

## CASO PRÁTICO 3 – Tópicos de Resolução

### Estatística Descritiva

#### Exercício 1:

Para analisar o comportamento de compra de seus clientes, uma empresa selecionou aleatoriamente vários clientes e recolheu informação sobre as suas compras recentes. Duas das características analisadas foram a idade e o sexo. Analise e compare o comportamento dos clientes tendo em conta as duas características e considerando as seguintes tabelas.

Dados disponíveis no moodle, Caso Prático 3, nos ficheiros CP3.xlsx (Exerc1) e RScript CP3\_1

Categoria	Freq_Abs	Percentagem
18-34 anos	237	47.4
35-54 anos	124	24.8
55 ou mais	139	27.8

Categoria	Freq_Abs	Percentagem
Feminino	240	48
Masculino	260	52

	Feminino	Masculino
18-34 anos	131	106
35-54 anos	43	81
55 ou mais	66	73

	Feminino	Masculino
18-34 anos	26.2	21.2
35-54 anos	8.6	16.2
55 ou mais	13.2	14.6

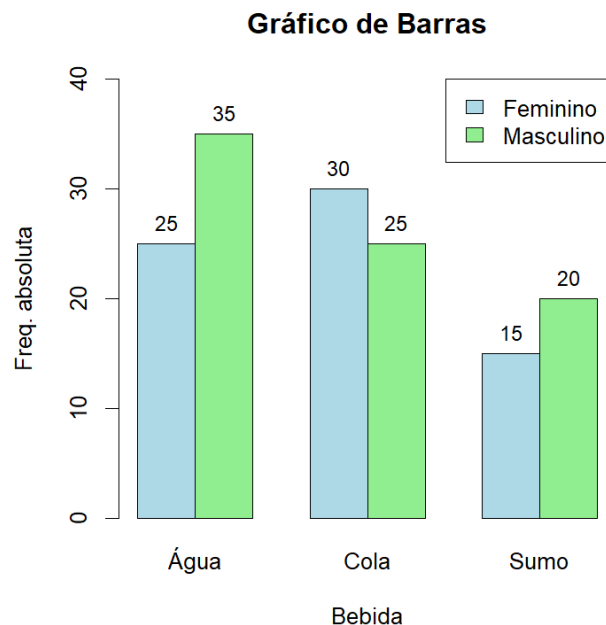
Dos 500 clientes que efectuaram compras,

- 240 (48%) são do sexo feminino e 260 (52%) do sexo masculino → mais clientes do sexo masculino;
- 237 (47,4%) têm 18 a 34 anos, 124 (24,8%) 35 a 54 anos e os restantes 139 (27,8%) têm 55 ou mais anos → mais clientes na faixa etária dos 18 aos 34 anos;
- 26,2% (131) são do sexo feminino na faixa etária dos 18 aos 34 anos, 21,2% (106) são do sexo masculino na mesma faixa etária;
- 16,2% (81) são do sexo masculino e têm 35 a 54 anos;
- 14,6% (73) são do sexo masculino com 55 ou mais anos;
- 13,2% (66) são do sexo feminino com 55 ou mais anos;
- 8,6% são do sexo feminino na faixa etária dos 35 aos 54 anos;

### Exercício 2:

Considerando o gráfico de barras seguinte, analise a preferência dos alunos de uma escola relativamente ao tipo de bebida (Cola/Sumo/Água) que geralmente consomem ao almoço.

Dados disponíveis no moodle, Caso Prático 3, nos ficheiros CP3.xlsx (Exerc2) e RScript CP3\_2



Participaram no estudo 150 alunos, dos quais 70 são do sexo feminino e 80 do sexo masculino.

60 inquiridos indicaram que geralmente bebem água ao almoço, 55 Cola e 35 Sumo. A bebida preferida dos inquiridos é a Cola, seguida pela Água e a bebida menos consumida é o Sumo.

No grupo dos indivíduos do sexo feminino a bebida preferida é a Cola (30), seguida pela Água (25) e a menos consumida é o Sumo (15).

A bebida preferida dos indivíduos do sexo masculino é a Água (35), seguida pela Cola (25) verificando-se também que a bebida menos consumida é o Sumo (20).

