BI-PST

Domáca úloha

Autori: Šimon Minárik, Matúš Botek, Matej Šutý

Obsah

Dáta a parametry

Reprezentant: Matej Šutý
K = 27
L = 4
M = ((27+4)*47)mod(11)+1 = (1457 % 11)+1 = 6
Dáta - case0302: koncentrace dioxinu dle vojenského působiště
Úlohy sme spracovávali pomocou programu RStudio

Úloha č. 1 - Načítanie a preskúmanie dát

Dáta obsahujú merania hladiny dioxínu TCDD (parts per trillion) v krvi vojakov. Merania sú rozdelené do 2 skupín, 1. skupina sú vojaci slúžiaci vo vojne vo Vietname a 2. skupina sú vojaci slúžiaci inde.
Skupina meraní vojakov, ktorí slúžili vo Vietname obsahuje 646 záznamov a skupina vojakov slúžiacich inde obsahuje 97 záznamov.

Najprv sme dáta načítali a rozdelili do skupín:

```
library(Sleuth2)
vietnam <- subset(case0302, Veteran=='Vietnam')[,1]
other <- subset(case0302, Veteran=='Other')[,1]
```

Pre obidve skupiny sme následne odhadli strednú hodnotu, rozptyl a medián:

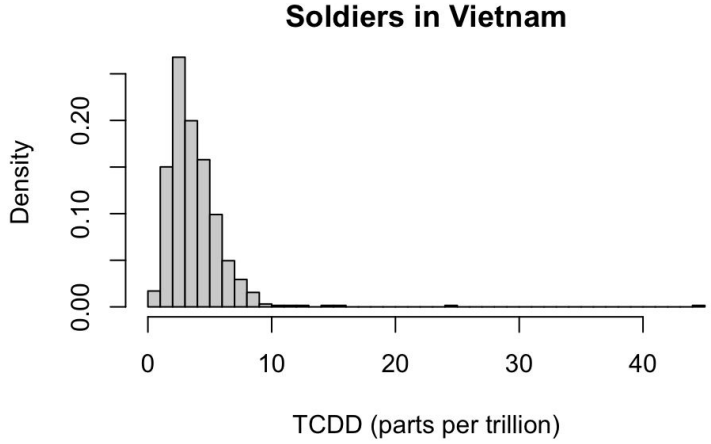
```
mean(vietnam) = 4.260062   |   mean(other) = 4.185567
var(vietnam) = 6.983426    |   var(other) = 5.29854
median(vietnam) = 4        |   median(other) = 4
```

Dáta nasvedčujú tomu, že vojaci, ktorí neboli prítomní vo Vietname, boli dioxínu TCDD vystavení menej. TCDD sa nachádzal v herbicíde Agent Orange, ktorý bol používaní americkou armádou v džungliach Vietnamu, takže je to odôvodniteľný objav.

Úloha č. 2 - Hustota a distribučná funkcia

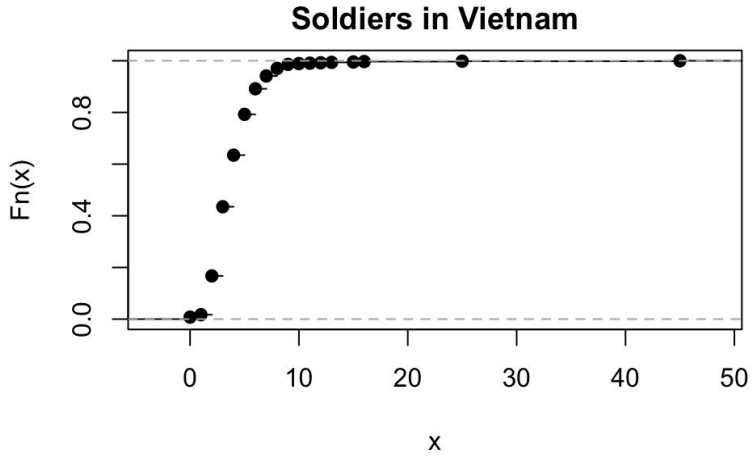
Hustotu skupiny vojakov vo Vietname sme odhadli pomocou histogramu. Počet breakpointov histogramu sme nastavili na 45, aby boli vidieť aj vysoké hodnoty TCDD a graf lepšie vykresloval skutočnosť. Rko ponúka zmenu parametra *freq*, vďaka ktorému histogram zobrazuje hustotu rozdelenia.

```
hist(vietnam,
     breaks = 45,
     freq=FALSE,
     main = "Soldiers in Vietnam",
     xlab = "TCDD (parts per trillion)",
     ylab = "Density")
```



