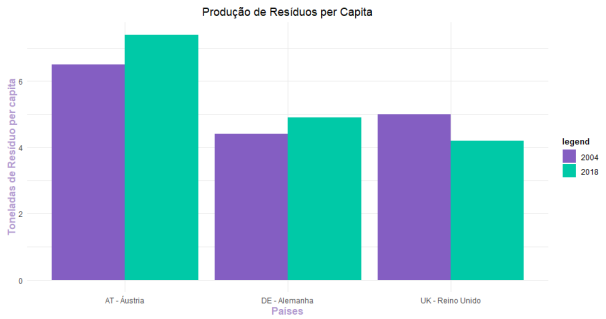
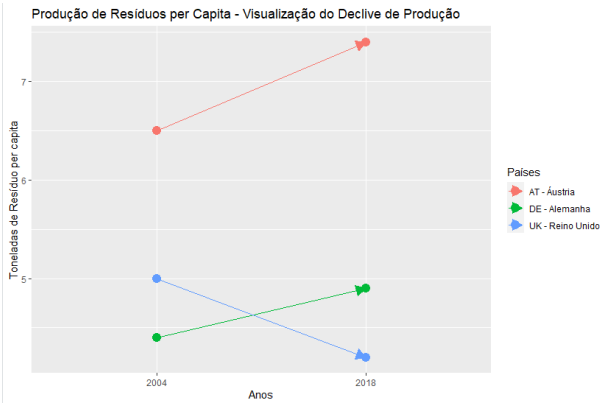
```r
library(tidyr)
library(readxl)
library(ggplot2)

#Define a diretoria com a localização do ficheiro a ser lido
setwd("C:\\Users\\TERESA NOGUEIRA\\Desktop\\R_excel_files\\1_exercise")

#Obtem os dados do ficheiro excel
data = read_excel("ResiduosPerCapita.xlsx",range = cell_rows(c(12:43)))

#Tratamento de dados -> compacta e retira dados desnecessários, insere a matriz resultante numa
    data frame
data <- data[-c(1,4:30),]
data.df <- as.data.frame(data)
colnames(data.df) <- c("Países", "2004", "2018")

final.df <- gather(data.df,"Anos", "Toneladas de Resíduo per capita",2:3)

#Gera o grafico de barras
ggplot(final.df, aes(x = Países, y =`Toneladas de Resíduo per capita`, fill = Anos)) +
        geom_bar(stat = "identity", position = position_dodge()) +
        labs(title = "Produção de Resíduos per Capita") +

#Tratamento estético do gráfico
        scale_fill_manual("legend", values = c("2004" = "#845EC2", "2018" = "#00C9A7")) +
        theme_minimal() +
        theme(plot.title = element_text(hjust = 0.5),
        axis.title.x = element_text(face="bold", colour="#B39CD0", size = 12),
        axis.title.y = element_text(face="bold", colour="#B39CD0", size = 12),
        legend.title = element_text(face="bold", size = 10))


#Gráfico extra para visualização do declive de Produção
ggplot(final.df, aes(x = Anos, y =`Toneladas de Resíduo per capita`, color = Países)) +
      geom_point(size = 4) +
      geom_line(aes(group=interaction(Países)),arrow = arrow(length=unit(0.40,"cm"),type = "
    closed")) +
      labs(title = "Produção de Resíduos per Capita - Visualização do Declive de Produção")
```

## Pergunta 1




### GRÁFICO DE BARRAS

O anterior gráfico de barras pretende evidenciar a comparação entre valores de diferentes subgrupos de dados, nomeadamente a produção de resíduos per capita de 3 países entre 2014 e 2018. Da avaliação direta do gráfico verificamos que a produção aumentou para a Áustria e para Alemanha e diminuiu para o Reino Unido.

