# Estudo de caso 01 - exemplo de solução

**Felipe Campelo**[a]

[a]Departamento de Engenharia Elétrica, UFMG.

This version was compiled on September 24, 2018

**Este documento apresenta um modelo de solução do Estudo de Caso 1 da disciplina de Planejamento e Análise de Experimentos, semestre 2018-2. Atenção: isto NÃO é um exemplo de relatório.**

## Parte 1: teste sobre a média

**Informações do problema.** As seguintes informações são dadas na definição do problema:

- Parâmetro considerado como conhecido da distribuição de custos da versão atual do software: $\mu_0 = 50$.
- Questão de interesse: a nova versão apresenta *ganhos* em termos de custo médio?
- Nível de confiança desejado: $1 - \alpha = 0.99$.
- Tamanho de efeito de mínima relevância prática: $\delta^* = 4$.
- Potência desejada para efeitos iguais ou maiores que $\delta^*$: $\pi = 1 - \beta = 0.8$.

**Definição das hipóteses de interesse.** Os pontos acima resultam na seguinte formulação para as hipóteses de teste:

$$\begin{cases} H_0 : \mu = 50 \\ H_1 : \mu < 50 \end{cases}$$

**Cálculo do tamanho amostral.** Com base nas propriedades desejadas para o experimento, pode-se realizar a estimativa do tamanho amostral de três formas. A primeira envolve a realização de um experimento preliminar para estimação da variância dos dados, que poderá então ser utilizada para a derivação de um tamanho amostral razoável para o experimento. A segunda envolve a utilização de informações preliminares sobre o problema para gerar uma estimativa razoável de variância, que pode então ser utilizada no cálculo do número de observações necessárias. O terceiro seria a utilização de um tamanho amostral predefinido (p.ex., 30, 50 ou 100 execuções), o que não garantiria a potência desejada.

Neste exemplo de solução seguiremos a abordagem número 2, utilizando a seguinte fundamentação para o mesmo: a variância do software *atual* é dada como $\sigma^2 = 100$, e espera-se que o novo sistema possa trazer *ganhos* de variância. Mesmo que tais ganhos não sejam observados, a variância do sistema atual pode ser considerada como uma primeira estimativa (possivelmente sobre-estimada) da variância dos custos do novo sistema. Esta premissa técnica pode (e deve) ser avaliada posteriormente.

Assumindo $\sigma^2 = 100$ e tendo $\delta^* = 4$, $\alpha = 0.01$, $\pi = 0.8$ e uma hipótese alternativa unilateral, o tamanho amostral pode ser estimado como:

```
(my.sscalc <- power.t.test(delta       = 4,
                           sd          = 10,
                           sig.level   = 0.01,
                           power       = 0.8,
                           alternative = "one.sided",
                           type        = "one.sample"))
```

```
#
#      One-sample t test power calculation
#
#               n = 65.45847
#           delta = 4
#              sd = 10
#       sig.level = 0.01
#           power = 0.8
#     alternative = one.sided
```

```
N <- ceiling(my.sscalc$n)
```

Ou seja, uma estimativa de $N = 66$ observações.

**Coleta das observações.** Para coletar os os dados do experimento, basta seguir as instruções dadas na definição do estudo de caso:

```
suppressPackageStartupMessages(library(ExpDE))
mre <- list(name  = "recombination_bin", cr = 0.9)
mmu <- list(name  = "mutation_rand", f = 2)
mpo <- 100
mse <- list(name  = "selection_standard")
mst <- list(names = "stop_maxeval", maxevals = 10000)
mpr <- list(name  = "sphere",
            xmin = -seq(1, 20),
            xmax = 20 + 5 * seq(5, 24))

set.seed(1234) # <- isso não estava na definição do trabalho

my.sample <- numeric(N)
for (i in seq(N)){
  my.sample[i] <- ExpDE(mpo, mmu, mre, mse, mst, mpr,
                        showpars = list(show.iters = "none"))$Fbest
}
```

Antes de proceder com o teste de hipóteses, é interessante realizar uma análise exploratória dos dados. Isto pode ser feito facilmente com os gráficos básicos do R, mas a biblioteca *ggplot2* tende a gerar visualizações mais esteticamente agradáveis:

```
# Some summary statistics
summary(my.sample)
```

