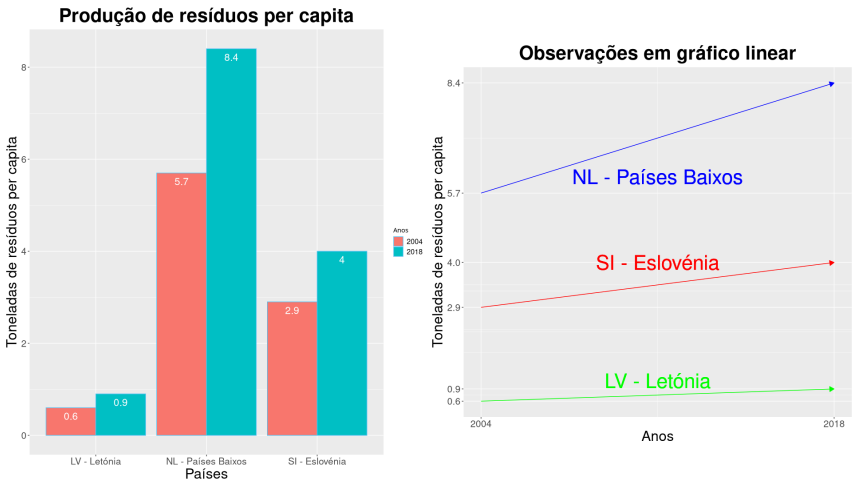
Pergunta 1

```r
pacman::p_load(pacman, rio, tidyverse, ggplot2, cowplot) # Load req. packages
## Dados ---------------------------------------------------------------------
residuos <- import("~/Projects/git/personal/PEst/Code_env/ResiduosPerCapita.xlsx")
req <- c("NL - Países Baixos", "SI - Eslovénia", "LV - Letónia")
temp <- filter(residuos, grepl(paste(req, collapse="|"), residuos$...1))
result <- temp[, c(1,2,3)] # Selecionar as colunas
aux <- data.frame(
  Anos=c(2004,2018),
  NL=as.numeric(c(result[1,2:3])),
  SI=as.numeric(c(result[2,2:3])),
  LV=as.numeric(c(result[3,2:3])),
  stringsAsFactors = FALSE
)
names(result)[names(result) == "...1"] <- "Países"
names(result)[names(result) == "...3"] <- "2018"
names(result)[names(result) == "Produção de resíduos per capita"] <- "2004"
# Formatar os dados para o gráfico de barras
data <- gather(result, "Anos", "Toneladas de resíduos per capita", 2:3)
data[3] = as.numeric(c(data[[3]]))
## Gráfico de barras ---------------------------------------------------------
p1 <- ggplot(data=data, aes(x=Países, y=`Toneladas de resíduos per capita`, fill=Anos)) +
      geom_bar(stat="identity", position=position_dodge()) +
      ggtitle("Produção de resíduos per capita") +
      theme(plot.title=element_text(hjust=0.5, size=32, face="bold"),
            axis.title=element_text(size=24),
            axis.text=element_text(size=16),
            legend.title=element_text((size=16)),
            legend.text=element_text(size=12)) +
      geom_text(aes(label=`Toneladas de resíduos per capita`), vjust=1.6, color="white",
                position=position_dodge(0.9), size=6)
## Gráfico auxiliar ----------------------------------------------------------
p2 <- ggplot(data=aux, aes(x=Anos), fill=NULL) + ggtitle("Observações em gráfico linear") +
      theme(plot.title=element_text(hjust=0.5, size=32, face="bold"),
            axis.title=element_text(size=24),
            axis.text=element_text(size=16),
            aspect.ratio=0.9) +
      scale_x_continuous(breaks=as.numeric(unlist(aux[,1]))) +
      scale_y_continuous(breaks=as.numeric(unlist(aux[,2:4]))) +
      geom_line(aes(y=NL), color="red", arrow = arrow(length=unit(0.30,"cm"), type = "closed")) +
      geom_line(aes(y=SI), color="green", arrow = arrow(length=unit(0.30,"cm"), type = "closed")) +
      geom_line(aes(y=LV), color="blue", arrow = arrow(length=unit(0.30,"cm"), type = "closed")) +
      geom_text(aes(x=2011, y=4.0), label="SI - Eslovénia", color="red", size=12) +
      geom_text(aes(x=2011, y=1.1), label="LV - Letónia", color="green", size=12) +
      geom_text(aes(x=2011, y=6.1), label="NL - Países Baixos", color="blue", size=12) +
      labs(x="Anos", y="Toneladas de resíduos per capita")
## Plot do gráfico de barras e gráfico auxiliar ------------------------------
plot <- plot_grid(p1, p2, nrow=1, ncol= 2, labels=NULL); plot
## Média de todos os países --------------------------------------------------
mean04 = mean(as.numeric(residuos$`Produção de resíduos per capita`[7:36]))
mean18 = mean(as.numeric(residuos$...3[7:36]))
sprintf("Média prod. resíduos em 2004: %.2f", mean04)
sprintf("Média prod. resíduos em 2018: %.2f", mean18)
```