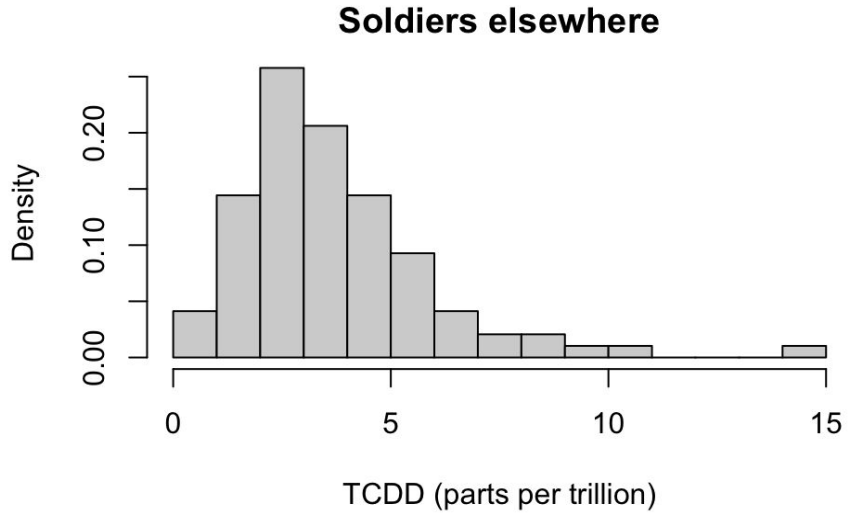
Distribučnú funkciu sme odhadli pomocou empirickej distribučnej funkcie.

```
plot(ecdf(vietnam),
     main = "Soldiers in Vietnam")
```