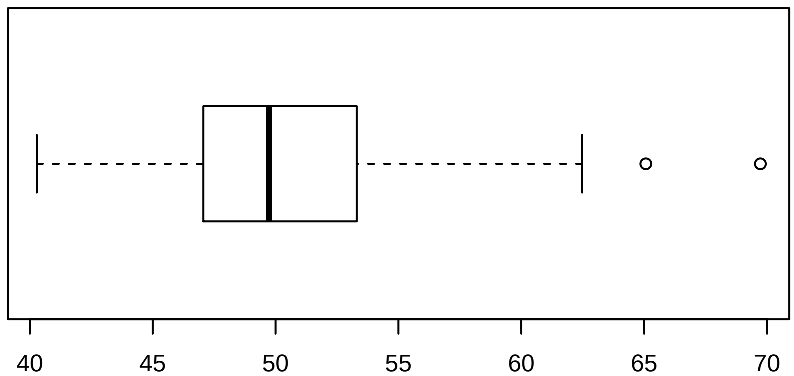
```
#     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#    40.28   47.19   49.74   50.71   53.26   69.73
```

```
var(my.sample)
```

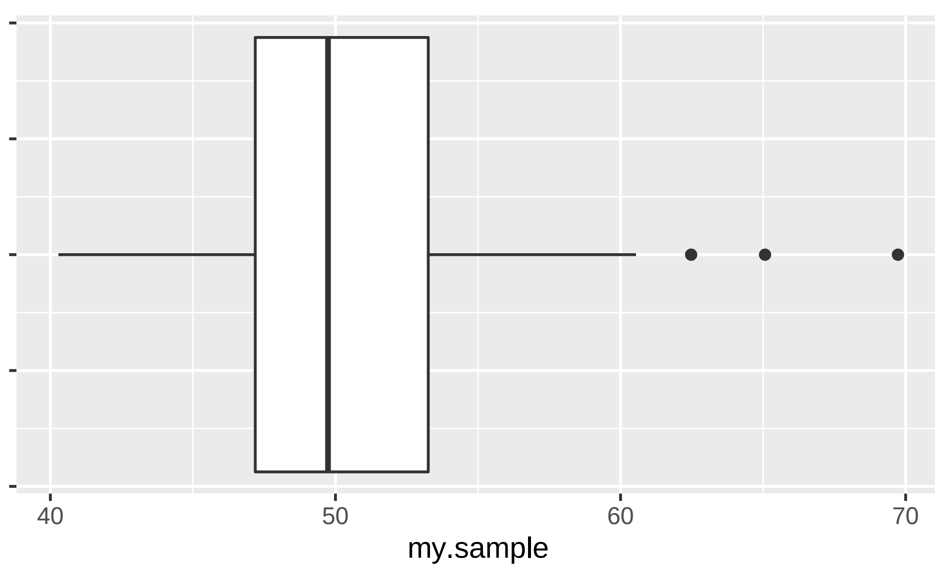
```
#  [1] 30.14134
```

Note que a variância amostral é substancialmente inferior à considerada no cálculo do tamanho amostral, o que sugere que nosso teste terá uma potência superior à nominal (assumindo que as premissas dos testes estejam válidas).

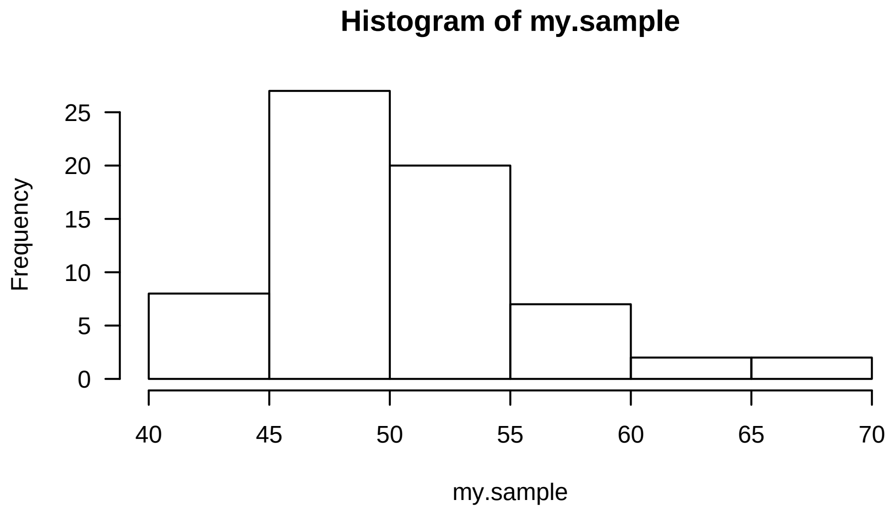
```
# Plot a boxplot (basic)
boxplot(my.sample, horizontal = TRUE)
```