### DECLIVE DE PRODUÇÃO

Por sua vez, o gráfico de declives ajuda apenas a tornar óbvio o que já é visualizado no anterior, sabemos que a produção em ambos os anos é mais elevada na Áustria (superior a 6 toneladas) e é onde se verifica o crescimento mais acentuado (maior declive visualizado). Os restantes países permanecem no mesmo intervalo de produção para ambos os anos, $[\,0,\,5]$ toneladas, apresentando flutuações de produção semelhantes, mas contrárias, dado que a produção de um aumenta e a do outro diminui.

```r
library(tidyr)
library(readxl)
library(ggplot2)

#Define a diretoria com a localização do ficheiro a ser lido
setwd("C:\\Users\\TERESA NOGUEIRA\\Desktop\\R_excel_files\\2_exercise")

#Obtém os dados do ficheiro excel
data = read_excel("EsperancaVida.xlsx",range = cell_rows(c(9:70)))

#Tratamento de dados -> compacta e retira dados desnecessários, insere a matriz resultante numa
    data frame
data <- data[-c(1:42),-c(2:42, 44:52, 54:65, 67:76, 78:86, 88:99, 101:103)]

data.df <- as.data.frame(data)

colnames(data.df) <- c("Anos","CY - Chipre ♂","IS - Islandia ♂","HU - Hungria ♂",
                       "CY - Chipre ♀","IS - Islandia ♀","HU - Hungria ♀")

final.df <- gather(data.df,"Países","Esperança de Vida",2:7)

#Constrói o gráfico temporal
ggplot(final.df, aes(x=Anos, y=`Esperança de Vida`, color = Países)) +
      geom_point(shape = 18, colour = "#B39CD0", size = 4, stroke = 2) +
      geom_line(size = 1) +
      xlab("Anos")+
      theme_minimal() +
      scale_color_manual(values=c("#D65DB1", "#9270D3","#007ED9",
                          "#0082C1","#007F93", "#FFA967"))
```

## Pergunta 2

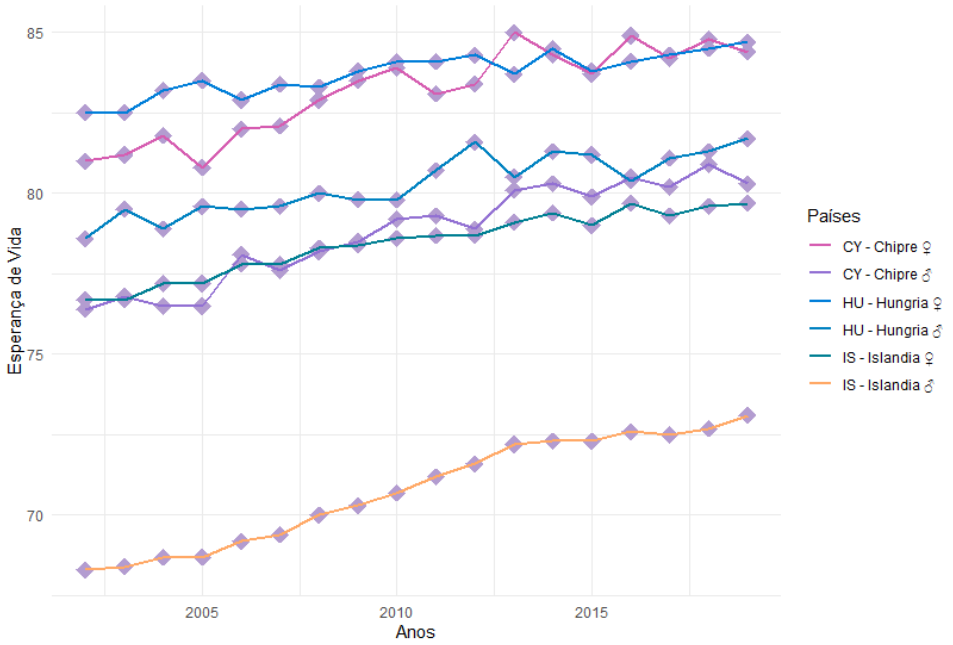



**GRÁFICO TEMPORAL**

O anterior gráfico temporal pretende analisar a evolução da esperança de vida dos homens e das mulheres da população de 3 países diferentes. A partir de uma análise meramente visual concluímos que a tendência de crescimento é positiva nos 3 países, que o género feminino tem uma expetativa de vida superior ao masculino e que a Islândia apresenta o conjunto de valores mais baixos de vida útil dos 3, especialmente para os homens, oposto a este país, evidencio o Chipre com uma expetativa de vida de 84.4 anos para as mulheres.

