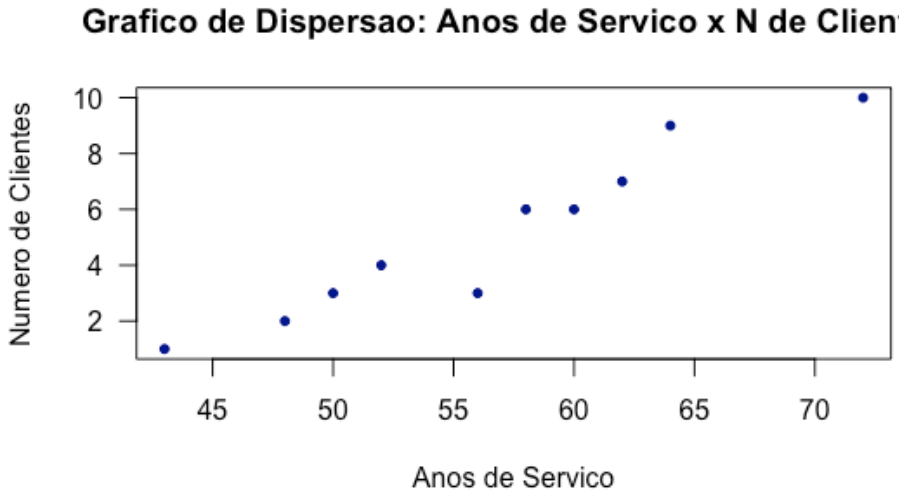
Noções de Estatística

Aula 4

- Regressão Linear
- QQ Plot

4.1 Introdução a Regressão Linear

Vamos retomar um gráfico de dispersão visto na aula sobre análise bidimensional:



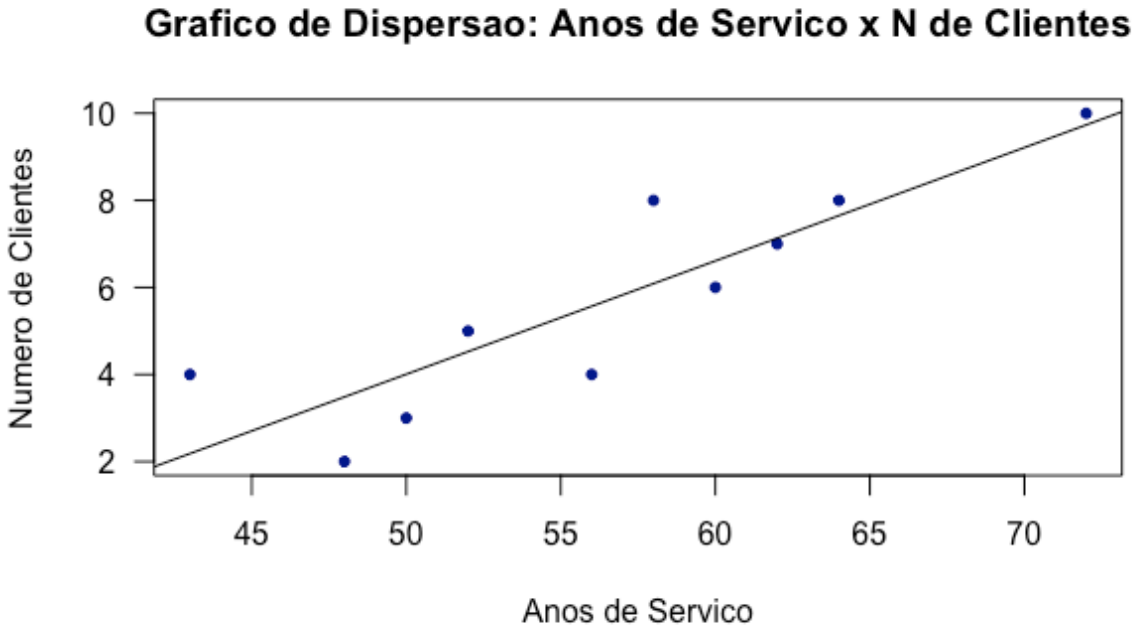
Observe que aparentemente existe uma relação linear entre o número de clientes e os anos de serviço para o nosso funcionário hipotético.

