Estudo de Caso 01: Desempenho de uma nova versão de Software

André Boechat, Mateus Pongelupe, Samuel Leite

24 de Setembro de 2018

Resumo

Este trabalho consiste no primeiro estudo de caso da disciplina de Planejamento e Análise de Experimentos. Nele, foram executados testes para a avaliar se o desempenho de uma nova versão de software é superior ao da anterior. A média e a variância do custo de execução foram as variáveis escolhidas para fazer essa medida, sendo que a versão atual do software possui um custo de execução dado por uma distribuição conhecida. Os resultados alcançados <!– !TODO - Adicionar resultados ao resumo –> ...

Planejamento do Experimento

Nesse experimento, está sendo avaliado o desempenho de uma nova versão de software. É conhecido que a versão atual do software possui uma distribuição para seu custo de execução com média populacional $\mu = 50$ e variância populacional $\sigma^2 = 100$. Para a nova versão do softare deseja-se investigar seus resultados quanto a melhorias de desempenho, isto é, menor custo médio de execução e/ou menor variância. Com esse intuito, foram desenvolvidos dois experimentos: um para availar a média e outro para availar a variância.

Teste da média

Para avaliar se o desempenho do novo software é melhor que a versão antiga, está sendo observado se a média do custo de execução é menor. Assim, ao definir a hipótese H_1, podemos fazer com que ela seja unidirecional, isto é, a região de interesse do teste está na direção em que a média de execução da nova versão seja menor que a média atual. Dessa forma, a hipótese nula H_0 e a hipótese H_1 podem ser definidas como:

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

Para esse teste, definiu-se um nível de significância de $\alpha = 0.01$, um efeito de relevância mínimo $\delta^* = 4$ e uma potência desejada de $\pi = 1 - \beta = 0.8$.

```
h0.mean = 50
h0.sd = sqrt(100)

t1.alpha = 0.01
t1.delta = 4
t1.beta = 0.2
t1.power = 1 - t1.beta
```

Assumindo que a hipótese nula H_0 se comporte com uma distribuição de média populacional $\mu = 50$ e variância populacional $\sigma^2 = 100$, pode-se calcular o número de amostras mínimo a partir do teste Z, haja vista que a variância da hipótese nula é conhecida. Para fazer esse cálculo, foi usado o pacote *asbio*:

```
library(asbio)
n <- power.z.test(power = t1.power, alpha = t1.alpha, effect = t1.delta, sigma = h0.sd, test = "one.tail
t1.N <- ceiling(n)
cat("N: ", t1.N)
```

```
## N:  63
```

Assim, fazendo uso do teste Z, precisaremos de uma amostra de tamanho $N = 63$ para executar o nosso teste com uma potência $\pi = 0.8$. Por nossa hipótese H_1 ser unidirecional, a região crítica do teste Z pode ser determinada como:

$$P(z_\alpha \leq Z_0 | H_0 éverdadeira)$$

Isto é, para que a hipótese nula seja rejeitada com um nível de confiança de 99% é preciso que $z_\alpha > Z_0$.

Teste da variância

A section detailing the experimental setup. This is the place where you will define your test hypotheses, e.g.:

$$\begin{cases} H_0 : \mu = 10 \\ H_1 : \mu < 10 \end{cases}$$

including the reasons behind your choices of the value for H_0 and the directionality (or not) of H_1.

This is also the place where you should discuss (whenever necessary) your definitions of minimally relevant effects (δ^*), sample size calculations, choice of power and significance levels, and any other relevant information about specificities in your data collection procedures.

Coleta de Dados

A coleta de dados foi simulada a partir da rotina sugerida no caso de uso, com uma pequena modificação: uma *seed* foi definida para a execução do programa, de forma a garantir sua reproducibilidade.

```r
# Loading required package
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))

# Setting seed so the program can be reproduced.
set.seed(1998)

# One sample
ExpDE(mpo, mmu, mre, mse, mst, mpr,
               showpars = list(show.iters = "none"))$Fbest
```

Em nossos experimentos, precisaremos coletar um número arbitrário **N** de amostras. Portanto, a partir das rotinas acima, foram criadas duas funções para essa coleta:

- *generate_sample* : Coleta uma única amostra.
- *generate_n_samples* : Coleta **n** amostras no formato de um *data.frame.*

Segue abaixo a codificação dessas funções, bem como um exemplo da chamada de *generate_n_samples* para $N = 10$:

```r
#Generates one sample
generate_sample <- function() {
  return(ExpDE(mpo, mmu, mre, mse, mst, mpr,
               showpars = list(show.iters = "none"))$Fbest);
```

```
}

#Generates n samples on the data.frame format
generate_n_samples <- function(n) {
  cost <- replicate(n, generate_sample())
  return(data.frame(cost))
}

#Example for N=10
generate_n_samples(10)
```

Análise Estatístiica

Teste da Média

Dados os parâmetros definidos na seção *Planejamento do Experimento* para o teste da média, foram recolhidas $N = 63$ amostras e o teste foi executado nas linhas abaixo. O intervalo de confiança também foi calculado, considerando uma distribuição normal cuja variância populacional $\sigma^2 = 100$ é conhecida.

```
## Getting the samples
t1.samples <- generate_n_samples(t1.N)
## Writing samples to csv file
write.csv(t1.samples, 'test-one.csv')

## Test Z Execution
t1.mean <- mean(t1.samples$cost)
t1.sd <- sd(t1.samples$cost)
z0 <- (t1.mean - h0.mean)/(h0.sd/sqrt(t1.N))
t1.z_alpha <- qnorm(t1.alpha)

## Confidence interval
t1.error <- qnorm(1-(t1.alpha/2)) * h0.sd / sqrt(n)


cat("\n",
"Mean: ", t1.mean, "\n",
"Z0: ", z0 ,"\n",
"Zalpha: ",  t1.z_alpha ,"\n",
"Confidence Interval: ", t1.mean - t1.error, " <= ", t1.mean, " <= ", t1.mean + t1.error, "\n")
```

```
##
##  Mean:  50.78709
##  Z0:  0.6247342
##  Zalpha:  -2.326348
##  Confidence Interval:  47.53475  <=  50.78709  <=  54.03943
```

Como $Z_\alpha < Z_0$, conclui-se que não há evidências suficientes para rejeitar H_0 a um nível de confiança de 99%. Falta validar e discutir as assumptions e discutir a potência do teste -> um nível de confiança muito elevado pode levar a uma dificuldade de detectar que a hipótese nula é falsa. <!– !TODO - Fazer análise da média –> ...

Checking Model Assumptions

The assumptions of your test should also be validated, and possible effects of violations should also be explored.

Conclusions and Recommendations

The discussion of your results, and the scientific/technical meaning of the effects detected, should be placed here. Always be sure to tie your results back to the original question of interest!