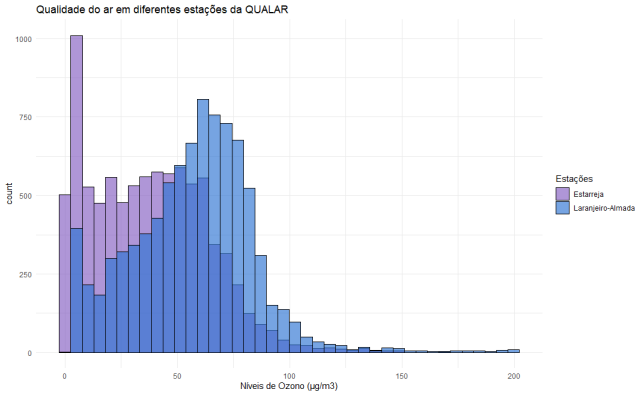
Posso ainda verificar que não existem outliers nem shifts súbitos nos dados (mantém a trend linear) e que não apresentam comportamentos cíclicos nem sazonais, o esperado já que a esperança de vida é maioritariamente dependente (do desenvolvimento) do país (um outro fator mais evidente, especialmente no gráfico em questão é o género).

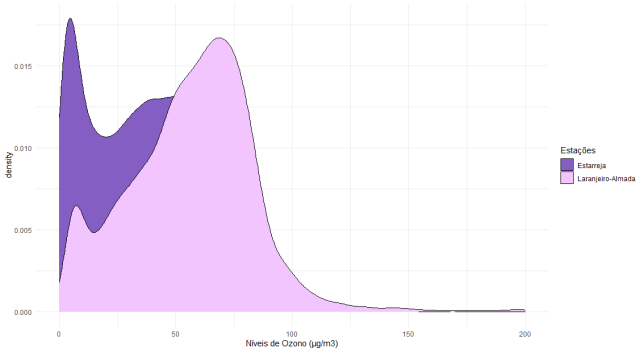
**"Where we live seems to influence how long we might live."[1]**

[1] https://www.disabled-world.com/fitness/longevity/

```r
library(tidyr)
library(readxl)
library(ggplot2)

#Define a diretoria com a localização do ficheiro a ser lido
setwd("C:\\Users\\TERESA NOGUEIRA\\Desktop\\R_excel_files\\3_exercise")

#Obtém os dados do ficheiro excel
data = read_excel("QualidadeAR03.xlsx",range = cell_cols("C:E"), col_types = c("text","text","text"))

#Tratamento de dados -> compacta, retira dados desnecessários e converte a matriz para numérico, insere a matriz resultante numa data frame
data <-data[,-2]

data_numeric = apply(as.matrix.noquote(data),2,as.numeric)

final.df = as.data.frame(data_numeric)
final.df <- gather(final.df,Estações,Value,1:2)


#Constrói o histograma
ggplot(final.df, aes(x = Value, fill = Estações)) +
      geom_histogram(color = 1, alpha = 0.65, position = "identity", bins = 40, ) +
      scale_fill_manual(values = c("#845EC2", "#2C73D2")) +
      xlab("Níveis de Ozono (µg/m3)") +
      ggtitle("Qualidade do ar em diferentes estações da QUALAR") +
      theme_minimal()

#Constrói o plot de densidade
ggplot(final.df, aes(x=Value, group=Estações, fill=Estações)) +
      geom_density(adjust=1.5) +
      scale_fill_manual(values = c("#845EC2", "#F3C5FF")) +
      xlab("Níveis de Ozono (µg/m3)") +
      theme_minimal()
```