### Exercício 3:

Uma empresa analisou a eficácia de uma campanha de marketing efectuada por e-mail em termos do número de cliques num link (Clica/Não clica) e o tipo de conteúdo do e-mail (promocional/informativo). A empresa enviou 100 e-mails promocionais e 100 e-mails informativos e registou o número de destinatários que clicaram num link do e-mail.

Dados disponíveis no moodle, Caso Prático 3, ficheiro CP3.xlsx (Exerc3).

a) Obtenha a tabela de contingência e o respectivo gráfico e barras para as duas variáveis.

Importar os dados do separador **Exerc3** do ficheiro em excel **CP3.xlsx**

```
library(readxl)
```

```
CP3 <- read_excel("C:/...(caminho).../CP3.xlsx", sheet = "Exerc3")
```

```
view(CP3)
```

	Click	Email
23	Clicou	Promocional
24	Clicou	Promocional
25	Clicou	Promocional
26	Não clicou	Promocional
27	Não clicou	Promocional
28	Não clicou	Promocional
29	Não clicou	Promocional

Identificar o nº de observações por variável

```
library(dplyr)
```

```
glimpse(CP3)
```

```
Rows: 200
```

```
Columns: 2
```

```
$ Click <chr> "Clicou", "Clicou", "Clicou", "Clicou", "Clicou", "Clicou", "..."
```

```
$ Email <chr> "Promocional", "Promocional", "Promocional", "Promocional", "..."
```

2 variáveis com 200 observações cada.

```
table(CP3$Click, CP3$Email)
```

	Informativo	Promocional
Clicou	20	25
Não clicou	80	75

```
prop.table(table(CP3$Click, CP3$Email))
```

	Informativo	Promocional
Clicou	0.100	0.125
Não clicou	0.400	0.375

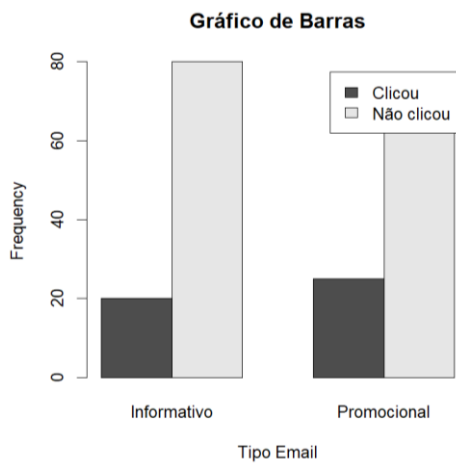
```
prop.table(table(CP3$Click, CP3$Email))*100
```

	Informativo	Promocional
Clicou	10.0	12.5
Não clicou	40.0	37.5

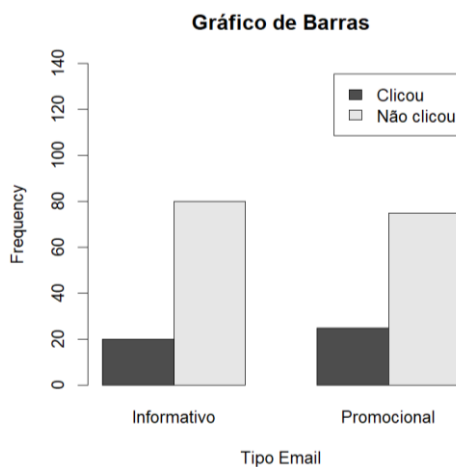
```
data.frame(Categoria1 = rep(rownames(table(CP3$Click, CP3$Email)), each = ncol(table(CP3$Click, CP3$Email))),
           Categoria2 = rep(colnames(table(CP3$Click, CP3$Email)), times = nrow(table(CP3$Click, CP3$Email))),
           Frequencia_Absoluta = as.vector(table(CP3$Click, CP3$Email)),
           Frequencia_Relativa = as.vector(prop.table(table(CP3$Click, CP3$Email))))
```

	Categoria1	Categoria2	Frequencia_Absoluta	Frequencia_Relativa
1	Clicou	Informativo	20	0.100
2	Clicou	Promocional	25	0.125
3	Não clicou	Informativo	80	0.400
4	Não clicou	Promocional	75	0.375