Esse comportamento linear muito nos interessa pois, em primeiro lugar, facilita a nossa interpretação quando observamos os dados. Em segundo lugar, nos ajuda a realizar uma possível previsão para dados que não foram observados. Em outras palavras, suponha que

exista um funcionário com 20 anos de serviço, ou 10, ou 80. Esses valores não foram observados nos nossos dados, porém ainda assim, dado o caráter linear do que observamos, é possível fazer uma previsão.

Se pudermos traçar uma reta que represente esse comportamento, poderemos estimar o número de clientes de um funcionário qualquer, dada apenas a informação dos anos de serviço.

Nosso objetivo é traçar uma reta como a reta abaixo:



Vamos relembrar a equação de uma reta?

$$Y = a + bx$$

Veja que a obtenção de Y depende de X, ou seja, está em função de X. E o que são **a** e **b** mesmo?

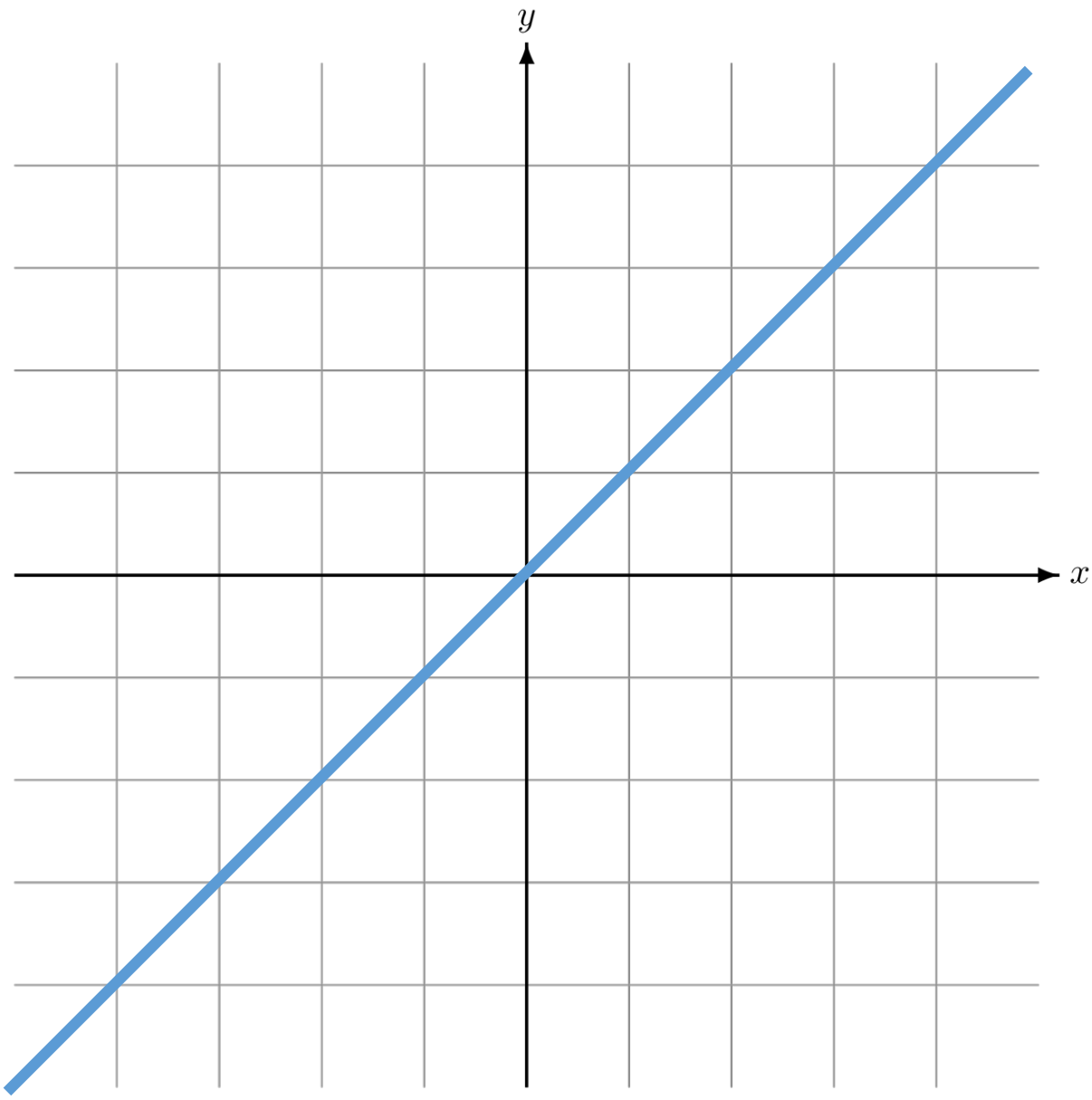
O coeficiente angular está representado pela letra **b**. Ele nos dá a informação relativa ao ângulo da nossa reta

no plano cartesiano. Obs: não é o ângulo em graus, veja os exemplos para entender melhor.