## Pergunta 3




### HISTOGRAMA

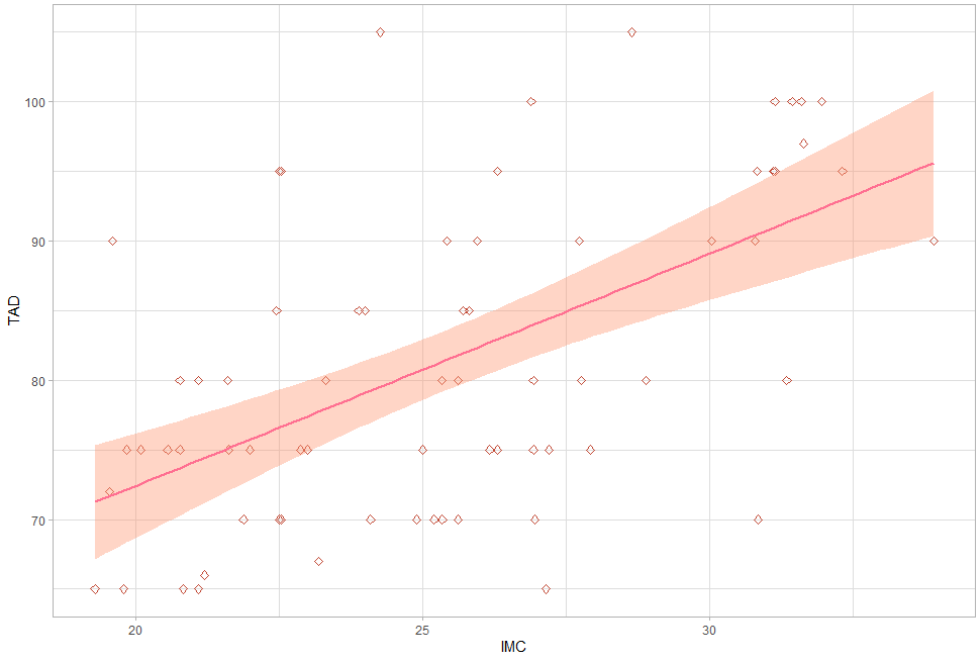
Por análise direta do histograma relativo à qualidade de ar de duas estações de QUALAR podemos numa primeira observação admitir que ambas possuem distribuições assimétricas, o expectável, tendo em conta que a variável em observação está dependente de um conjunto de fatores relativamente versáteis (nomeadamente, o meio geográfico, o paradigma social, disposição arquitetónica do local, etc). Numa segunda observação verificamos que a mancha de concentração dos níveis de ozono para a Estarrajeira encontra-se especialmente no intervalo [0, 50] µg/m3 e a de Laranjeira-Almada em [50,100] µg/m3. É então de fácil afirmação dizer que a Estarrajeira será a menos poluída do duo. De relevância, aponto apenas o pico de concentração perto de 5 µg/m3

### GRÁFICO DE DENSIDADE

O respetivo gráfico de densidade corrobora a afirmação anterior e garante uma melhor visualização da mancha de concetração (através da inspeção da curva contínua das amostras), de facto, Laranjeira-Almada aparenta possuir resultados mais drásticos.

```r
library(tidyr)
library(readxl)
library(ggplot2)

#Define a diretoria com a localização do ficheiro a ser lido
setwd("C:\\Users\\TERESA NOGUEIRA\\Desktop\\R_excel_files\\4_exercise")

#Obtém os dados do ficheiro excel
data = read_excel("Utentes.xlsx", range = cell_cols("C:D"))

final.df = as.data.frame(data)

#Cálculo da covariância e do coeficiente de correlação linear para comentários posteriores
IMC = final.df$IMC
TAD = final.df$TAD
cov(IMC, TAD)
cor(IMC,TAD)

#Constrói o Gráfico de Dispersão
ggplot(final.df, aes(x = IMC, y = TAD)) +
      geom_point(size = 2, shape = 23, colour = "#C34A36")  +
      geom_smooth(formula = y ~ x, method = "lm", colour = "#FF6F91", fill="#FF9671") +
      theme_minimal()
```

## Pergunta 4




| Covariância $[s_{xy}]$ | Coeficiente de Correlação Linear $[r_{xy}]$ |
|---|---|
| 24.94318 | 0.5662385 |

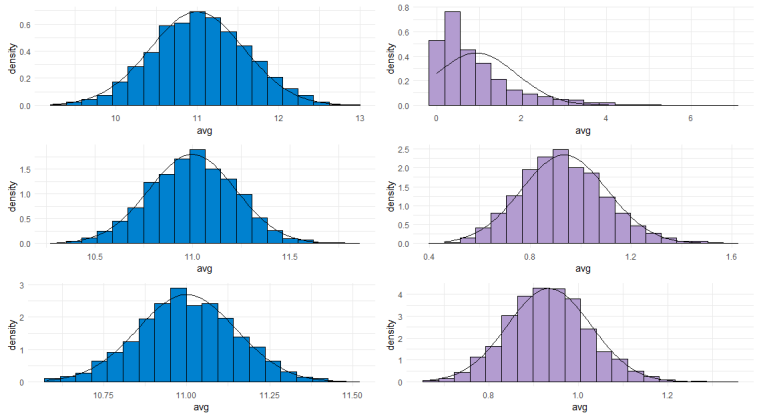
Por análise direta do gráfico, rapidamente concluímos existir uma relação linear crescente entre o aumento do índice de massa corporal - IMC e a Tensão Arterial Diastólica - TAD. Podemos até mesmo invocar a seguinte citação, **"Body mass index (BMI) is positively associated with both systolic blood pressure (SBP) and diastolic blood pressure (DBP)." [2018,Linderman]**[2] para corroborar os resultados obtidos.