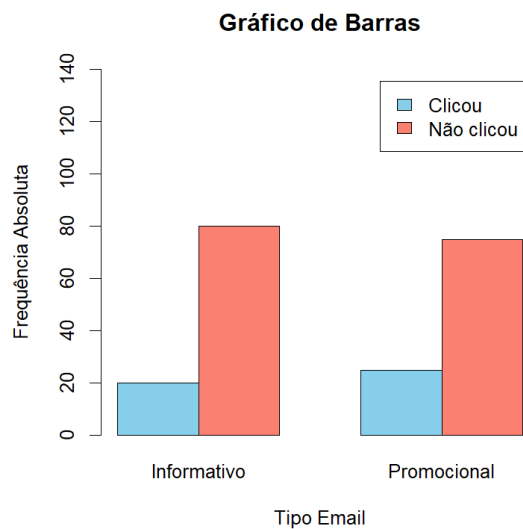
```
barplot(table(CP3$Click, CP3$Email), beside = TRUE,
        legend = rownames(table(CP3$Click, CP3$Email)),
        main = "Gráfico de Barras",
        xlab = "Tipo Email",
        ylab = "Frequency")
```



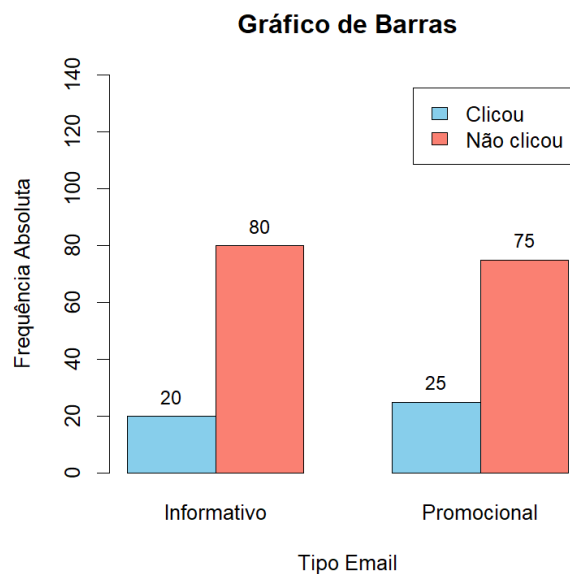
```
barplot(table(CP3$Click, CP3$Email),
        beside = TRUE,
        legend = rownames(table(CP3$Click, CP3$Email)),
        main = "Gráfico de Barras",
        xlab = "Tipo Email",
        ylab = "Frequency",
        ylim = c(0,140))
```



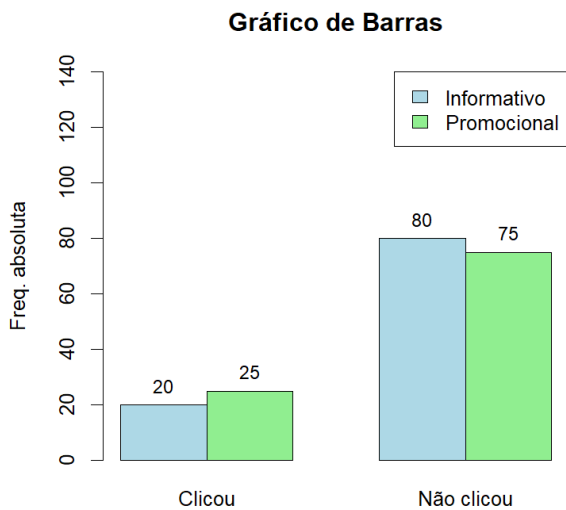
```
barplot(table(CP3$Click, CP3$Email),
        beside = TRUE,
        legend = rownames(table(CP3$Click, CP3$Email)),
        main = "Gráfico de Barras",
        xlab = "Tipo Email",
        ylab = "Frequência Absoluta",
        col = c("skyblue", "salmon"),
        ylim = c(0,140))
```



```
text(x=barplot(table(CP3$Click, CP3$Email),
  beside = TRUE,
  legend = rownames(table(CP3$Click, CP3$Email)),
  main = "Gráfico de Barras",
  xlab = "Tipo Email",
  ylab = "Frequência Absoluta",
  col = c("skyblue", "salmon"),
  ylim = c(0,140)), y = table(CP3$Click, CP3$Email),
  label = table(CP3$Click, CP3$Email),
  pos = 3, cex = 0.9, col = "black"
)
```



```
text(x = barplot(table(CP3$Email, CP3$Click), beside = TRUE,
  col = c("lightblue", "lightgreen"),
  legend.text = rownames(table(CP3$Email, CP3$Click)),
  args.legend = list(x = "topright"),
  ylim = c(0,140),
  main = "Gráfico de Barras",
  xlab = "",
  ylab = "Freq. absoluta"),
  y = table(CP3$Email, CP3$Click),
  label = table(CP3$Email, CP3$Click),
  pos = 3, cex = 0.9, col = "black")
```



b) Analise a tabela de contingência e o gráfico de barras obtidos.