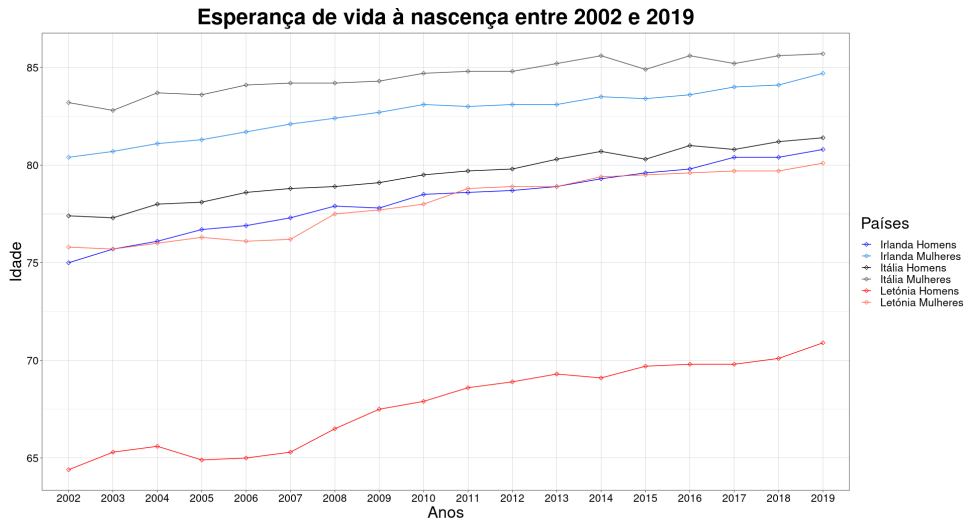


Comentário: Por observação direta do diagrama de barras e do gráfico linear dos três países europeus no ano 2004 e 2018, é necessáriamente possível concluir que ocorreu um aumento de produção de resíduos nos três países. Salienta-se que os Países Baixos foram o maior produtor de resíduos (com o maior acrescimo também), enquanto a Letónia o menor.

Naturalmente, foi calculada a média da produção de resíduos de todos os países nos anos 2004 e 2018 por efeitos comentaristas. Os valores obtidos foram 6.14 e 6.21 toneladas de resíduos per capita, respetivamente para 2004 e 2018. Note-se que a Letónia se encontra tremendamente abaixo destes valores, e que em 2004 o grupo dos três países encontrava-se abaixo da média. Em 2018, os Países Baixos tiveram uma produção acima da média.

