Estatística Computacional 2 - Método de Monte Carlo pra inferência estatística.

Henrique Valaski de Mello - GRR20185621

UFPR - Universidade Federal do Paraná

Antônio José Hamerschmidt - GRR20185613

UFPR - Universidade Federal do Paraná

Introdução

Em estatística quando se trabalha com uma base de dados, é padrão criar uma hipótese que assume que a ocorrência de qualquer amostra dentro da base tem a mesma probabilidade (hipótese nula ou H0). Em seguida é feito o teste de hipótese, que vai nos ajudar a rejeitar ou não a hipótese nula. Com Monte Carlo, temos apenas uma amostra, então a ideia é calcular a estatística de teste para esse conjunto de dados (supondo sempre que H0 é verdadeiro), e em seguida é simulado valores com as mesmas características e calculado as suas estatísticas.

Com isso teremos N estatísticas de teste, uma dos valores reais e N-1 de valores simulados. Com isso, é calculado a proporção de valores iguais ou mais extremos que a estatística de teste da amostra. Proporções altas indicam que as estatísticas de teste da amostra não são extremas, o que reforça H0. Já o oposto indica que a estatística de teste tem pouca probabilidade de ocorrer ao acaso, o que refuta H0.

Neste trabalho vamos utilizar os Métodos de Monte Carlo para fazer duas inferências estatísticas, que será referente a uma base de dados que compara a presença ou não de minhocas à quantidade de zeatina na planta, e compara também a presença de solo ou solução hidropônica e a quantidade de zeatina. A zeatina é basicamente um dos hormônios de crescimento presente em diversas plantas, promovendo o crescimento dos brotos laterais e estimulando a divisão celular para produzir plantas mais espessas. Será realizado duas inferências estatísticas, uma sobre a presença ou não de minhocas, e uma sobre a presença ou não de solo.

Para fazer a inferência é suposto que vai se ter uma concentração média maior de zeatina nas plantas que cresceram com a presença de solo. Supondo que mi1 é a concentração média de zeatina em plantas que cresceram com solo, e mi2 a concentração média de zeatina em plantas que cresceram na hidroponia, temos que:

```
db <- read.csv('../data/table.csv', sep = ';')
```

Considera-se que

$$\mu_1$$

seja a média em miligramas do elemento Zeatina nas plantas de solo, e que

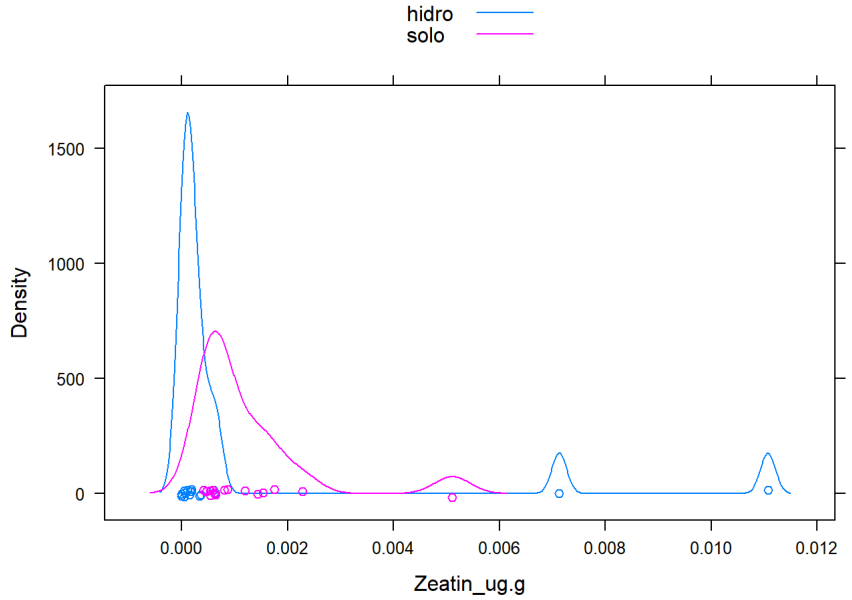
$$\mu_2$$

se refere quantidade do mesmo elemento porém para as plantas hidroponicas. Pretende-se desta forma identificar se de algum modo possuimos diferença, ou seja, a presença de zeatina é maior no solo? Dado o modo de cultivo para isso temos um teste de hipótese que pode ser formulado com as seguintes hipóteses

$$H_0 : \mu_1 - \mu_2 = 0 \Rightarrow \mu_1 = \mu_2$$
$$H_a : \mu_1 - \mu_2 > 0 \Rightarrow \mu_1 > \mu_2$$