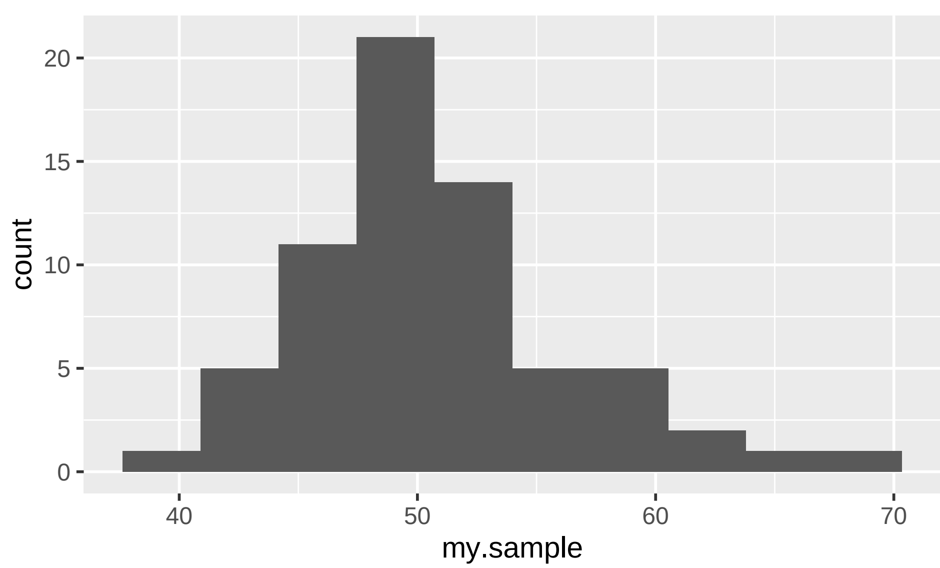

```
# Plot a boxplot (ggplot2)
suppressPackageStartupMessages(library(ggplot2))
p1 <- ggplot(as.data.frame(my.sample), aes(y = my.sample))
p1 + geom_boxplot() + coord_flip() + theme(axis.text.y = element_blank())
```



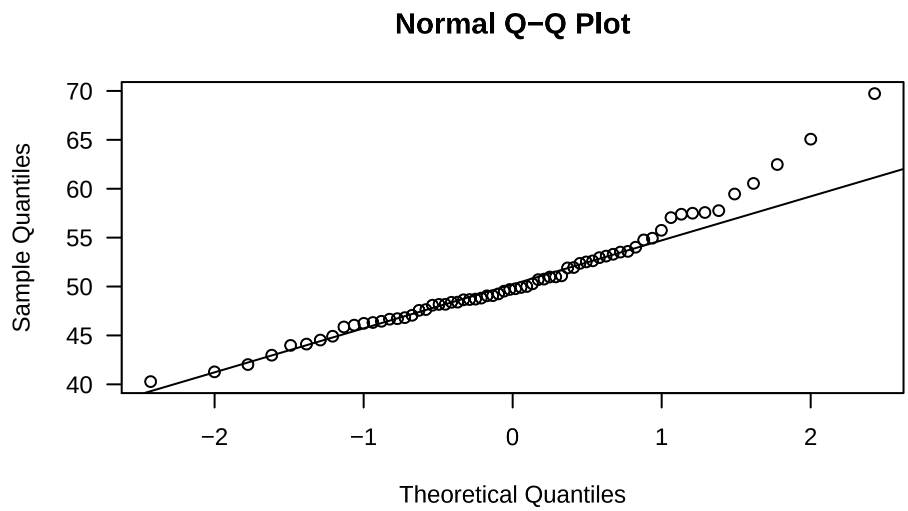

```
# plot a histogram (basic)
hist(my.sample, las = 1, breaks = 10)
```