Pergunta 2

```
pacman::p_load(pacman, rio, tidyverse, ggplot2, hrbrthemes) # Load req. packages
## Dados ----------------------------------------------------------------------
esperanca_vida <- import("~/Projects/git/personal/PEst/Code_env/EsperancaVida.xlsx")
result <- esperanca_vida[c(48:65), c(1, 54, 55, 56, 88, 89, 90)]
result$...54 <- as.numeric(esperanca_vida$...54[48:65])
result$...55 <- as.numeric(esperanca_vida$...55[48:65])
result$...56 <- as.numeric(esperanca_vida$...56[48:65])
result$...88 <- as.numeric(esperanca_vida$...88[48:65])
result$...89 <- as.numeric(esperanca_vida$...89[48:65])
result$...90 <- as.numeric(esperanca_vida$...90[48:65])
# Rename das colunas
names(result)[names(result) == "...1"] <- "Anos"
names(result)[names(result) == "...56"] <- "Letónia Homens"
names(result)[names(result) == "...55"] <- "Itália Homens"
names(result)[names(result) == "...54"] <- "Irlanda Homens"
names(result)[names(result) == "...90"] <- "Letónia Mulheres"
names(result)[names(result) == "...89"] <- "Itália Mulheres"
names(result)[names(result) == "...88"] <- "Irlanda Mulheres"
gtemporal <- gather(result, "Países", "Esperança de vida à nascença",2:7)
## Gráfico temporal -----------------------------------------------------------
ggplot(gtemporal, aes(x=`Anos`, y=`Esperança de vida à nascença`, color=`Países`,
                      group =`Países`)) +
  geom_point(size=2, shape=23) +
  geom_line() +
  labs(x="Anos", y="Idade") +
  scale_color_manual(values=c('blue', 'dodgerblue2', 'black', 'gray31', 'red',
                              'tomato1')) +
  labs(color='Países',fill='Países') +
  ggtitle("Esperança de vida à nascença entre 2002 e 2019")  +
  theme_linedraw() +
  theme(plot.title = element_text(hjust = 0.5, size=32, face="bold"),
        axis.title=element_text(size=24),
        axis.text=element_text(size=16),
        legend.title=element_text(size=24),
        legend.text=element_text(size=16))
```



Comentário: Independemente do género e do país, é natural concluir que a esperança média de vida tem aumentado ao longo dos anos, graças aos dados desencadeados. É verificável que a aparente trend linear se mantém e que não se apresentam comportamentos cíclicos nem sazonais - o fenómeno é expectável, visto que o país de nascença afeta estridulamente a esperança de vida[1].

Por observação direta do gráfico temporal apresentado em seguida, destaca-se também que a esperança média de vida à nascença é superior para as mulheres em relação à dos homens em todos os casos.

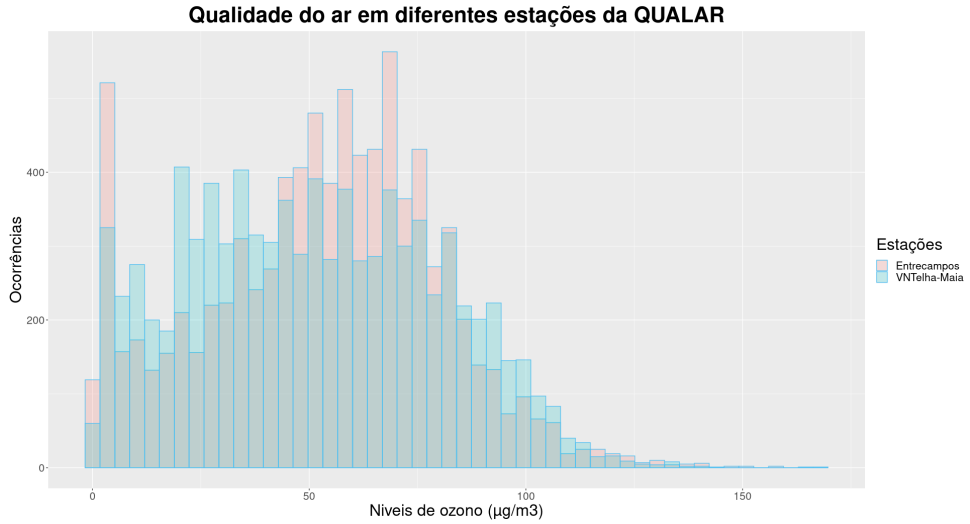
É também salientado que a Itália é o país com maior esperança média de vida do grupo analizado, seguido da Irlanda e, por final, pela Eslováquia (especialmente no caso dos homens, em que a diferença é aproximadamente 10 anos, em comparação com os outros dois países).