Inicalmente podemos verificar com o gráfico abaixo a comparação entre as plantas de solo e hidroponicas dado o nível de zeatina presente nelas:

```
solo <- db %>% filter(Matrix. == "Soil") %>% select(Zeatin_ug.g)
hidro <- db %>% filter(Matrix. == "Hydroponic") %>% select(Zeatin_ug.g)
hidro$tipo = 'hidro'
solo$tipo = 'solo'
solo <- solo[0:16,]
db_ <- rbind(solo, hidro)
densityplot(~Zeatin_ug.g, groups = tipo, data = db_, auto.key = TRUE)
```



Nota-se uma presença maior em valores menores nas plantas hidroponicas, porém podemos notar abaixo a diferença das médias, temos que as médias são:

```
# Médias
tapply(db_$Zeatin_ug.g, db_$tipo, mean)
```

```
##       hidro        solo
## 0.001316650 0.001212178
```

e sua diferença é dada por

```
diff(tapply(db_$Zeatin_ug.g, db_$tipo, mean))
```

```
##          solo
## -0.0001044719
```

Para prosseguir iremos realizar o procedimento padrão para futuramente realizar a comparação com o método de monte carlo, podemos realizar o Teste F para igualdade de variâncias,

```
var.test(x = solo$Zeatin_ug.g, y = hidro$Zeatin_ug.g)
```

```
##
##  F test to compare two variances
##
## data:  solo$Zeatin_ug.g and hidro$Zeatin_ug.g
## F = 0.14026, num df = 15, denom df = 15, p-value = 0.000472
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.04900573 0.40143403
## sample estimates:
## ratio of variances
##          0.1402589
```

```
n1 <- length(solo$Zeatin_ug.g)
v1 <- var(solo$Zeatin_ug.g)
n2 <- length(hidro$Zeatin_ug.g)
v2 <- var(hidro$Zeatin_ug.g)
(s.pond <- sqrt(((n1 - 1) * v1 + (n2 - 1) * v2)/(n1 + n2 - 2)))
```

```
## [1] 0.002361259
```

Abaixo podemos obter o resultado para o teste de hipotese:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 > \mu_2$$

```
mu0  <- 0
t.test(x = solo$Zeatin_ug.g, y = hidro$Zeatin_ug.g, alternative = "greater",
       var.equal = TRUE, mu = mu0)
```

```
##
##  Two Sample t-test
##
## data:  solo$Zeatin_ug.g and hidro$Zeatin_ug.g
## t = -0.12514, df = 30, p-value = 0.5494
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.001521398          Inf
## sample estimates:
##   mean of x   mean of y
## 0.001212178 0.001316650
```

Obtém-se pelo metodo padrão um p-value = 0.54, com um nível de significancia de 5% rejeita-se a hipotese nula, ou seja possuímos evidencias estatísticas para concluir que a quantidade de Zeatina é maior em plantas de solo.

```
m1 <- mean(solo$Zeatin_ug.g)
m2 <- mean(hidro$Zeatin_ug.g)
## Estatística de teste
(tcalc  <- (m1 - m2)/(s.pond * sqrt(1/n1 + 1/n2)))
```

```
## [1] -0.1251414
```

```
## Valor crítico
(tcrit  <-  qt(.025, df = n1 + n2 - 2, lower.tail = FALSE))
```

```
## [1] 2.042272
```

```
## p-valor
pt(tcalc, df = n1 + n2 - 2, lower.tail = FALSE)
```

```
## [1] 0.5493768
```