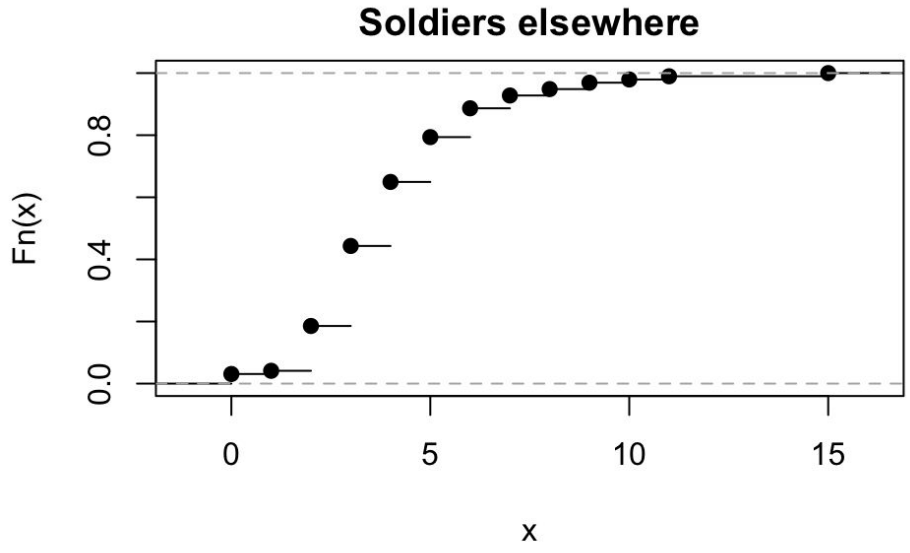


Obdobne sme hustotu a distribučnú funkciu odhadli aj pre skupinu vojakov slúžiacich mimo Vietnam. Jedinou zmenou je počet breakpointov histogramu, pretože dáta neobsahujú toľko rozdielnych hodnôt ako predošlá skupina.

```
hist(other,
     breaks = 15,
     freq=FALSE,
     main = "Soldiers elsewhere",
     xlab = "TCDD (parts per trillion)",
     ylab = "Density")
```




```
plot(ecdf(other),
     main = "Soldiers elsewhere")
```



Úloha č. 3 - Najbližšie rozdelenie

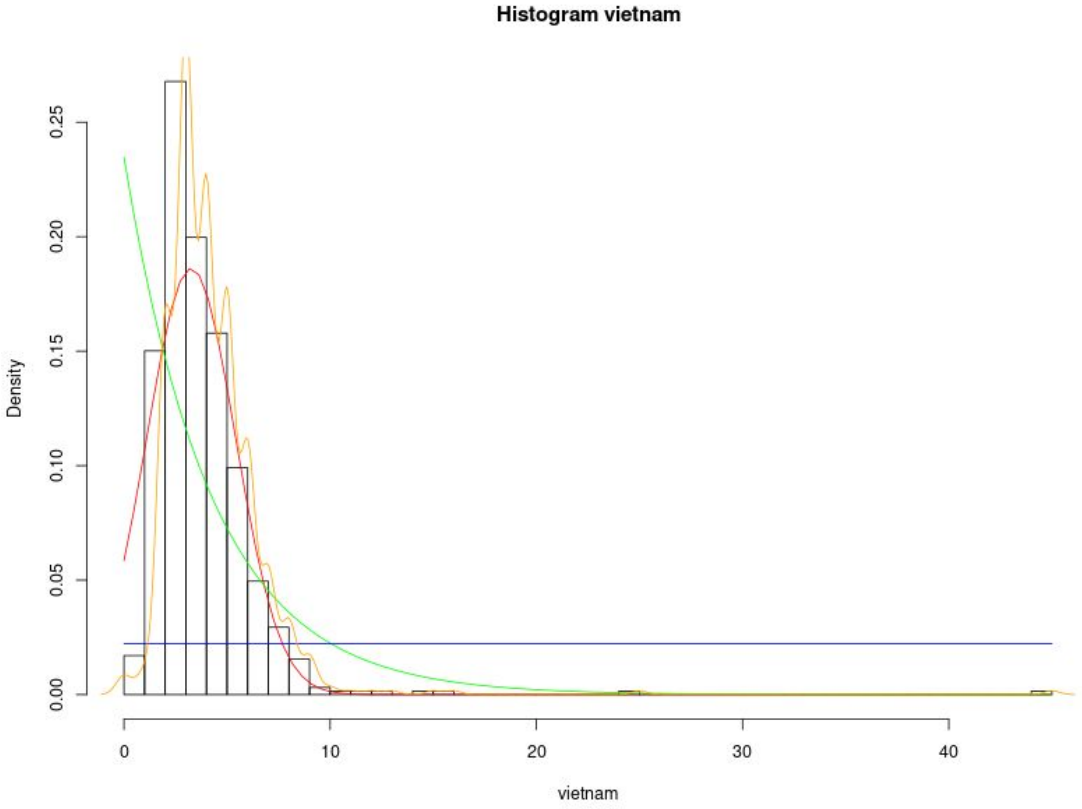
Z datasetu **vietnam** som vypočítal strednú hodnotu *Evietnam* a smerodatnú odchýľku *sd(vietnam)*. Pomocou týchto hodnôt som do grafu vykreslil krivku normálneho rozdelenia(červená). V nej som skúšal zvyšovať a znižovať hodnoty *Evietnam, sd(vietnam)* aby čo najlepšie vystihovali histogram.

