
Total	**300**	**200**	**500**

Podemos extrair várias informações a partir dos dados da tabela:

As mulheres representam 40% do total das pessoas inquiridas, pois

$$\frac{200}{500} \times 100\% = 40\%$$

Entre os homens, que são ao todo 300, cerca de 33.33% fumam:

$$\frac{100}{300} \times 100\% \approx 33.33\%$$

Entre as mulheres, apenas 25% fumam:

$$\frac{50}{200} \times 100\% = 25\%$$

Muitas outras conclusões poderiam ser tiradas a partir dos dados da tabela de contingência dada.

Exemplo 2:

A tabela refere-se a 200 alunos que frequentam o 10º ano.

	Rapaz	Rapariga	TOTAL
Frequenta Matemática A			
Não frequenta Matemática A			
TOTAL			

Complete a tabela acima tendo em conta que:

60 % dos alunos frequentam Matemática A.

$\frac{3}{8}$ dos alunos são rapazes.

40% das raparigas não frequenta Matemática A

Resolução:

Como 60% dos alunos frequenta matemática A, estes correspondem ao total:

$0.6 \times 200 = 120$.

Como $\frac{3}{8}$ dos alunos são rapazes, fazemos $\frac{3}{8} \times 200 = 75$ (total de rapazes).

Sendo 75 rapazes, então o número de raparigas é 200-75=125.

Atendendo a que 40% das raparigas não frequentam Matemática A, estas serão ao todo: $0.4 \times 125 = 50$.

Neste momento, já podemos preencher os seguintes valores da tabela:

	Rapaz	Rapariga	TOTAL
Frequenta Matemática A			120
Não frequenta Matemática A		50	
TOTAL	75	125	200

Podemos agora completar a tabela, completando as linhas e as colunas:

	Rapaz	Rapariga	TOTAL
Frequenta Matemática A	45	75	120
Não frequenta Matemática A	30	50	80
TOTAL	75	125	200

Exemplo 3.

Na tabela seguinte, temos dados referentes à cor do cabelo e à cor dos olhos de 500 pessoas.

Sabemos que: 1/4 das pessoas com cabelo castanho tem olhos verdes. 20% das pessoas com cabelo louro tem olhos castanhos. 62,5% das pessoas com olhos azuis tem cabelo louro. Há mais 70 pessoas com olhos castanhos do que pessoas com olhos verdes.

Cabelo/Olhos	Azuis	Castanhos	Verdes	Total
Castanho			***a*** =	300
Louro	***b =***			150
Ruivo		10		50
Total	80		***c =***	500

Calcule os valores de ***a***, ***b*** e ***c***, indicando todos os cálculos e raciocínios. Se estiver mal justificado ou fizer apenas por tentativas, será considerado errado.Depois, copie a tabela para a sua folha de respostas e preencha-a.

Resolução:

1/4 de 300 é 300/4 = 75, logo **a=75**

62.5% de 80 é 0.625×80=50 logo **b=50**

Como 500-80 = 420 e 70+c+c=420 ⇔ 2c=420-70 ⇔ 2c=350⇔ c=350/2 ⇔ **c= 175**

Cabelo/Olhos	Azuis	Castanhos	Verdes	Total
Castanho	20	205	75	300
Louro	50	30	70	150
Ruivo	10	10	30	50
Total	80	245	175	500