[1] Eric Neumayer and Thomas Plümper. Inequalities of income and inequalities of longevity: A cross-country study. American Journal of Public Health, 106(1):160–165, 2016. PMID: 26562120.

Pergunta 3

```r
pacman::p_load(pacman, rio, tidyverse, ggplot2, hrbrthemes) # Load req. packages
## Dados ----------------------------------------------------------------------
qualidade_ar <- import("~/Projects/git/personal/PEst/Code_env/QualidadeAR03.xlsx")
result <- qualidade_ar[, c(2,10)]
result$Entrecampos <- as.numeric(qualidade_ar$Entrecampos)
result$`VNTelha-Maia` <- as.numeric(qualidade_ar$`VNTelha-Maia`)
dataframe <- gather(result, "VNTelha-Maia", "Entrecampos",1:2)
## Histograma -----------------------------------------------------------------
ggplot(dataframe, aes(x = `Entrecampos`, fill = `VNTelha-Maia`)) +
  geom_histogram(position = "identity", alpha = 0.2, bins = 50) +
  labs(fill="Estações", x="Niveis de ozono (μg/m3)", y="Ocorrências") +
  ggtitle("Qualidade do ar em diferentes estações da QUALAR") +
  theme(plot.title = element_text(hjust = 0.5, size=32, face="bold"),
        axis.title=element_text(size=24),
        axis.text=element_text(size=16),
        legend.title=element_text(size=24),
        legend.text=element_text(size=16))

## Valor médio da qualidade do ar nas estações ---------------------------------
E_avg = mean(result$Entrecampos); print(E_avg)
V_avg = mean(result$`VNTelha-Maia`); print(V_avg)
```

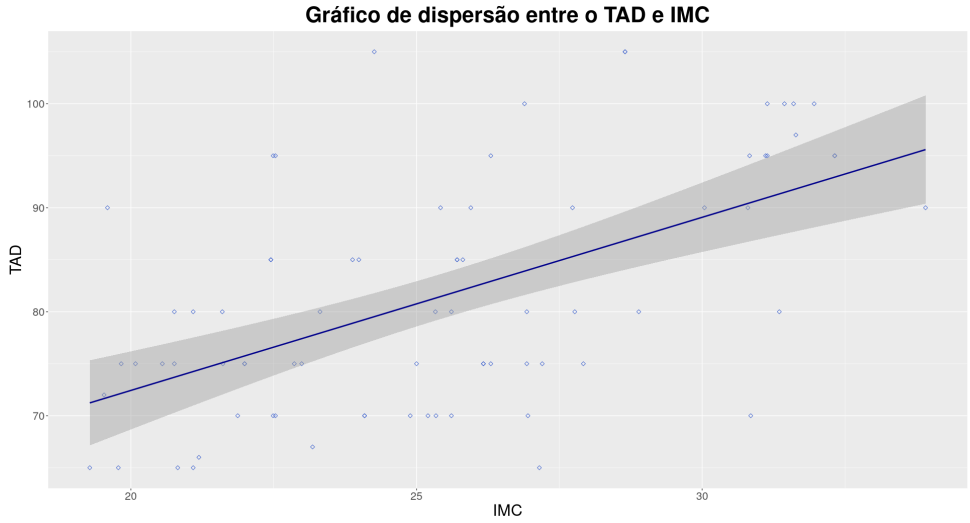


Comentário: A partir do histograma formulado através das observações horárias de níveis de ozono em microgramas por metro cúbico em duas estações específicas - Entrecampos e VNTelha-Maia - no ano de 2020, é possível obter a conclusão de que a estação VNTelha-Maia teve mais horas com níveis baixos de ozono (0 a 50 $\mu g/m^3$) em comparação à estação Entrecampos. No entanto, para níveis intermédios de ozono (50 a 100 $\mu g/m^3$), suscita-se o contrário, i.e., a estação Entrecampos tem um nível predominante de ocorrências. A níveis elevados de ozono (100 a 200 $\mu g/m^3$), ambas as estações apresentam um número de ocorrências bastante diminuto.