Na grafe vidno aj krivky rovnomerného rozdelenia(modrá), a exponenciálneho rozdelenia(zelená). Oranžová krivka popisuje hustotu rozdelenia samotného datasetu.

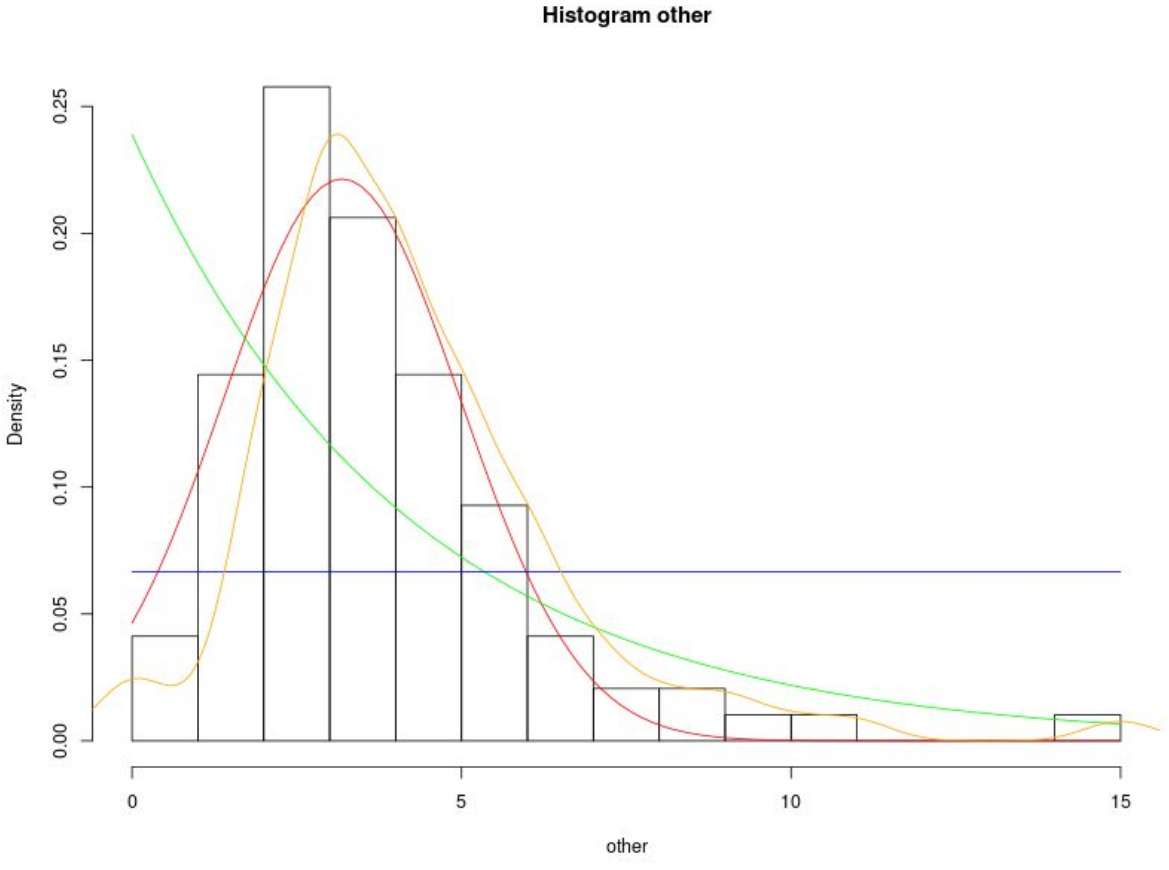
Rovnako som postupoval aj pri datasete **other.**

```
Evietnam <- mean(vietnam)
minvietnam <- min(vietnam)
maxvietnam <- max(vietnam)
vecvietnam <- seq(minvietnam, maxvietnam, length.out=100)

vietnamUnif <- dunif(vecvietnam, minvietnam, maxvietnam)
vietnamNorm <- dnorm(vecvietnam, Evietnam - 1, sd(vietnam) - 0.5)

lambdavietnam <- 1/Evietnam
vietnamExp <- dexp(vecvietnam, lambdavietnam)

hist(vietnam, breaks = 45, freq=FALSE, main = "Histogram vietnam")
lines(vecvietnam, vietnamNorm, col="red")
lines(vecvietnam, vietnamExp, col="green")
lines(vecvietnam, vietnamUnif, col="blue")
lines(density(vietnam), col='orange')
```