Foram enviados 200 emails, 100 informativos e 100 promocionais.

77,5% dos destinatários não clicaram num link, apenas 22,5% dos destinatários clicaram num link.

O número de destinatários que clicaram num link de um email promocional (25) é superior ao número de destinatários que clicaram num link dum email informativo (20),

ou

a percentagem de destinatários que clicaram num link de um email promocional (12,5%) é superior à percentagem de destinatários que clicaram num link dum email informativo (10%).

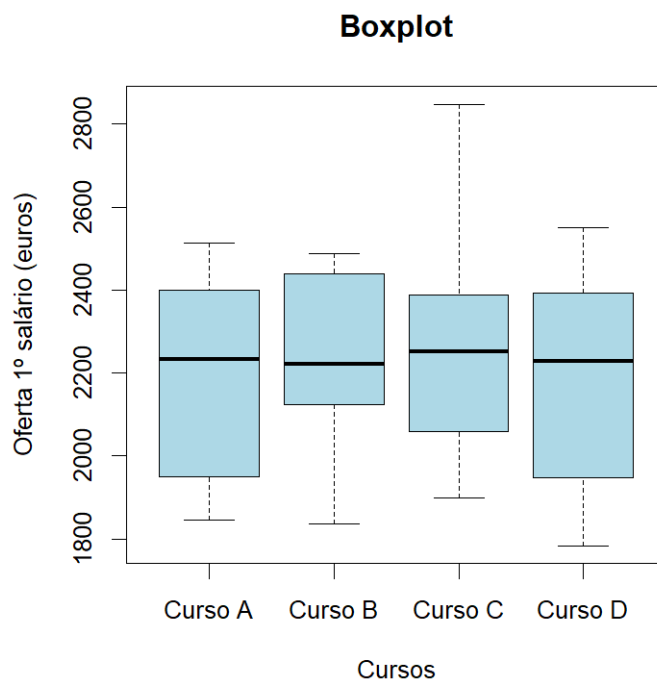
#### Exercício 4:

Uma instituição do ensino superior realizou um estudo com o objetivo de analisar algumas características dos alunos que concluíram o 1º ciclo. Algumas das variáveis analisadas foram:

OFERTA Maior salário oferecido para o primeiro trabalho após a conclusão do curso (em euros)

CURSO (Curso A/Curso B/Curso C/ Curso D). Analise as Box-plot.

**Dados disponíveis no moodle, Caso Prático 3, nos ficheiros CP3.xlsx (Exerc4) e RScript CP3\_4.**



As medianas dos maiores salários oferecidos para o 1º trabalho após a conclusão do curso dos quatro cursos são semelhantes (aproximadamente 2200 euros), sendo a do curso C ligeiramente superior às restantes.

O menor salário para o 1º trabalho após a conclusão do curso foi oferecido a um aluno do curso D (ligeiramente inferior a 1800 euros) e o maior a um aluno do curso C (superior a 2800 euros).

A dispersão (em termos da amplitude total) dos maiores salários para o 1º trabalho após a conclusão do curso C é superior à dispersão dos maiores salários oferecidos para o 1º trabalho após a conclusão dos restantes cursos.

As amplitudes interquartis dos maiores salários oferecidos ao 1º trabalho após a conclusão dos cursos A e D (são semelhantes) são superiores às amplitudes interquartis dos maiores salários oferecidos ao 1º trabalho após a conclusão dos cursos B e C (que são semelhantes).

A distribuição dos maiores salários oferecidos após a conclusão do curso A é quase simétrica (maior concentração nos valores centrais).

A distribuição dos maiores salários oferecidos ao 1º trabalho após a conclusão do curso D é quase simétrica (maior concentração nos valores centrais).

A distribuição dos maiores salários oferecidos ao 1º trabalho após a conclusão do curso B é assimétrica à esquerda (maior concentração nos maiores valores).

A distribuição dos maiores salários oferecidos ao 1º trabalho após a conclusão do curso C é assimétrica à direita (maior concentração nos menores valores).