Com o observado, seria de esperar que a estação Entrecampos fosse a mais poluída. Tal é corroborado, ao analizar os níveis de poluição médias das duas estações ao longo do ano, visto que Entrecampos apresenta um valor de 52.7 $\mu g/m^3$ e VNTelha-Maia um valor de 50.9 $\mu g/m^3$.

Pergunta 4

```
pacman::p_load(pacman, rio, tidyverse, ggplot2, hrbrthemes) # Load req. packages
## Dados ----------------------------------------------------------------------
utentes <- import("~/Projects/git/personal/PEst/Code_env/Utentes.xlsx")
result <- utentes[, c(3,4)]
# CovariÃ¢ncia e coeficiente de correlação linear
IMC=result$IMC
TAD=result$TAD
cov=cov(IMC, TAD); sprintf("CovariÃ¢ncia: %.2f", cov)
corr=cor(IMC, TAD); sprintf("Coeficiente de correlação linear: %.2f", corr)
## Gráfico de dispersão -------------------------------------------------------
ggplot(result, aes(x=IMC, y=TAD)) +
  geom_point(size=2, shape=23, color = 'royalblue3') +
  ggtitle("Gráfico de dispersão entre o TAD e IMC")  +
  theme(plot.title=element_text(hjust=0.5, size=32, face="bold"),
        axis.title=element_text(size=24),
        axis.text=element_text(size=16))+
  geom_smooth(method=lm, color = 'darkblue')
```



Comentário: Note-se que apesar da aproximação linear, há uma fraca correlação entre o IMC e o TAD, tal pode ser observado gráficamente através dos diversos pontos dos dados recolhidos que não seguem este padrão. É então possível concluir que apesar de se associar[2]que um aumento do IMC leva a um aumento do TAD, os diversos pontos dispersos, provenientes dos dados recolhidos, suscitam uma relação que não é perfeitamente linear - sendo este o motivo pelo qual foi considerado que a correlação entre as duas variáveis é fraca/não ideal, por análise gráfica. A associação positiva entre os dados é amplamente estudada[1], e portanto, esperada.

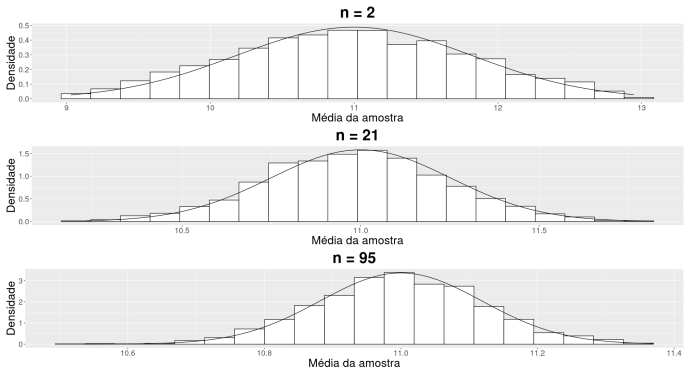
Covariância $[s_{xy}]$	**Coeficiente de correlação linear** $[r_{xy}]$
24.94	0.57

Assim, analizando a covariância e do coeficiente de correlação linear, é facilmente corroborado o visualizado, e expectável entre o binómio de conjuntos de dados. Visto que $s_{xy} > 0$, é natural a associação positiva observada (e documentada[1]); no entanto, verifica-se $r_{xy} \approx 0.57$, que ainda se encontra longe do valor ideal unitário.

[2]Dua S, Bhuker M, Sharma P, Dhall M, Kapoor S. Body mass index relates to blood pressure among adults. N Am J Med Sci. 2014 Feb;6(2):89-95. doi: 10.4103/1947-2714.127751. PMID: 24696830; PMCID: PMC3968571.