Histogram datasetu **vietnam.**



Histogram datasetu **other.**



Úloha č. 4 - Náhodný výber

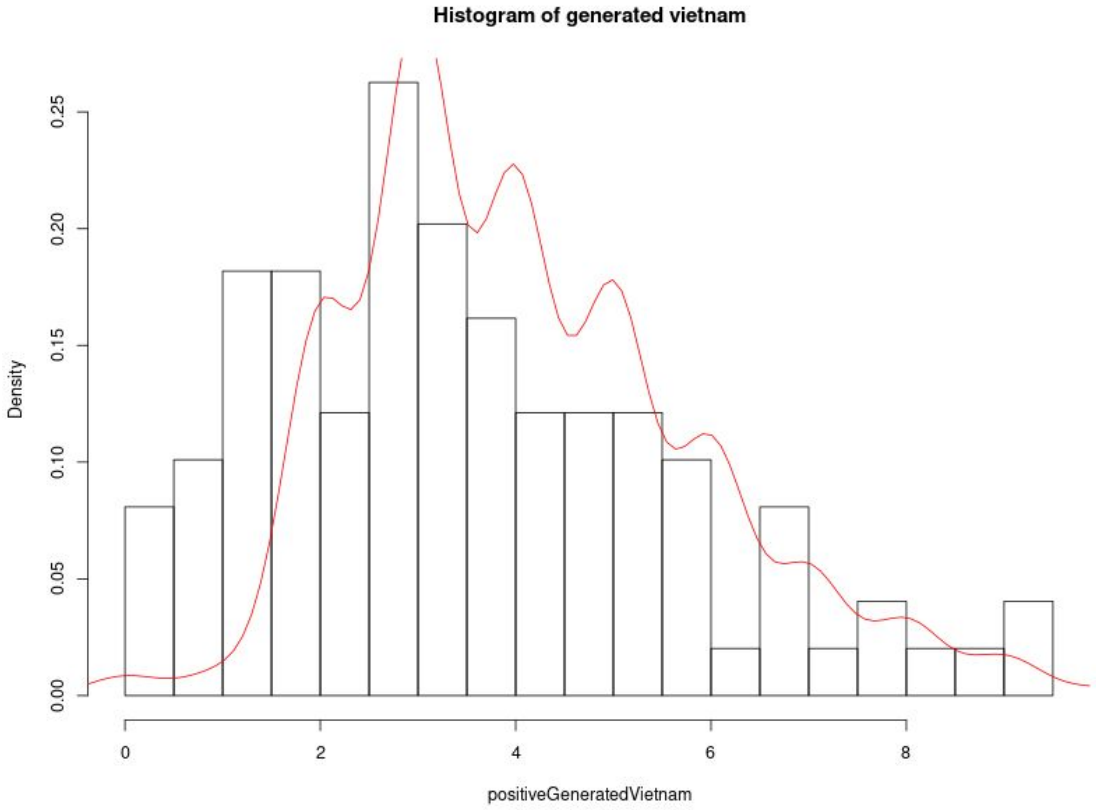
Pomocou odhadnutých parametrov strednej hodnoty (*Evietnam - 1)* a smerodatnej odchýľky *(sd(vietnam) - 0.5)* som vytvoril normálne rozdelenie a generoval som pomocou neho 100 kladných čísel.

```
n <- 0
positiveGeneratedVietnam = c()
while (n < 100 ){
  generatedVietnam <- rnorm(1, mean = Evietnam - 1, sd =
sd(vietnam) - 0.5)
  if (generatedVietnam > 0){
      positiveGeneratedVietnam[n] <- generatedVietnam
      n <- n + 1
  }
}
hist(positiveGeneratedVietnam, breaks = 15, freq=FALSE, main =
"Histogram of generated vietnam")
lines(density(vietnam), col='red')
```