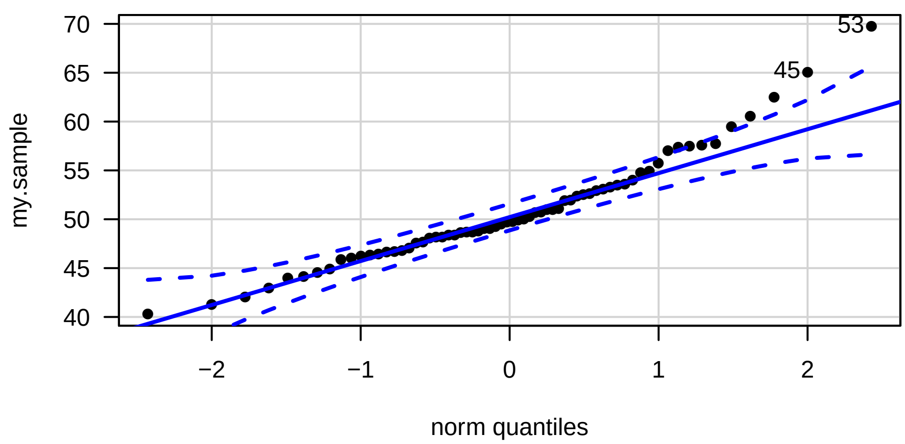

```r
# plot a histogram (ggplot2)
p2 <- ggplot(as.data.frame(my.sample), aes(x = my.sample))
p2 + geom_histogram(bins = 10)
```




```r
# Plot a Normal qq-plot (basic)
qqnorm(my.sample, las = 1)
qqline(my.sample)
```

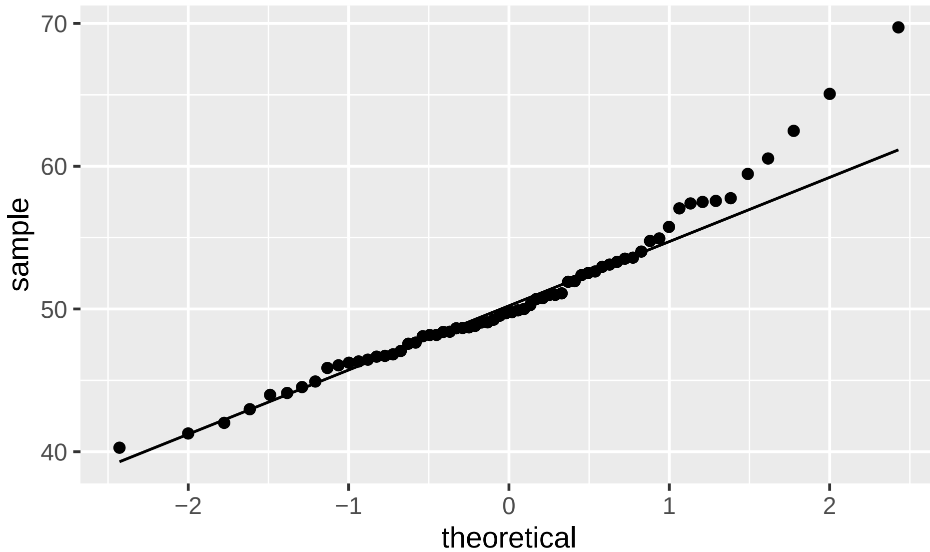



```
# Plot a Normal qq-plot (a little better)
suppressPackageStartupMessages(library(car))
qqPlot(my.sample, las = 1, pch = 16)
```




```
#  [1] 53 45
```

```
# Plot a Normal qq-plot (ggplot2)
p3 <- ggplot(as.data.frame(my.sample), aes(sample = my.sample))
p3 + geom_qq() + geom_qq_line()
```




A análise exploratória é muito importante para revelar possíveis assimetrias, outliers, ou outras peculiaridades dos dados. Neste caso específico, vemos que a distribuição das observações aparenta ter uma cauda mais pesada à direita, o que pode comprometer as premissas do teste-t. Contudo, o gráfico quantil-quantil sugere que este desvio é pequeno, e o tamanho amostral (66) é alto o bastante para que a distribuição amostral das médias já esteja suficientemente próxima da Normal (mais sobre isso na parte de teste de premissas).