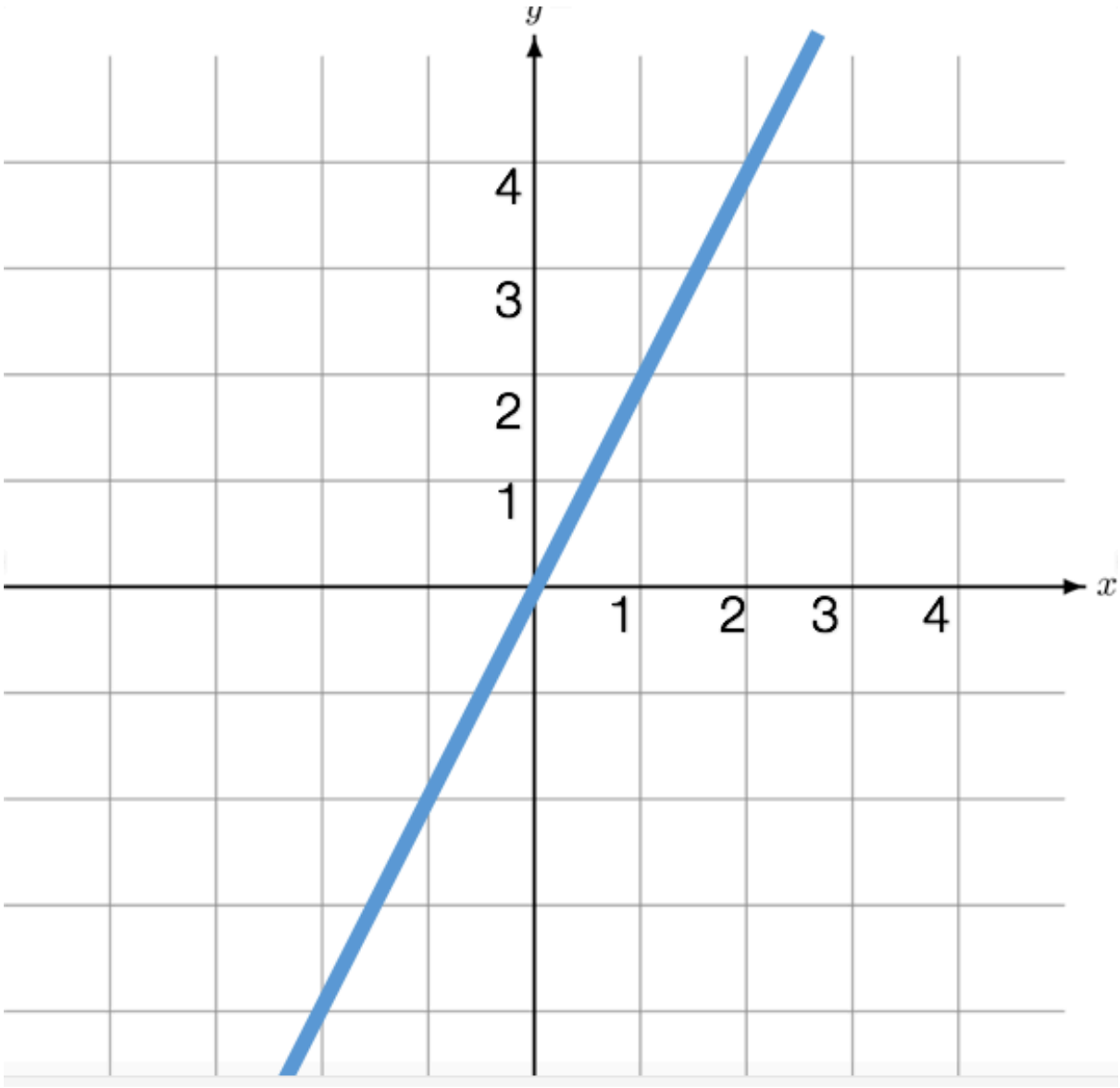
Suponha que **b** seja igual a 1. (E suponha que **a** seja igual a zero, só para compreendermos inicialmente o **b**)
Então teremos a equação:

$$Y = X$$

No plano cartesiano, essa reta corresponde a uma reta de 45º. E todo Y é exatamente igual ao X.