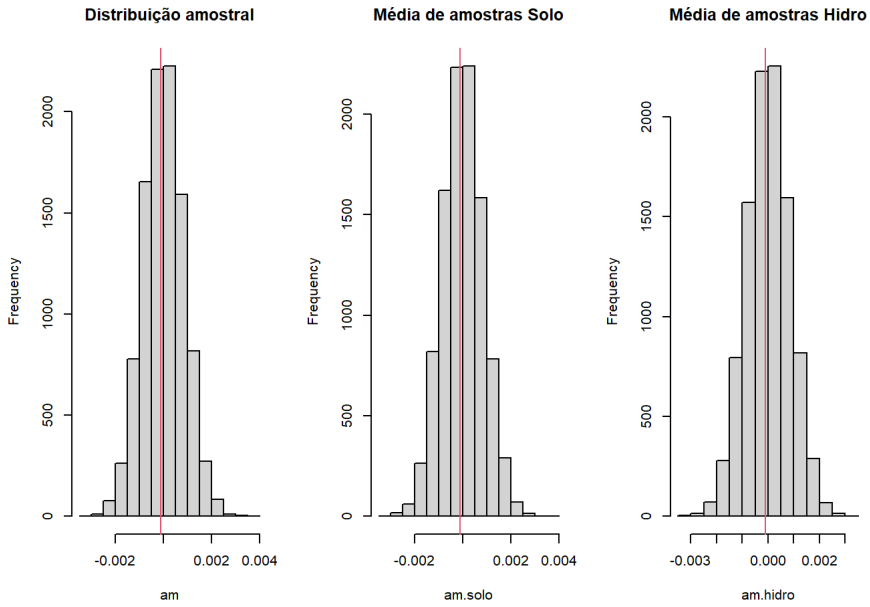
Abaixo realiza-se o teste por simulação de Monte Carlo para assim comparar com os testes realizados anteriormente:

```
## Teste por simulação de Monte Carlo
N <- 1e4
## Simula direto da distribuição amostral
am <- replicate(N, rnorm(1, mu0, s.pond * sqrt(1/n1 + 1/n2)))
```

```
am.solo <- replicate(
    N, diff(tapply(rnorm(32, m1, s.pond), db_$tipo, mean))
)
am.hidro <- replicate(
    N, diff(tapply(rnorm(32, m2, s.pond), db_$tipo, mean))
)
## Visualização
(med.amostral <- m1 - m2)
```

```
## [1] -0.0001044719
```

```
par(mfrow = c(1, 3))
{hist(am, main = "Distribuição amostral")
abline(v = med.amostral, col = 2)}
{hist(am.solo, main = "Média de amostras Solo")
abline(v = med.amostral, col = 2)}
{hist(am.hidro, main = "Média de amostras Hidro")
abline(v = med.amostral, col = 2)}
```




```
par(mfrow = c(1, 1))
## p-valor empírico
sum(am >= med.amostral)/N
```

```
## [1] 0.5494
```

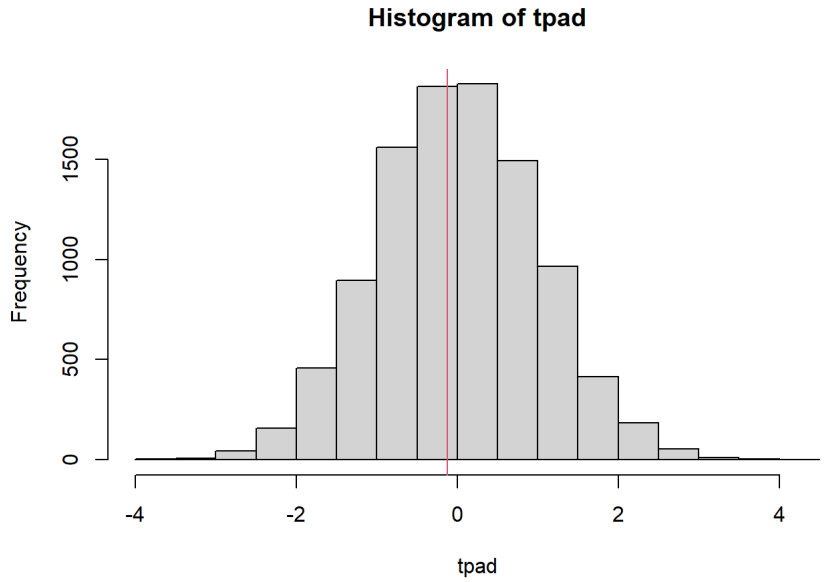
```
sum(am.solo >= med.amostral)/N
```

```
## [1] 0.5487
```

```
sum(am.hidro >= med.amostral)/N
```

```
## [1] 0.5533
```

```
tpad <- (am - mu0)/(s.pond * sqrt(1/n1 + 1/n2))
{hist(tpad)
abline(v = tcalc, col = 2)}
```




```
sum(tpad >= tcalc)/N
```

```
## [1] 0.5494
```