### Exercício 5:

Uma empresa pretende investigar a relação entre a produtividade e o tempo de formação (horas) dos seus funcionários. A empresa mede a produtividade dos funcionários pelo número de unidades produzidas por hora e regista o tempo que cada funcionário gasta em programas de formação durante um período de três meses. Os resultados obtidos foram os que seguintes:

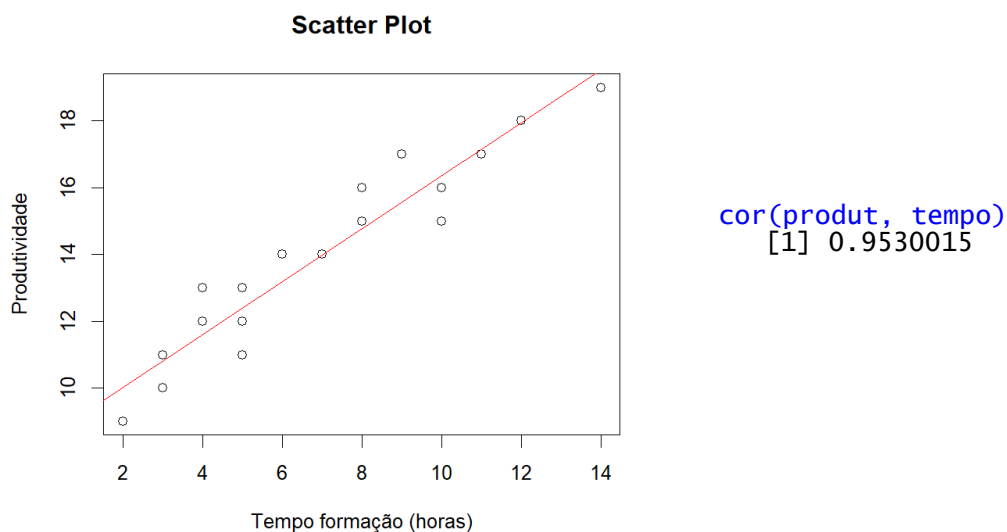
a) Identifique e classifique as variáveis.

Número de unidades produzidas/hora → quantitativa discreta

Tempo que cada funcionário gasta em programas de formação → quantitativa contínua

b) Analise o gráfico de dispersão e o valor do coeficiente de *Pearson*

Dados disponíveis no moodle, Caso Prático 3, no ficheiro Script CP3\_5



Existe associação linear positiva muito forte ( $0,9530015 \approx 1$ ) entre o tempo de formação e a produtividade (nº unidades produzidas/hora) dos funcionários.

Aos maiores tempos de formação correspondem, em média, maior produtividade.

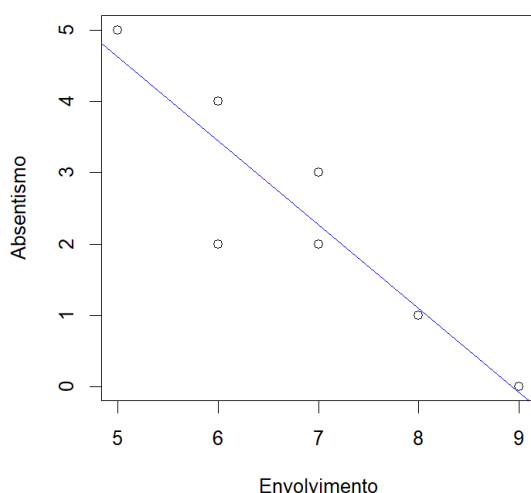


### Exercício 6:

Uma empresa pretende investigar a relação entre o envolvimento dos funcionários com a empresa e o absentismo ao trabalho. O envolvimento dos funcionários é medido numa escala ordinal (de 1=nada envolvido até 7=completamente envolvido) tendo em conta a motivação, o compromisso com a empresa, etc. O absentismo é registado através do número de dias que cada funcionário está ausente devido a doença ou motivos pessoais. Para o efeito, foram recolhidos dados das duas variáveis durante um período de seis meses e obtiveram-se os seguintes resultados.