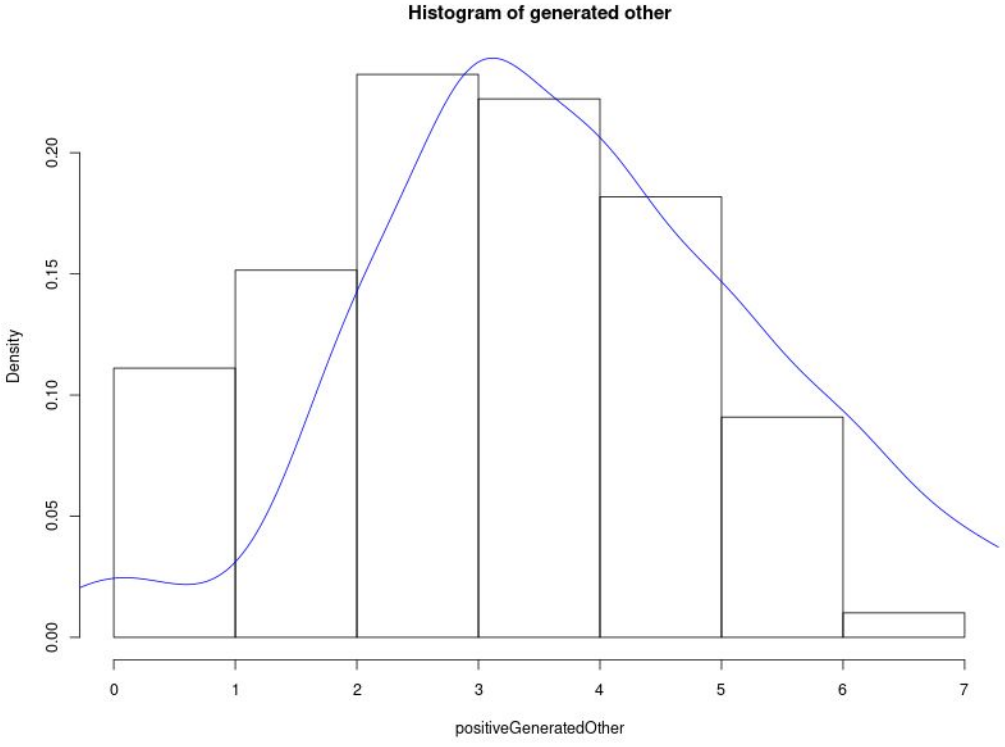
Tie som zobrazil v histograme a preložil som cez neho krivku hustoty pravdepodobnosti pôvodného datasetu **vietnam** resp. **other.**

Na grafe vidíme, že odhadnuté rozdelenie pomocou strednej hodnoty datasetu a smerodatnej odchýľky je *podobné* rozdeleniu podobného datasetu. Pri generovaní vyššieho počtu hodnôt by sa výsledok spresnil.