Nota-se que os valores estão próximos, chegando assim ao mesmo resultado. ## Presença de minhoca no solo e sua influência na quantidade de Zeatina

```
knitr::kable(sample(db[0:5,]))
```

Outlier_LOD..3	ABA_ug.g	Ade_ug.g	iP_ug.g	Sample_name	Plants.	Matrix.	IAA_.D.U.	Outlier_LOD.	Outlier_LOD..1	Outlier_LOD..2	Zea
LOD	8.0e-06	0.0364793	3.590.316.479	Lterr 1.1	Yes	Soil	U	N	N	N	0.
LOD	8.0e-06	0.0256643	4.591.537.723	Lterr 1.2	Yes	Soil	U	N	N	N	0.
LOD	8.0e-06	0.0199014	584.602.388	Lterr 1.3	Yes	Soil	U	N	N	N	0.
N	2.9e-03	0.0246555	0.00112033	Lterr 1.4	Yes	Soil	U	N	N	N	0.
LOD	8.0e-06	0.0281940	0.008660618	Lterr 1.5	Yes	Soil	U	N	N	N	0.

Iremos agora realizar o estudo nas amostras de solo, e estudar se a presença ou não de minhocas aumenta/diminui a quantidade de zeatina na amostra,
Considera-se que

$$\mu_1$$

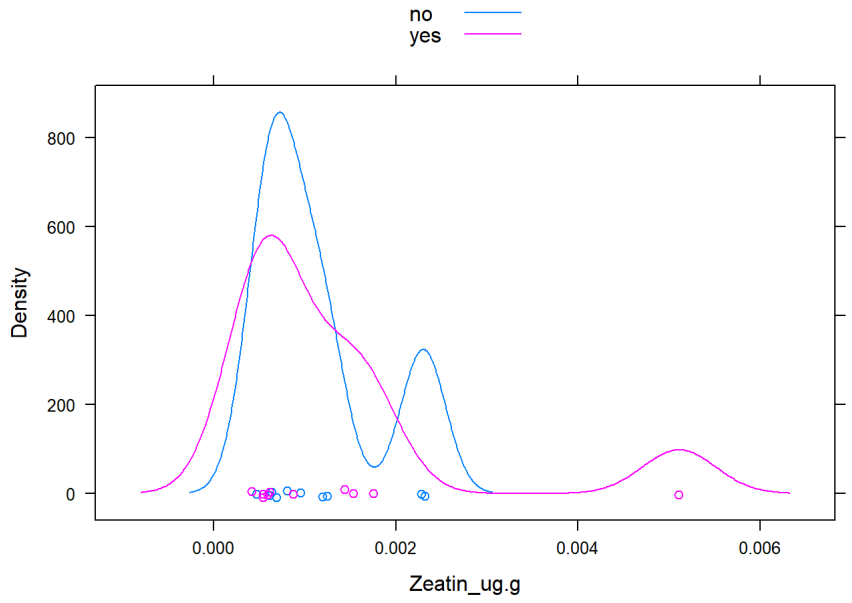
seja a média em miligramas do elemento Zeatina nas plantas de solo com presença de minhoca, e que

$$\mu_2$$

se refere quantidade do mesmo elemento porém para as plantas que não apresentaram presença de minhoca. Temos um teste de hipótese que pode ser formulado com as seguintes hipóteses:

$$H_0 : \mu_1 - \mu_2 = 0 \Rightarrow \mu_1 = \mu_2$$
$$H_a : \mu_1 - \mu_2 > 0 \Rightarrow \mu_1 > \mu_2$$

```
db <- db %>% filter(Matrix. == 'Soil')
worm <- db %>% filter(Earthworms. == "Yes") %>% select(Zeatin_ug.g)
noworm <- db %>% filter(Earthworms. == "No") %>% select(Zeatin_ug.g)
worm$tipo = 'yes'
noworm$tipo = 'no'
#solo <- solo[0:16,]
db_ <- rbind(worm, noworm)
densityplot(~Zeatin_ug.g, groups = tipo, data = db_, auto.key = TRUE)
```