**Teste de hipóteses.** Utilizando os dados coletados, o teste de hipóteses pode ser feito de forma simples:

```
(my.test <- t.test(my.sample, mu = 50,
                   alternative = "less",
                   conf.level  = 0.99))
```

```
#
#   One Sample t-test
#
#  data:  my.sample
#  t = 1.0453, df = 65, p-value = 0.8501
#  alternative hypothesis: true mean is less than 50
#  99 percent confidence interval:
#       -Inf 52.31823
#  sample estimates:
#  mean of x
#   50.70642
```

Com base nestes dados, não é possível rejeitar a hipótese nula ao nível de confiança de 99% (note o valor-p obtido). O intervalo de confiança fornecido por este teste é um intervalo aberto (posto que a alternativa era unilateral), mas é possível estimar um intervalo fechado para o valor da média utilizando:

```
t.test(my.sample, mu = 50, conf.level  = 0.99)$conf.int
```

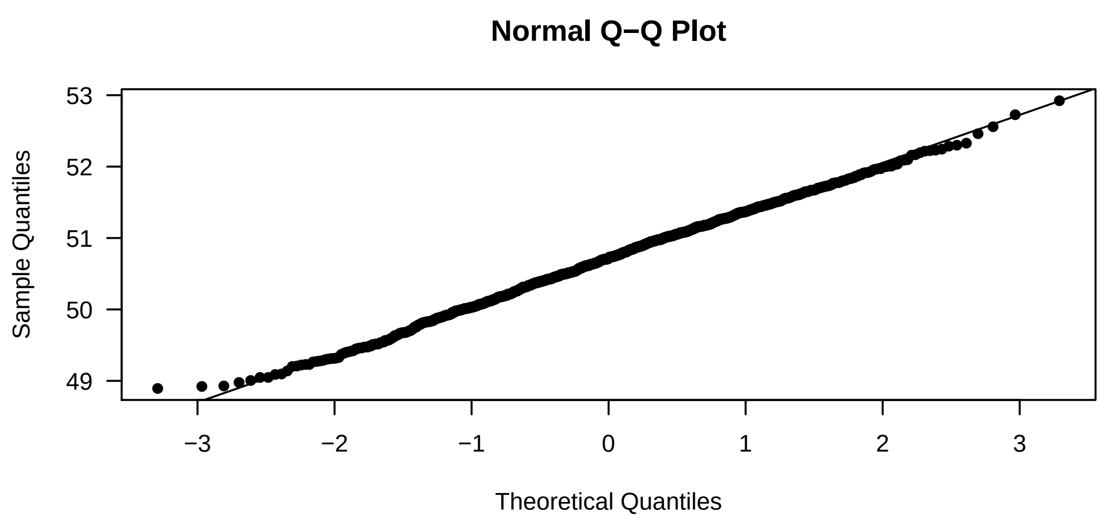
```
#  [1] 48.91315 52.49969
#  attr(,"conf.level")
#  [1] 0.99
```

**Validação das premissas.** Assumindo independência no processo de obtenção dos dados (garantido pelo planejamento) e a ausência de outras co-variáveis de importância nos resíduos (o que pode ser assumido neste caso específico, dada a simplicidade do teste), a única premissa a ser validada é a de normalidade. Os gráficos quantil-quantil gerados anteriormente sugerem uma cauda ligeiramente pesada à direita.

Dado o tamanho amostral é possível que esta leve assimetria não impacte na normalidade da distribuição amostral das médias. Uma forma simples de testar esta possibilidade envolve a estimação qualitativa da distribuição amostral das médias através de reamostragem (*bootstrap*):

```
K <- 999
boot.means <- numeric(K)
for (i in seq(K)){
  boot.sample <- sample(my.sample, replace = TRUE) # sample with replacement
  boot.means[i] <- mean(boot.sample)
}
```

```
qqnorm(boot.means, las = 1, pch = 16)
qqline(boot.means)
```




Esta estimativa da distribuição amostral das médias sugere que a mesma se aproxima bastante do que seria esperado de uma Normal, e consequentemente quaisquer efeitos de degradação nas taxas de erro devem ser pequenas o bastante para não requerer a modificação do teste estatístico utilizado. Alternativamente o teste de Wilcoxon (*Wilcoxon signed-ranks test*) poderia ser utilizado, uma vez que o mesmo não depende da premissa de normalidade.