100 generovaných hodnot v histograme, červená krivka zobrazuje hustotu pravdepodobnosti datasetu **vietnam.**

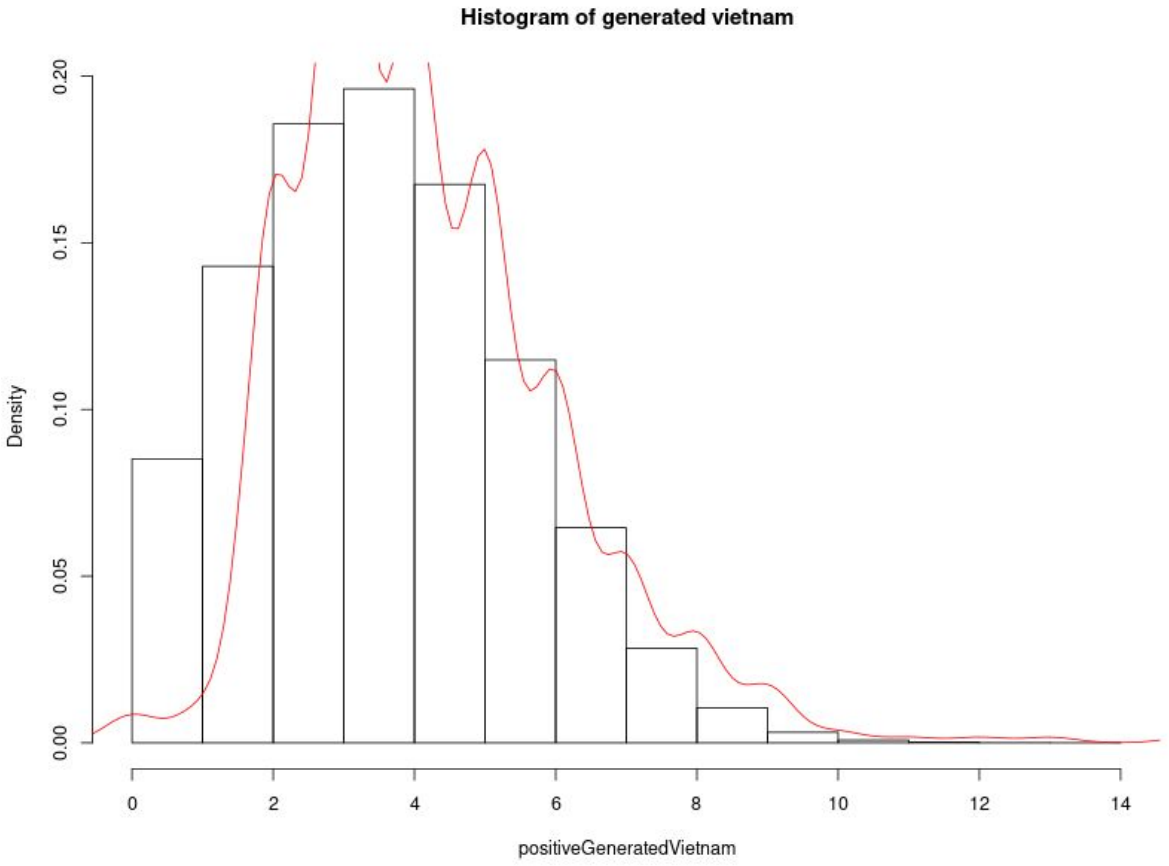


100 generovaných hodnot v histograme, modrá krivka zobrazuje hustotu pravdepodobnosti datasetu **other.**



Pre porovnanie:

100 000 generovaných hodnot v histograme, červená krivka zobrazuje hustotu pravdepodobnosti datasetu **vietnam**



Úloha č. 5 - Konfidenčný interval

Na výpočet oboustranného 95% konfidenčního intervalu pro střední hodnotu som použil vstavanú funkciu t.test.

```
confIntVietnam <- t.test(vietnam, conf.level = 0.95)$"conf.int"
confIntOther <- t.test(other, conf.level = 0.95)$"conf.int"
print("95 percent confidence interval: ")
print("VIETNAM")
print(confIntVietnam[1:2])
print("OTHER")
print(confIntOther[1:2])
```

```
[1] "95 percent confidence interval: "
[1] "VIETNAM"
[1] 4.055897 4.464227
[1] "OTHER"
[1] 3.721640 4.649494
```

Vzorec na výpočet intervalového odhadu som prevzal zo študijných materiálov pre predmet BI-PST.