Pergunta 6

```r
pacman::p_load(pacman, rio, cowplot, tidyverse, datasets, ggplot2, dplyr,
               data.table, hrbrthemes) # Load req. packages
options(digits=15)
amostras=1940
a=9; b=13
E=(a+b)/2 # Valor teórico para o valor esperado
V=((b-a)^2)/12 # Valor teórico para a variÃ¢ncia
## Simulação ------------------------------------------------------------------
sim <- function(n){
  set.seed(731)
  Vn=V/n
  data=rep(0, amostras)
  data_avg=data.frame(rep(0, amostras))
  avg=rep(0, amostras)
  for(i in 1:amostras){
    data[i] = data.frame(X=runif(n, a, b))
    avg[i] = mean(data[[i]])
  }
  data_avg <- data.frame(avg)
  # Plot do gráfico
  plot <- ggplot(data_avg, aes(x = avg)) +
    geom_histogram(aes(y = ..density..), colour = 1, fill = "white", bins=20) +
    labs(x="Média da amostra", y="Densidade") +
    stat_function(fun = dnorm, args = list(mean = E, sd = sqrt(Vn))) +
    theme(plot.title=element_text(hjust=0.5, size=32, face="bold"),
          axis.title=element_text(size=24),
          axis.text=element_text(size=16)) + ggtitle(paste0("n = ", n))
  return(plot)
}
## Histogramas sobrepostos com as curvas com a distribuição normal -------------
plot_grid(sim(2), sim(21), sim(95), nrow=3, ncol=1)
```



Parâmetros: Seed: 731; Número de amostras: 1940; Uniforme no intervalo: $[9, 13]$

Comentário: Com base nas múltiplas simulações, onde foi calculada a média de uma distribuição uniforme (para três valores da dimensão de amostras), foi elaborado o histograma de densidades relativas apresentado, ao qual foi sobreposto uma curva com distribuição normal.

Visto que as variáveis aleatórias são independentes e identicamente distribuídas, estamos perante uma aplicação possível do Teorema do Limite Central, e deste modo é aceitável aproximar a distribuição uniforme a uma normal de valor esperado $E(\mathbb{X}) = \mu$ e variância $V_n(\mathbb{X}) = \sigma^2/n$. Salienta-se a aproximação vai melhorando com o aumento da dimensão das amostras, no entanto prova-se aceitável para os valores baixos requisitados (inferiores a 30).

Ao aumentar a dimensão de amostras é aparente a tendência de um estreitamento e maior semelhança com a curva com distribuição normal. Este comportamento é necessariamente esperado, visto que ocorre uma menor dispersão da média à medida que a dimensão de amostras aumenta - a variância é inversamente proporcional a este valor: $V_n(\mathbb{X}) = \sigma^2/n$.

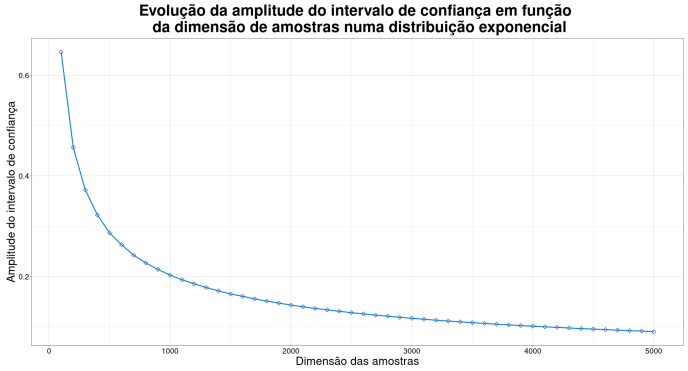
Então para amostras de maiores dimensões, a média das amostras da distribuição uniforme, aproxima-se, de facto, a uma distribuição normal centrada no valor esperado de cada amostra e com variância que diminui com o aumento da dimensão das amostras.