**Discussão sobre a potência do teste.** Uma vez que a variância amostral é substancialmente inferior à variância de referência utilizada para a estimativa do tamanho amostral, é esperado que a potência efetiva para a detecção de diferenças maiores que o tamanho de efeito de mínima relevância prática seja superior aos 80% utilizados no cálculo de $N$. Um limitante inferior para a potência estatística pode ser derivado a partir da margem de confiança superior para o valor da variância. Um intervalo de confiança unilateral superior para a variância (ao nível de confiança $1-\alpha$) pode ser estimado como:

$$\sigma^2 \leq \frac{(n-1)s^2}{\chi^{2(n-1)}_{\alpha}}$$

Para a amostra disponível:

```
(CI_u <- (N - 1) * var(my.sample) / qchisq(p = 0.01, df = N - 1))
```

```
#  [1] 47.27356
```

ou seja, o intervalo entre 0 (limite inferior de variâncias, por definição) e 47.2735604 contém o valor real da variância com 99% de confiança[1]. Utilizando este limitante superior podemos estimar uma "potência de pior caso" para efeitos maiores ou iguais a $\delta^*$ (ou seja, diferenças em relação à hipótese nula de magnitude tão grande ou maior que $\delta^* = 4$) utilizando:

```
(my.pwr <- power.t.test(n           = N,
                        delta       = 4,
                        sd          = sqrt(CI_u),
                        sig.level   = 0.01,
                        type        = "one.sample",
                        alternative = "one.sided"))
```

```
#
#      One-sample t test power calculation
#
#               n = 66
#           delta = 4
#              sd = 6.875577
#       sig.level = 0.01
#           power = 0.9892823
#     alternative = one.sided
```

ou seja, teríamos uma potência de no mínimo 0.9893 para efeitos iguais ou maiores que $\delta^*$, o que é absolutamente aceitável para este experimento.

## Parte 2: teste sobre a variância

### Informações do problema.

- Parâmetro considerado como conhecido da distribuição de custos da versão atual do software: $\sigma_0^2 = 100$.
- Questão de interesse: a nova versão apresenta *ganhos* em termos de variância média?
- Nível de confiança desejado: $1 - \alpha = 0.95$.
- Uso da amostra disponível para o teste das médias.

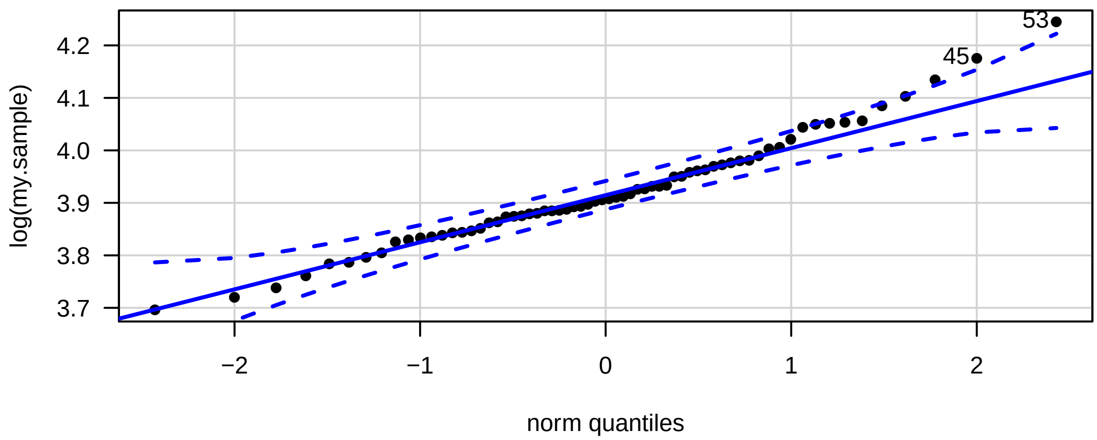
**Definição das hipóteses de interesse.** Os pontos acima resultam na seguinte formulação para as hipóteses de teste:

$$\begin{cases} H_0 : \sigma^2 = 100 \\ H_1 : \sigma^2 < 100 \end{cases}$$

[1] Note que este intervalo assume normalidade, e consequentemente é somente uma aproximação útil. Um intervalo melhor é estimado na seção sobre o teste da variância