Dados disponíveis no moodle, Caso Prático 3, no ficheiro Script CP3\_6

Gráfico dispersão



```
cor(A, E, method = "spearman")  
[1] -0.9518386
```

Como uma das variáveis é qualitativa ordinal → coeficiente de correlação de Spearman

Absentismo (número de dias que cada funcionário está ausente devido a doença ou motivos pessoais) → quantitativa discreta

Envolvimento dos funcionários (de 1=nada envolvido até 7=completamente envolvido) → qualitativa ordinal

Existe uma associação linear forte negativa (-0,9518386) entre o número de dias que cada funcionário está ausente do trabalho e o seu envolvimento com a empresa. A maior grau de envolvimento corresponde menor absentismo.

### Exercício 7:

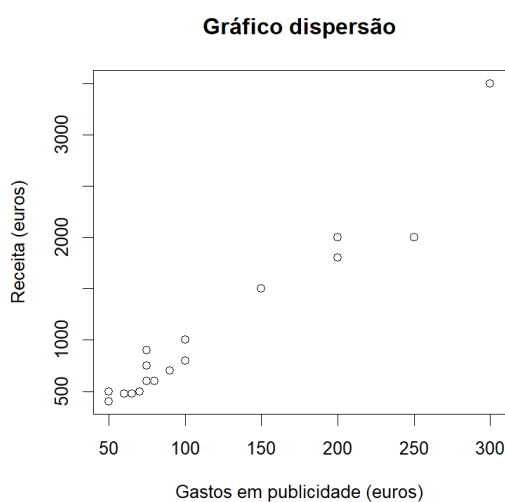
Foi realizado um estudo com o objectivo de analisar se existe associação linear entre os gastos das empresas com publicidade (em TV, anúncios online,...) e a receita das vendas. Utilize as ferramentas estatísticas adequadas para analisar o problema.

**Dados disponíveis no moodle, Caso Prático 3, no ficheiro Script CP3\_7**

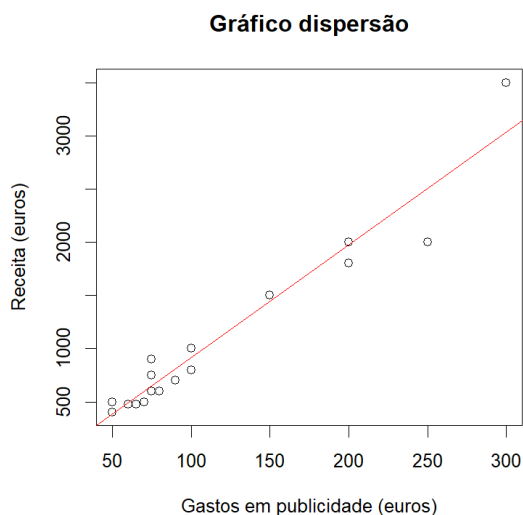
Como as variáveis são quantitativas → coeficiente de correlação linear de Pearson

```
cor(Gastos,Receita)
[1] 0.969452
```

```
plot(Gastos,Receita, main = "Gráfico dispersão", xlab = "Gastos em publicidade (euros)", ylab = "Receita (euros)")
```

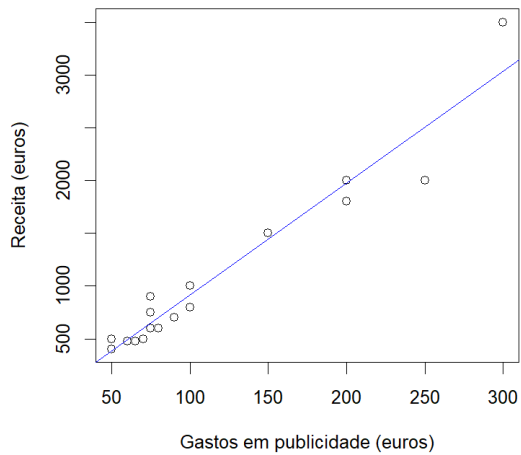


```
r1 <- lm(R~G)
abline(r1, col = "red")
```