Ao alterarmos os valores de **b**, estaremos mudando a angulação desta reta.



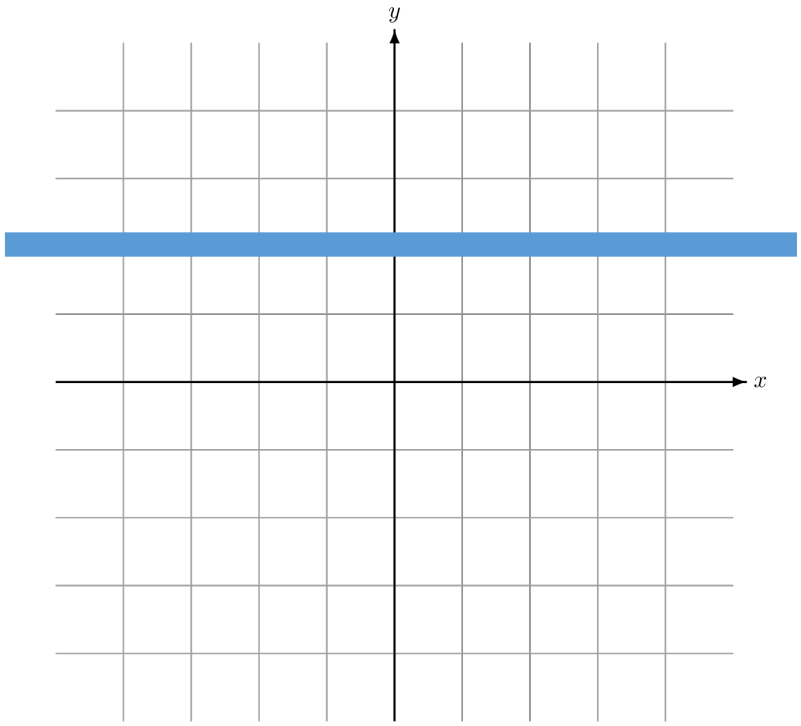
Y = 2x

E o **a**?

O **a** corresponde ao intercepto. O intercepto é o valor de Y quando x = 0. Ou melhor, quando a reta cruza o eixo y.

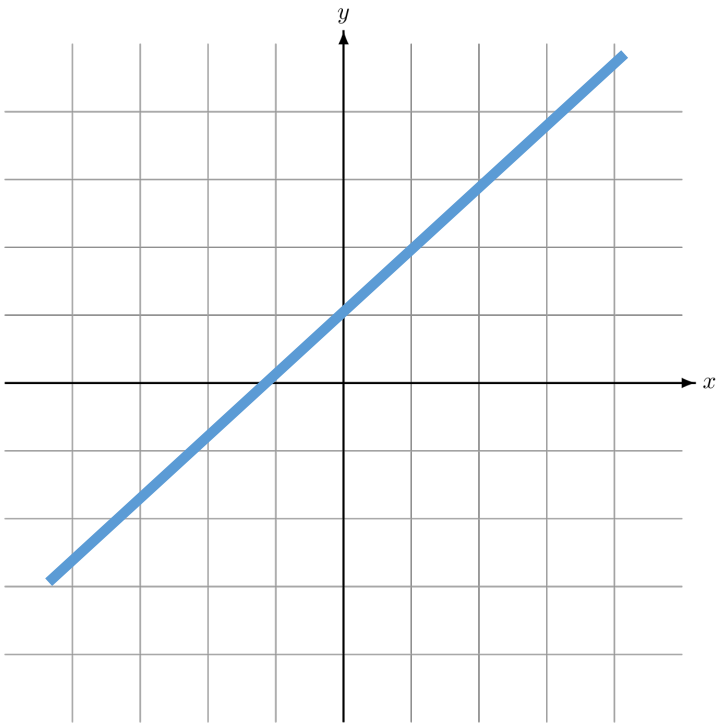
Quando o coeficiente angular é igual a zero, obtemos uma reta constante. Para todo x, y é o mesmo.