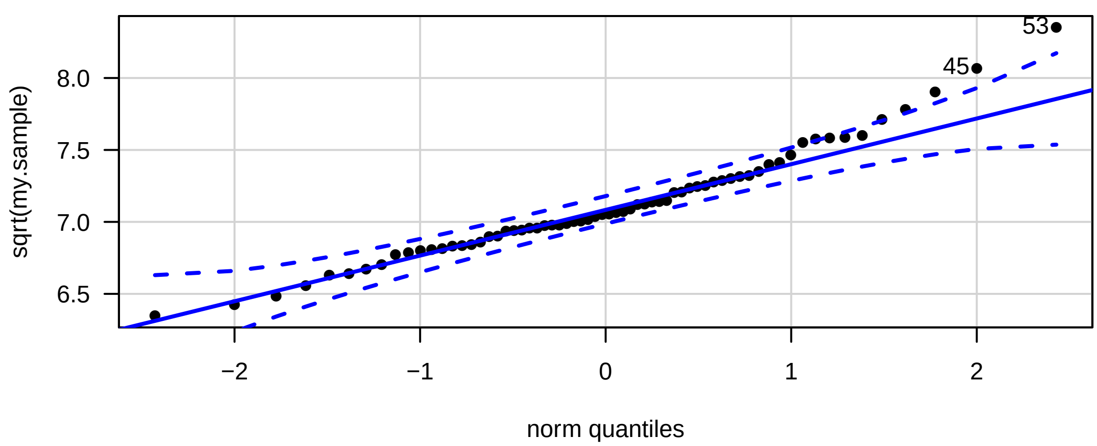
No caso dos testes sobre a variância, a premissa é requerida diretamente sobre os dados, e não somente na distribuição amostral do estimador (o o Teorema do Limite Central não se aplica ao estimador variância amostral). Assim, precisamos de i) encontrar uma transformação dos dados que leve os mesmos à normalidade, ou ii) utilizar uma abordagem que não dependa desta premissa. As transformações mais usuais para dados com caudas pesadas à direita são a logarítmica e a raiz quadrada:

```
qqPlot(log(my.sample), pch = 16, las = 1)
```




```
#  [1] 53 45
```

```
qqPlot(sqrt(my.sample), pch = 16, las = 1)
```




```
#  [1] 53 45
```

Infelizmente nenhum dos dois casos leva a distribuição para a normalidade. Uma forma de testar as hipóteses sobre a variância neste caso envolve novamente o uso de *bootstrap*, juntamente com a correspondência entre a região coberta por um intervalo de confiança e a rejeição (ou não) de uma dada hipótese relativa a um parâmetro de interesse. Desta vez utilizaremos uma implementação mais eficiente[2]

[2] E com melhores propriedades estatísticas - ver, p.ex., [http://users.stat.umn.edu/ helwig/notes/bootci-Notes.pdf](http://users.stat.umn.edu/ helwig/notes/bootci-Notes.pdf).

```r
suppressPackageStartupMessages(library(boot))
boot.out <- boot(my.sample, statistic = function(x, i){var(x[i])}, R = 999)
(my.boot.var <- boot.ci(boot.out, conf = 0.9, type = "bca"))
```

```
#  BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
#  Based on 999 bootstrap replicates
#
#  CALL :
#  boot.ci(boot.out = boot.out, conf = 0.9, type = "bca")
#
#  Intervals :
#  Level       BCa
#  90%   (21.90, 47.24 )
#  Calculations and Intervals on Original Scale
```

Note que o nível de confiança foi ajustado para 0.9, de forma a ter uma taxa de erros de 0.05 para cada lado do intervalo. Como estamos interessados somente no limitante superior, podemos ignorar o limitante inferior e declarar que, com 95% de confiança, a variância da nova versão do software é inferior a 47.2408. Um intervalo bilateral de confiança ao nível de 95% pode ser derivado de forma igualmente simples,

```r
boot.ci(boot.out, conf = 0.95, type = "bca")
```

```
#  BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
#  Based on 999 bootstrap replicates
#
#  CALL :
#  boot.ci(boot.out = boot.out, conf = 0.95, type = "bca")
#
#  Intervals :
#  Level       BCa
#  95%   (20.46, 51.02 )
#  Calculations and Intervals on Original Scale
#  Some BCa intervals may be unstable
```

De qualquer forma, a variância da nova versão pode ser declarada significativamente inferior (ao nível de confiança de 95%) à da versão atual.