Por outro lado, a inspeção da covariância e do coeficiente de Correlação linear facilmente nos indicariam a relação expetável entre os dois conjuntos de dados: $s_{xy} > 0 \rightarrow s_{xy} = 24.94318$ logo era esperada uma associação linear positiva entre conjuntos, ainda, $r \approx 0.5662385 \approx 1$ portanto era expectável visionar uma distribuição de pontos próxima de uma resta com declive positivo, ambas estas observações são corroboradas pelo gráfico em questão.

[2] George C. Linderman, B. S. (2018, August 17). Association of Body Mass index with blood pressure among 1.7 million Chinese adults. JAMA Network Open. Retrieved Junho 6, 2022, de https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2696872

## Pergunta 6

```r
library(tidyr)
library(cowplot)
library(readxl)
library(ggplot2)
#---dados do problema---
set.seed(950)
a = 9
b = 13
n_samples = 1970
#Produz o vetor da média de amostras para cada n
vector_maker <- function(n){
  avg = rep(0,n_samples)
  samples = rep(0,n)

  for(i in 1:n_samples){
    samples = runif(n, a, b)
    avg[i] = mean(samples)
  }
  avg.df = as.data.frame(avg)
  return(avg.df)
}
#Produz os gráficos (o gráfico da distribuição exponencial não consta
# no código por simplicidade, é simplesmente uma avaliação extra)
plot_maker <- function(avg.df){
  mean_of_dist = (a + b)/2
  variance_of_dist = (b - a)^2/12
  Vn = variance_of_dist/n

  plot1 <- ggplot(avg.df, aes(x = avg)) +
          geom_histogram(aes(y = ..density..),
          colour = 1,
          fill = "#0081CF", bins = 15) +
          stat_function(fun = dnorm,
          args = list(mean = mean_of_dist,
          sd = sqrt(Vn))) +
          theme_minimal()

return(plot1)
}

n = 4
avg_1.df <- vector_maker(n)
plotFreqDen1 <- plot_maker(avg_1.df)

n = 27
avg_2.df <- vector_maker(n)
plotFreqDen2 <- plot_maker(avg_2.df)

n = 61
avg_3.df <- vector_maker(n)
plotFreqDen3 <- plot_maker(avg_3.df)

plot_grid(plotFreqDen1, plotFreqDen2,
    plotFreqDen3, nrow = 3, ncol = 1)
```




### COMPORTAMENTO DAS DISTRIBUIÇÕES UNIFORME E EXPONENCIAL

Na lateral encontram-se 6 gráficos correspondentes à distribuição da média de duas populações, uma de distribuição normal e outra de distribuição exponencial, para diferentes valores de n (4, 27 e 61; 1, 30 e 100, respetivamente). A observação direta e a natureza do conjunto ilícita um análise do teorema do limite Central:

Invocando o teorema do limite central, sabemos que a média de uma amostra de dados estará mais próxima da média da população geral em questão, à medida que o tamanho da amostra aumentar, independentemente da distribuição real dos dados. Afirma assim que com a extensão do tamanho da amostra, a distribuição da sua média aproxima-se de uma distribuição normal.