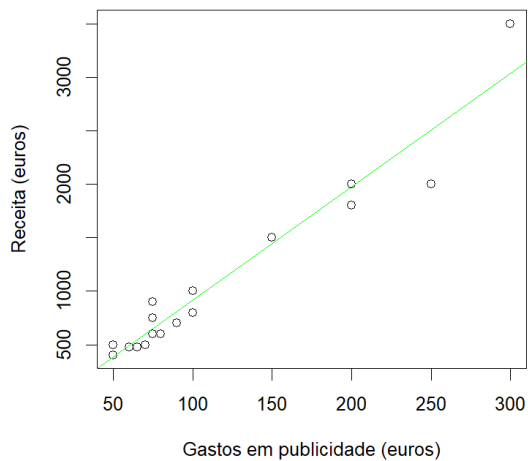
```
abline(r1, col = "blue")
```

**Gráfico dispersão**



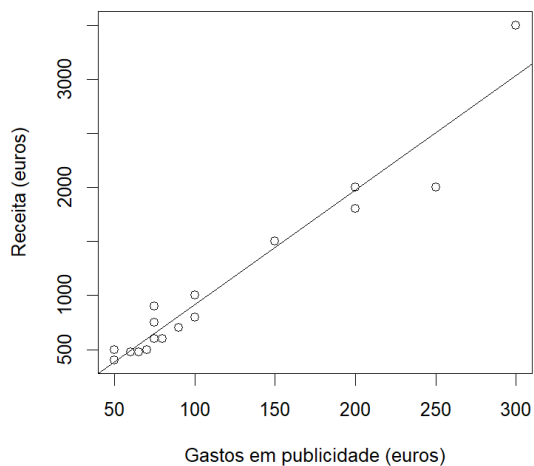
```
abline(r1, col = "green")
```

**Gráfico dispersão**



```
abline(r1)
```

**Gráfico dispersão**



Existe associação linear positiva muito forte (0,969452) entre os gastos em publicidade e a receita em vendas.

Aos maiores valores dos gastos em publicidade correspondem, em média, maiores valores das receitas.

### Exercício 8:

Uma empresa que comercializa um refrigerante pretende analisar a relação entre o consumo deste refrigerante (litros) e a idade do consumidor (anos). Analise o problema.

#### Dados disponíveis no moodle, Caso Prático3, ficheiro Script CP3\_8

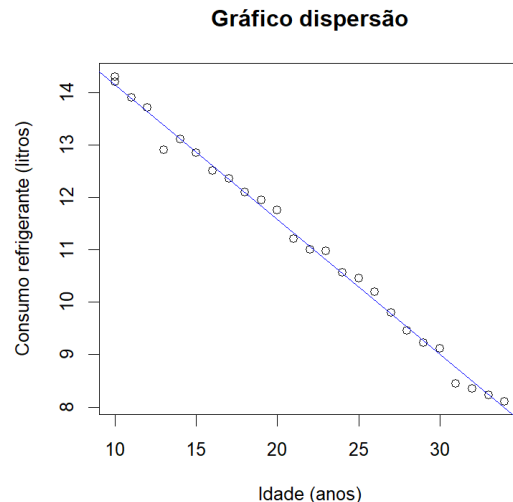
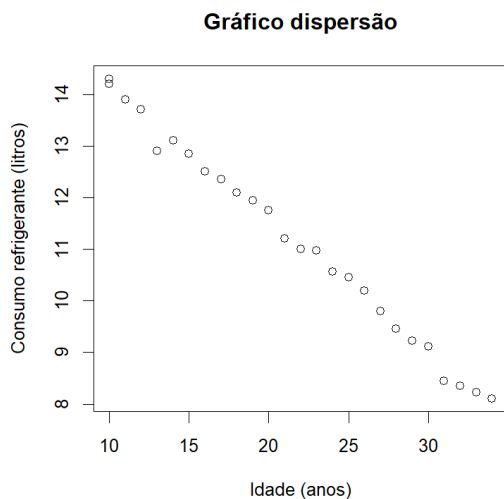
Como as variáveis Consumo de refrigerante e idade do consumidor são quantitativas (contínuas) → coeficiente de correlação linear de Pearson

Importar os dados do ficheiro **Caso Prático3**, separador **Exerc8**.

```
library(readxl)
> CP3 <- read_excel("C:/...caminho... /CP3.xlsx", sheet = "Exerc8")
> View(CP3)
```

```
cor(CP3$Idade, CP3$Consumo)
[1] -0.997104
```

```
plot(CP3$Idade, CP3$Consumo, main = "Gráfico dispersão", xlab = "Idade (anos)",
ylab = "Consumo refrigerante (litros)")
```



```
regr <- lm(CP3$Consumo~CP3$Idade)
> abline(regr, col = "blue")
```

Existe associação linear negativa muito forte entre a idade do consumidor e a quantidade de refrigerante consumida. A maiores consumos de refrigerante correspondem, em média, indivíduos mais novos.