Acima nota-se a distribuição de zeatina dado presença ou não de minhoca na amostra e não é possível notar a principio nenhum padrão ou diferença nitida na amostra.
Para prosseguir iremos verificar as médias e outras métricas e teste já realizados acima:

```
# Médias
tapply(db_$Zeatin_ug.g, db_$tipo, mean)
```

```
##          no         yes
## 0.001120022 0.001339961
```

```
diff(tapply(db_$Zeatin_ug.g, db_$tipo, mean))
```

```
##         yes
## 0.0002199391
```

```
var.test(x = worm$Zeatin_ug.g, y = noworm$Zeatin_ug.g)
```

```
##
##  F test to compare two variances
##
## data:  worm$Zeatin_ug.g and noworm$Zeatin_ug.g
## F = 4.4296, num df = 9, denom df = 9, p-value = 0.0371
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   1.100249 17.833534
## sample estimates:
## ratio of variances
##           4.429598
```

```
n1 <- length(worm$Zeatin_ug.g)
v1 <- var(worm$Zeatin_ug.g)
n2 <- length(noworm$Zeatin_ug.g)
v2 <- var(noworm$Zeatin_ug.g)
(s.pond <- sqrt(((n1 - 1) * v1 + (n2 - 1) * v2)/(n1 + n2 - 2)))
```

```
## [1] 0.001102814
```

Abaixo podemos obter o resultado para o teste de hipotese:

$$H_0 : \mu_1 = \mu_2$$
$$H_a : \mu_1 > \mu_2$$

```
mu0  <- 0
t.test(x = worm$Zeatin_ug.g, y = noworm$Zeatin_ug.g, alternative = "greater",
       var.equal = TRUE, mu = mu0)
```

```
##
##  Two Sample t-test
##
## data:  worm$Zeatin_ug.g and noworm$Zeatin_ug.g
## t = 0.44595, df = 18, p-value = 0.3305
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.0006352895           Inf
## sample estimates:
##   mean of x   mean of y
## 0.001339961 0.001120022
```

Obtém-se pelo metodo padrão um p-value = 0.33, com um nível de significancia de 5% rejeita-se a hipotese nula, ou seja possuímos evidencias estatísticas para concluir que a quantidade de Zeatina é maior em plantas de solo com presença de minhoca.

```
m1 <- mean(worm$Zeatin_ug.g)
m2 <- mean(noworm$Zeatin_ug.g)
## Estatística de teste
(tcalc  <- (m1 - m2)/(s.pond * sqrt(1/n1 + 1/n2)))
```

```
## [1] 0.4459491
```

```
## Valor crítico
(tcrit  <-  qt(.025, df = n1 + n2 - 2, lower.tail = FALSE))
```

```
## [1] 2.100922
```

```
## p-valor
pt(tcalc, df = n1 + n2 - 2, lower.tail = FALSE)
```

```
## [1] 0.3304759
```