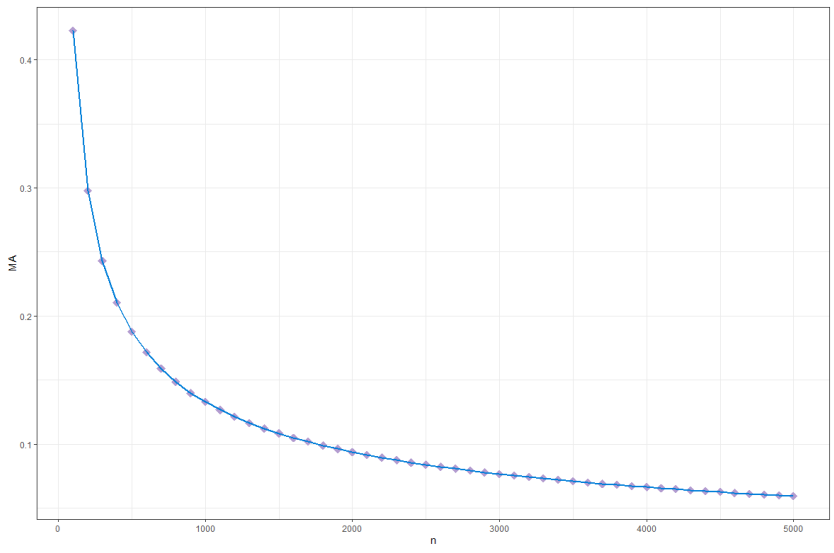
É exatamente isto que verificamos para os 6 conjuntos de amostras: para os primeiros 3, cuja população em análise é dotada de uma distribuição uniforme, é esperado verificar já a partir de n = 4 uma boa sobreposição entre a curva normal teórica e as probabilidades obtidas, verificando-se apenas um afunilamento do gráfico com o aumento do tamanho da amostra (tal era esperado já que se a distribuição subjacente for simétrica, não precisamos de um tamanho de amostra muito grande para atingir a distribuição normal), basta analisar o eixo das ordenadas para verificar o aumento do concentração perto do valor esperado da média, $E[X] = \frac{a+b}{2} = \frac{13+9}{2} = 11.0$. Para o segundo conjunto de 3 gráfico, o mesmo fenómeno verifica-se, pesa embora apenas a partir de n = 30 (a distribuição não possui simetria), onde passará a aproximar-se cada vez mais da média amostral teórica, (para $\lambda = 1.07$, $E[X] = 1\lambda \approx 0.93$).

**Nota:** Esta concentração/afunilamento é expectável já que a variância é inversamente proporcional ao tamanho da amostra $\rightarrow Var[\overline{X}] = \frac{\sigma^2}{n}$ e $Var[\overline{X}] = \frac{1}{\lambda^2 n}$ para cada uma das respetivas populações.

**Dados do problema** $\rightarrow \lambda = 1.17$, $m = 1350$, seed $= 269$

```
library(tidyr)
library(readxl)
library(ggplot2)
#---dados do problema---
set.seed(269)
m = 1350
lambda = 1.07
#-----------------------
n = rep(0,50)
n[1] = 100
sample_number = 100
i = 2
#Cálcula o vetor com o tamanho das samples
while (sample_number < 5000){
  n[i] = n[1] + sample_number
  sample_number = sample_number + 100
  i = i + 1
}
#Cálcula o vetor de intervalos para cada n e de a sua média
intervalo_maker <- function(n){
  x = 0
  for (i in 1:m){
    x[i] = mean(rexp(n,lambda))
  }
  gama = 0.95
  a = qnorm((1+gama)/2)
  amplitude = (2*a)/(x*sqrt(n))
  result = mean(amplitude)
  return(result)
}
#Produz o vetor de médias da amplitude
z = 1
MA = rep(0,50)
for(i in 1:50){
  MA[z] = intervalo_maker(n[i])
  z = z + 1
}

data.df <- data.frame(n, MA)

ggplot(data.df, aes(x=n, y = MA)) +
  geom_point(shape = 18, colour = "#B39CD0", size = 3, stroke = 2) +
  geom_line(size = 1, colour = "#007ED9") +
  theme_bw()
```

## Pergunta 9



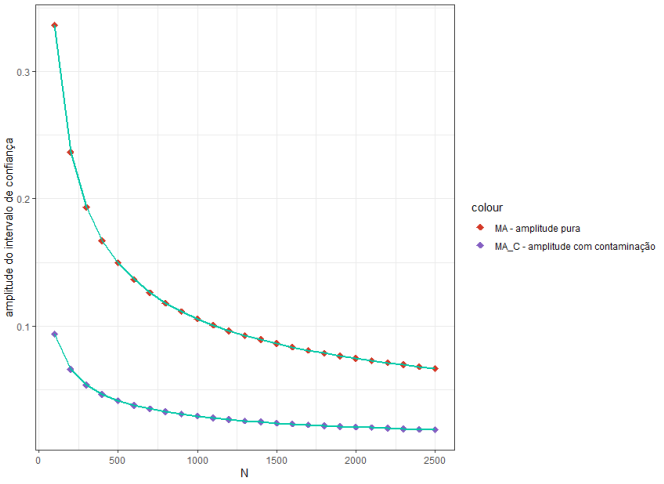

**MÉDIA DAS AMPLITUDES DOS INTERVALOS DE CONFIANÇA**

Por observação direta da progressão da amplitude de confiança em função da dimensão das amostras, admitimos que esta sofre um declínio com o aumento da extensão amostral. Tal é expectável uma vez que para obter a média das amplitudes utilizamos a seguinte fórmula: $MA = \frac{2a}{\overline{x}\sqrt{n}}$ onde $\overline{x}$ é a média da amostra gerada, $a = \frac{1+\gamma}{2}$ m é o tamanho do vetor final obtido de x e subsequentemente da própria amplitude.