Y = 2



Veja agora um exemplo com **b** diferente de 0.

$$y = \frac{1}{2}\,x + 1$$



Concluída a nossa revisão sobre retas, como obtemos a reta de regressão linear?

A reta de regressão linear é obtida pelo método de **mínimos quadrados**.

O método de mínimos quadrados visa garantir a formulação de uma reta cuja a distância entre todos os pontos e a própria reta seja mínima, gerando, portanto, uma reta ótima.

Sem entrar nos detalhes do método de mínimos quadrados, podemos apenas dizer que já existem equações prontas para a obtenção dos coeficientes desejados. São elas:

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{Y}\bar{X}}{(n-1)S_x^2}$$

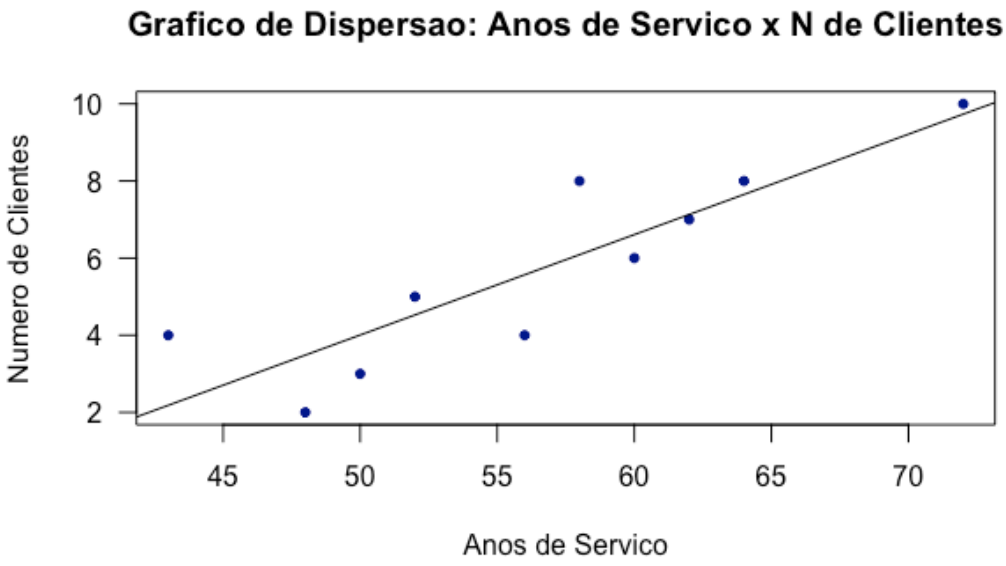
$$a = \bar{y} - b\bar{x}$$

Interpretação dos coeficientes:

Para um aumento de uma unidade na variável x, a variável y aumenta em média **b** unidades.

Quando a variável x é igual a zero, a variável y é **a**.

No nosso exemplo...



A reta acima foi obtida a partir do método dos mínimos quadrados com coeficientes:

b = 0.2604
a = -9.014

4.2 Gráficos q x q:

Observe abaixo uma tabela que compara as notas de alunos em duas provas de estatística:

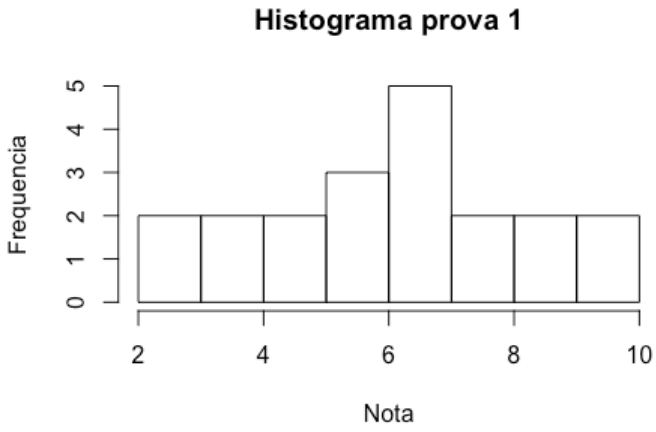
aluno	prova1	prova2
1	8.5	8
2	3.5	2.8
3	7.2	6.5
4	5.5	6.2
5	9.5	9
6	7	7.5
7	4.8	5.2
8	6.6	7.2
9	2.5	4
10	7	6.8
11	7.4	6.5
12	5.6	5
13	6.3	6.5
14	3	3
15	8.1	9
16	3.8	4
17	6.8	5.5
18	10	10
19	4.5	5.5
20	5.9	5