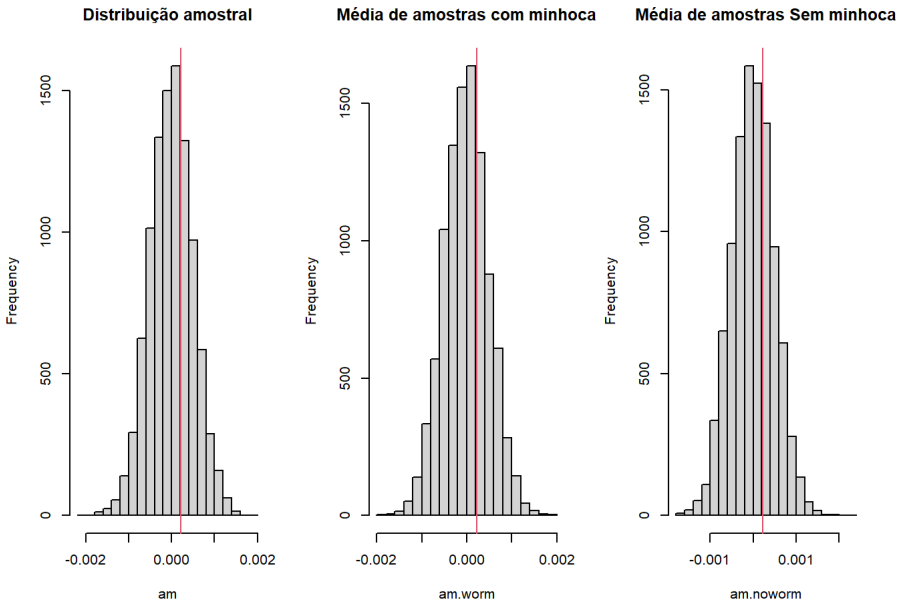
Como realizado anteriormente agora realiza-se o teste por simulação de Monte Carlo para assim comparar com os testes realizados anteriormente:

```
## Teste por simulação de Monte Carlo
N <- 1e4
## Simula direto da distribuição amostral
am <- replicate(N, rnorm(1, mu0, s.pond * sqrt(1/n1 + 1/n2)))
```

```
am.worm <- replicate(
    N, diff(tapply(rnorm(20, m1, s.pond), db_$tipo, mean))
)
am.noworm <- replicate(
    N, diff(tapply(rnorm(20, m2, s.pond), db_$tipo, mean))
)
## Visualização
(med.amostral <- m1 - m2)
```

```
## [1] 0.0002199391
```

```
par(mfrow = c(1, 3))
{hist(am, main = "Distribuição amostral")
abline(v = med.amostral, col = 2)}
{hist(am.worm, main = "Média de amostras com minhoca")
abline(v = med.amostral, col = 2)}
{hist(am.noworm, main = "Média de amostras Sem minhoca")
abline(v = med.amostral, col = 2)}
```




```
par(mfrow = c(1, 1))
## p-valor empírico
sum(am >= med.amostral)/N
```

```
## [1] 0.3251
```

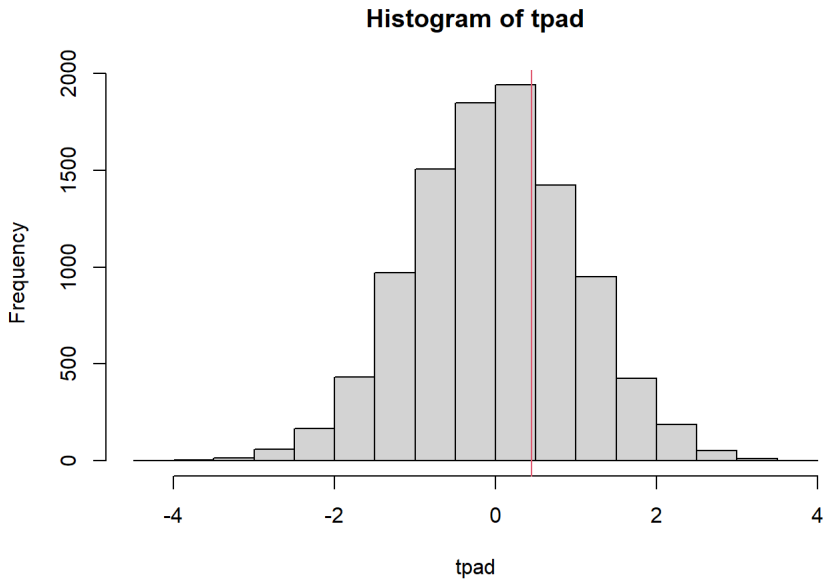
```
sum(am.worm >= med.amostral)/N
```

```
## [1] 0.317
```

```
sum(am.noworm >= med.amostral)/N
```

```
## [1] 0.3268
```

```
## Padroniza a distribuição para t(n1 + n2 - 2)
tpad <- (am - mu0)/(s.pond * sqrt(1/n1 + 1/n2))
hist(tpad)
abline(v = tcalc, col = 2)
```




```
sum(tpad >= tcalc)/N
```

```
## [1] 0.3251
```

Temos novamente o valor aproximado ao obtido anteriormente e concluimos o mesmo resultado de anteriormente, rejeita-se H0.