Pergunta 9

```r
pacman::p_load(pacman, rio, tidyverse, datasets, ggplot2, dplyr, data.table,
               hrbrthemes) # Load req. packages
set.seed(272)
amostras=850
size=5000/100
lambda=1.77
options(digits=15)
## Extremo da normal de lambda approx. ----------------------------------------
gama=0.93
a=qnorm((1+gama)/2)
q=(1-gama)/2

MA <- rep(0, size)
vec <- rep(0, size)
avec <- rep(0, amostras)

for(m in 1:50){
  n=m*100
  mle <- rep(0, amostras)
  # Simular a situação amostras vezes
  for (i in 1:amostras){
    mle[i] <- mean(rexp(n, lambda)) # Valor esperado (1/lambda)
    avec[i] <- (2*a)/(mle[i]*sqrt(n))
  }
  ## Amplitude
  MA[m]= mean(avec)
  vec[m] <- n # Array com no. de amostras utilizado
}
data = data.frame(vec, MA)
## Gráfico
ggplot(data, aes(x=vec, y=MA)) + theme_linedraw() +
  ggtitle("Evolução da amplitude do intervalo de confiança em função \n da dimensão de amostras numa distribuição
       exponencial") +
  geom_point(size=3, shape=23, color="dark blue") + geom_line(size = 1 , colour = "#007ED9" ) +
  labs(x="Dimensão das amostras", y="Amplitude do intervalo de confiança") +
  theme(plot.title=element_text(hjust=0.5, size=32, face="bold"),
        axis.title=element_text(size=24),
        axis.text=element_text(size=16))
```



Parâmetros: Semente: 272; Número de amostras (m): 850; $\lambda = 1.77$; Nível de confiança $(1-\alpha)$: 0.93

Comentário: Por observação direta da evolução da amplitude do intervalo de confiança em função da dimensão de amostras, conclui-se, que de facto ocorre um decaimento de amplitude à medida que a dimensão de amostras da distribuição exponencial aumenta. Tal é corroborado teoricamente, visto que seria expectável um decaimento na ordem de $1/\sqrt{n}$.

Dado que a distribuição exponencial apresenta amostras de dimensão superior a 30 e as variáveis aleatórias são independentes e identicamente distribuídas, o Teorema do Limite Central, é aplicável, e então $\bar{X} \sim N(\mu, \sigma^2/n)$. Como tal, a normal padrão é facilmente obtida: $Z = (\bar{X} - E(\bar{X}))/(\sqrt{V(\bar{X})/n})$.

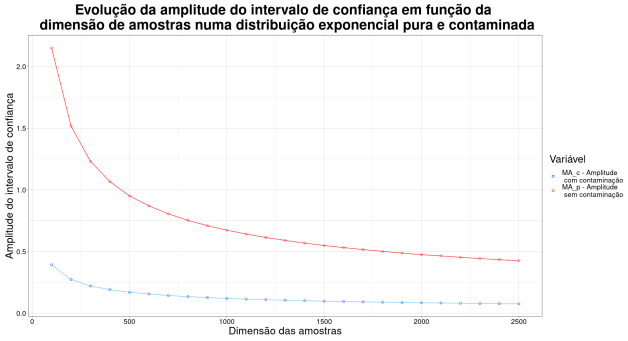
A amplitude do intervalo de confiança λ obtem-nos, então, a amplitude do intervalo: $A = \dfrac{2a}{\bar{x}\sqrt{n}}$, em que $a = \Phi^{-1}(\frac{1+(1-\alpha)}{2})$.