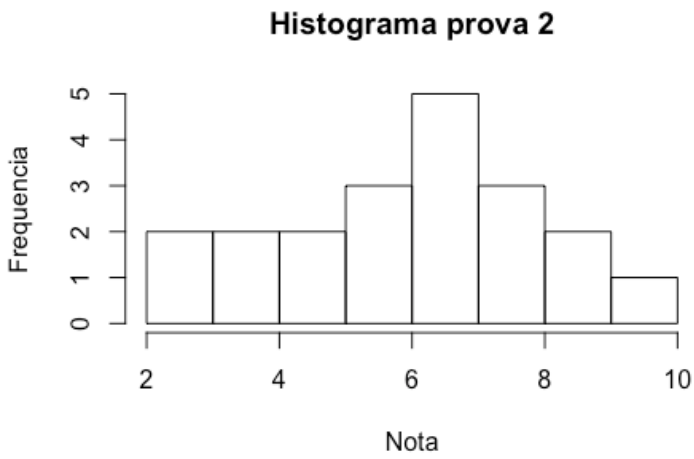
E, para cada prova, suas respectivas medidas-resumo:

```
     prova1           prova2
Min.   : 2.500   Min.   : 2.800
1st Qu.: 4.725   1st Qu.: 5.000
Median : 6.450   Median : 6.350
Mean   : 6.175   Mean   : 6.160
3rd Qu.: 7.250   3rd Qu.: 7.275
Max.   :10.000   Max.   :10.000
```

Pode ser mais interessante compararmos os desempenhos dos alunos através de uma análise referente a distribuição de frequências de notas para cada prova. Em outras palavras, queremos saber se os alunos tiveram (em geral) desempenho semelhante nas duas provas.

Para isso, podemos comparar as medidas resumo e os histogramas.





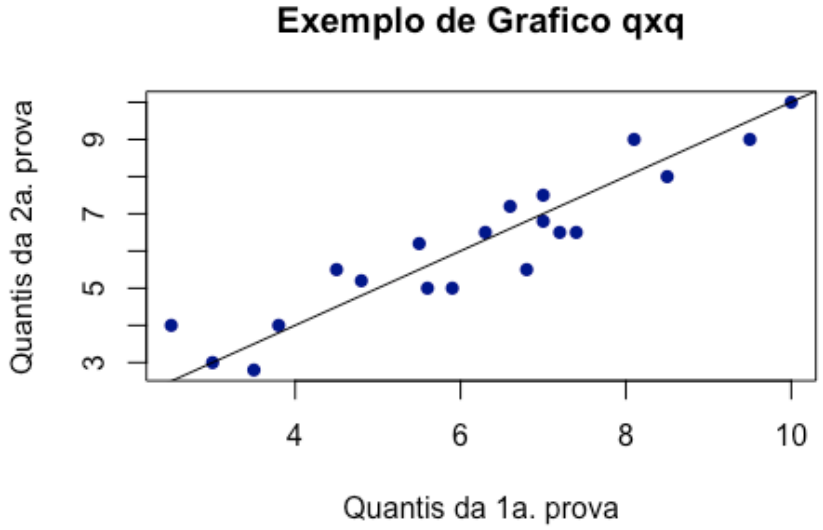
Observe que as duas distribuições são muito semelhantes.

Isso também pode ser verificado através do gráfico qxq.

O gráfico qxq é utilizado para compararmos duas distribuições diferentes.