É então esperado um decaimento na ordem dos $\frac{1}{\sqrt{n}}$ valor este suportado pela regressão calculada.

## Pergunta 10: Dados do problema → $\lambda = 0.92$, $\lambda_c = 0.08$, $m = 1050$, seed = 269

```
library(tidyr)
library(readxl)
library(ggplot2)
#----dados do problema----
set.seed(269)
m = 1050
lambda = 0.92
lambda_c = 0.08
#-------------------------
n = rep(0,25)
n[1] = 100
sample_number = 100

#declara o vetor das dimensões
i = 2
while (sample_number < 2500){
  n[i] = n[1] + sample_number
  sample_number = sample_number + 100
  i = i + 1
}

#provoca a contaminação do vetor puro
cont <- function(x_pure, x_other, n){
  for(i in 1:floor(n*0.25)){
    x_pure[i] = x_other[i]
  }
  return(x_pure)
}

#Cálcula as amplitudes dos intervalos de
    confiança
intervalo_maker <- function(n, x, lambda){
  gama = 0.93
  a = qnorm((1+gama)/2)
  amplitude = (2*a)/(x*sqrt(n))
  result = mean(amplitude)
  return(result)
}

#Contrói os vetores de MA e MA_C
x_pure = rep(0,m)
x_other = rep(0,m)
x_contaminated = rep(0,m)
MA = rep(0,25)
MA_C = rep(0,25)
z = 1
for(i in 1:25){
  for (c in 1:m){
    x = rexp(n[i],lambda)
    y = rexp(n[i],lambda_c)
    j = cont(x,y, n[i])

    x_pure[c] = mean(x)
    x_contaminated[c] = mean(j)
  }

  MA[z] = intervalo_maker(n[i], x_pure, lambda)
  MA_C[z] = intervalo_maker(n[i],
            x_contaminated, lambda_c)
  z = z + 1

}

#Constrói o gráfico das regressões lineares
data.df <- data.frame(n, MA,MA_C)

ggplot(data.df) + geom_point(aes(x = n, y = MA,
    color = "MA - amplitude pura"),
    shape = 18, size = 2, stroke = 2) +
    geom_point(aes(x = n, y = MA_C,
    color ="MA_C -amplitude com contaminação"),
    shape = 18, size = 2, stroke = 2) +
    scale_color_manual (
    values =c('#D43725','#845EC2')) +
    geom_line(aes(x = n, y = MA),
    size = 1, colour = "#00C9A7") +
    geom_line(aes(x = n, y = MA_C),
    size = 1, colour = "#00C9A7") +
    xlab(("N")) +
    ylab("amplitude do intervalo de confiança")
    + theme_bw()
```




### MÉDIA DAS AMPLITUDES DOS INTERVALOS DE CONFIANÇA COM E SEM CONTAMINAÇÃO

De forma semelhante ao exercício 9, é observado um declínio da média dos intervalos de confiança com a extensão do tamanho das amostras, nomeadamente na ordem de $\frac{1}{\sqrt{n}}$, já que (novamente) admitimos que $MA = \frac{2a}{\overline{x}\sqrt{n}}$.

De se fazer notar a discrepância entre a regressão pura e a regressão contaminada, fruto da aplicação de um $\lambda_c$ ($\lambda_c = 0.08 < \lambda = 0.92$ menor com contaminação da amostra inicial de 25%.

**Nota**: a regressão contaminada é menor do que a pura graças à aplicação de um *failure rate* menor, já como foi explicitado no parágrafo supra, tal deve-se ao próprio modelo da distribuição: $f(x) = \lambda \cdot e^{-\lambda \cdot x}$. Para o modelo contaminado teremos uma média de amostras dopada, $E[\overline{X}] = \frac{1}{\lambda} = \frac{1}{0.75\lambda_{puro}+0.25\lambda_c} > \frac{1}{\lambda_{puro}}$ que consequentemente provocará um decréscimo das amplitudes calculadas.