Věta 8.2. *Uvažujme náhodný výběr* $X_1, \ldots, X_n$ *z normálního rozdělení* $\mathrm{N}(\mu, \sigma^2)$ *a předpokládejme, že známe rozptyl* σ^2*. Oboustranný* $100 \cdot (1-\alpha)\%$ *interval spolehlivosti pro* μ *je*

$$\left(\bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

kde $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ *je* kritická hodnota *standardního normálního rozdělení, tj. číslo, pro které platí* $\mathrm{P}(Z > z_{\alpha/2}) = \alpha/2$ *pro* $Z \sim \mathrm{N}(0,1)$.

Úloha č. 6 - Testovanie hypotézy

Našou úlohou bolo pre každú skupinu zvlášť otestujtovať na hladině významnosti 5% hypotézu, že je stredná hodnota (μ) rovná hodnote K (parameter úlohy), proti obojstrannej alternatíve.
Na testovanie hypotézy potrebujeme nájsť obojstranný 95% konfidenčný interval (CI). Použil som CI, ktoré boli výstupom predošlého bodu úlohy, keďže sa jedná o 95% CI pre μ.

Testujeme hypotézu $\mathbf{H_0}$**:** μ = K proti $\mathbf{H_A}$**:** μ ≠ K na hladine významnosti α = 5%.
K = 27

Hypotézu $\mathbf{H_0}$ zamietam pokiaľ μ ∉ CI, nezamietam pokiaľ μ ∈ CI.

Konfidenčné intervaly pre každú skupinu z predošlého bodu úlohy:
CI Vietnam = (4.055897, 4.464227)
CI Other = (3.721640, 4.649494)

V oboch prípadoch platí, že K ∉ CI, takže obe hypotézy $\mathbf{H_0}$ zamietam a prijímam alternatívne hypotézy $\mathbf{H_A}$.

Úloha č. 7 - Testovanie strednej hodnoty

Našou úlohou bolo na hladine spoľahlivosti 5% otestovať, či majú pozorované skupiny rovnakú strednú hodnotu.

Testujeme hypotézu $\mathbf{H_0}$: $\mu_v = \mu_o$ proti $\mathbf{H_A}$: $\mu_v \neq \mu_o$ na hladine spoľahlivosti α = 5%.
V tejto situácii je vhodné použiť dvojvýberový t-test. Predpokladáme, že sú obe veličiny (V = Vietnam, O = Other) nezávislé a normálne rozdelené. Existujú dva typy dvojvýberového t-testu na základe toho, či sa rozptyly veličín rovnajú alebo nerovnajú.

Rovnosť rozptylov som zistil pomocou funkcie na test rozptylov dvoch veličín:

```
var.test(vietnam,other)
```

```
F test to compare two variances
data:  vietnam and other
F = 1.318, num df = 645, denom df = 96, p-value = 0.09203
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.9551141 1.7566600
sample estimates:
ratio of variances
          1.317991
```

Dôležitá je výsledná p-hodnota testu, ktorá vyšla 0.09203. Naša hladina spoľahlivosti α = 0.05 a teda p-hodnota > α. Na hladine spoľahlivosti 5% teda hypotézu, že majú veličiny rovnaký rozptyl, nezamietame.

Ďalej teda použijeme dvojvýberový test pre veličiny s rovnakými rozptylmi. Použil som na to funkciu t.test:

```
t.test(vietnam,other,paired=F,alternative="two.sided",var.equal=T)
```

```
Two Sample t-test
data:  vietnam and other
t = 0.26302, df = 741, p-value = 0.7926
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4815229  0.6305128
sample estimates:
mean of x mean of y
 4.260062  4.185567
```

Výsledná p-hodnota = 0.7926 je väčšia ako α = 0.05 a hypotézu $\mathbf{H_0}$, že pozorované skupiny majú rovnakú strednú hodnotu na hladine spoľahlivosti 5% nezamietame.