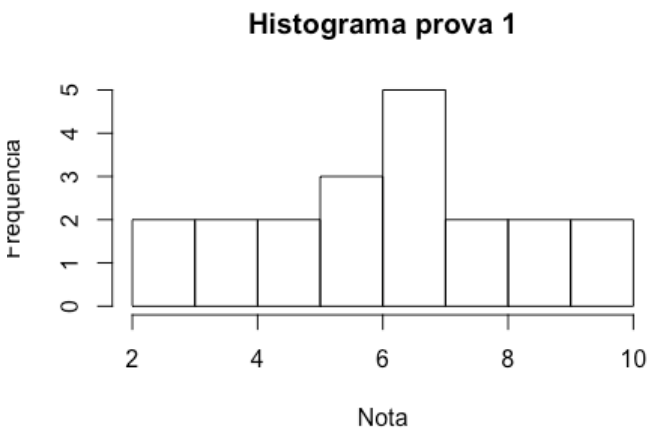
Para a construção de um gráfico qxq, é necessário encontrar os quantis de cada variável de interesse.

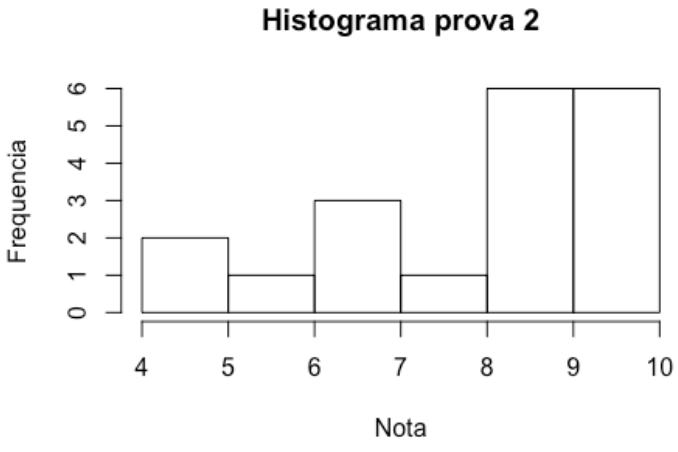
Cada ponto exibido no gráfico corresponde ao ponto (quantil na primeira prova, quantil na segunda prova). Se os pontos estiverem alinhados, isso significa que as duas variáveis comparadas tem comportamentos semelhantes.



Vejamos um exemplo para variáveis com comportamentos diferentes:

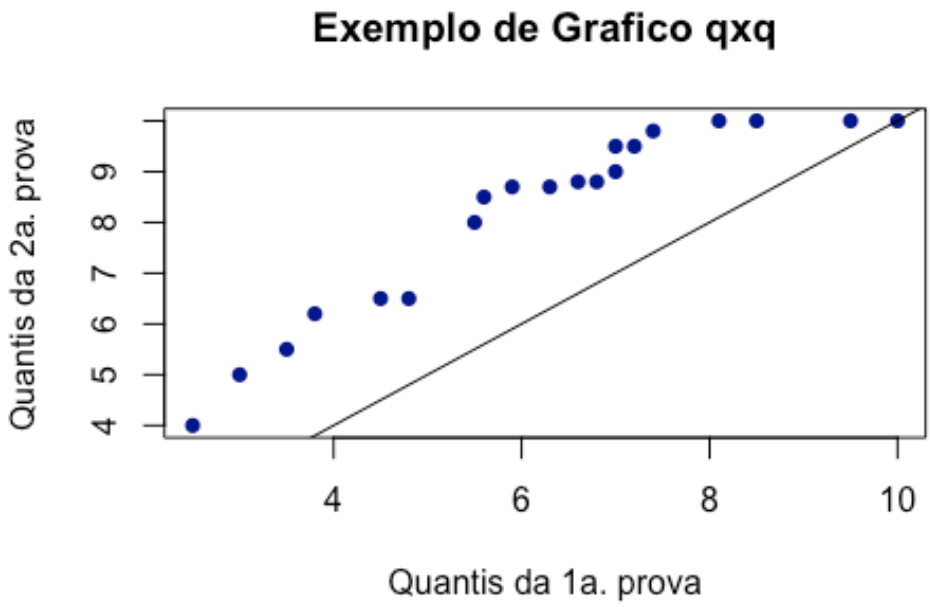
Vamos supor que para a segunda prova, tivéssemos observado uma distribuição diferente.





Os alunos tiveram desempenho muito melhor na segunda prova de acordo com a comparação dos histogramas.

Veja o que acontece com o gráfico qxq:



Como podemos observar, os pontos se encontram em sua grande maioria acima da reta de comparação. Isso reflete o melhor desempenho dos alunos na segunda prova com relação à primeira prova.