Pergunta 10

```
pacman::p_load(pacman, rio, tidyverse,
     datasets, ggplot2, dplyr, data.table,
                hrbrthemes) # Load req.
                     packages
options(digits=15)
set.seed(50)
amostras=1000
lambda1=4.57
lambda2=0.19
gama=0.98
rc=0.20 # Contaminados
size=2500/100
MA_p <- rep(0, size)
MA_c <- rep(0, size)
vec <- rep(0, size)
mlep <- rep(0,amostras)
mlec <- rep(0,amostras)
avec_p <- rep(0,amostras)
avec_c <- rep(0,amostras)
a=qnorm((1+gama)/2) # Extremo da normal de
     lambda aproximada
## Variar o valor de dimensão das amostras
for(m in 1:25){
  n=m*100
  # Simular a situação amostras vezes
  for (i in 1:amostras){
    data_p <- rexp(n, lambda1) #vetor não
         contaminado
    temp1 <- rexp(floor(n*rc), lambda2) #
         vetor contaminado
    #mlep é o valor esperado da data não
         contaminada
    mlep[i] <- mean(data_p)
    avec_p[i] <- (2*a)/(mlep[i]*sqrt(n))
    # Criar o vetor contaminado substituindo
         os primeiros 20% dos dados puros
         pelos contaminados
    temp2 = data_p
    for(k in 1:floor(n*rc)){
      temp2[k] <- temp1[k]
    }
    data_c = temp2
```

Continuação (...)

```
    mlec[i] <- mean(data_c) # Valor esperado
         da data contaminada
    avec_c[i] <- (2*a)/(mlec[i]*sqrt(n))
  }
  MA_p[m]= mean(avec_p) # Amplitude da
       amostra não contaminada
  MA_c[m]= mean(avec_c)   # Amplitude da
       amostra não contaminada
  vec[m] <- n # Array com no. de amostras
       utilizado
}
data_1 = data.frame(vec, MA_p)
data_2 = data.frame(vec, MA_c)
## Gráfico -------------------
ggplot() + theme_linedraw() +
  ggtitle("Evolução da amplitude do intervalo
       de confiança em função da \n dimensão
       de amostras numa distribuição
       exponencial pura e contaminada") +
  geom_point(data = data_1, aes(vec, MA_p,
       color = "MA_p - Amplitude \n sem
       contaminação"),
             shape=1, size=2, alpha = 1)  +
                  geom_line(data=data_1, aes(
                  x=vec, y=MA_p), color="red"
                  ) +
  geom_point(data = data_2, aes(vec, MA_c,
       color = "MA_c - Amplitude \n com
       contaminação"),
             shape=1, size=2, alpha = 1) +
                  geom_line(data=data_2, aes(
                  x=vec, y=MA_c)) +
  labs(x="Dimensão das amostras", y="
       Amplitude do intervalo de confiança",
       color='Variável',fill='Variável') +
  theme(plot.title=element_text(hjust=0.5,
       size=32, face="bold"),
        axis.title=element_text(size=24),
        axis.text=element_text(size=16),
        legend.title=element_text(size=24),
        legend.text=element_text(size=16)) +
  scale_color_manual(values=c('blue', 'red'))
```



Parâmetros: Semente: 50; Amostras m: 1000; $\lambda = 4.57$; $\lambda_C = 0.19$; $\epsilon = 20\%$; $(1-\alpha) = 0.98$

Comentário: À semelhança do exercício anterior, efetuou-se a média de amplitude de vários intervalos de confiança de uma distribuição exponencial com diferentes dimensões. Porém neste caso, compara-se uma distribuição exponencial "pura" com $\lambda = \lambda_1 = 4.57$ com uma distribuição exponencial "contaminada" com $\lambda_C = \lambda_2 = 0.19$ em que 20% foi substituído de modo a modelar a distribuição dos *outliers*.

Naturalmente, como no exercício anterior, observa-se diretamente no gráfico um decréscimo do intervalo de confiança com o aumento da dimensão das amostras (decaimento da ordem de $1/\sqrt{n}$). Na situação com amostras contaminadas por um λ_C baixo, faz-se o reparo de que a amplitude do intervalo de confiança é menor do que na situação com as amostras inalteradas ("puras").

Através do Teorema do Limite Central, é trivial concluir que a amplitude do intervalo de confiança de λ numa distribuição exponencial é dado por $A = \frac{2a}{\bar{x}\sqrt{n}}$ (com $a = \Phi^{-1}(\frac{1+(1-\alpha)}{2})$), i.e., a amplitude depende inversamente da média das amostras e da dimensão das amostras. O aumento da média das amostras também leva a uma redução na largura do intervalo de confiança, o que é igualmente observado no gráfico.