Délka humeru podle přežití vrabců

Vojtěch Tóth, Cuphead, Mugman

2022-12-08

```
  K = 4
  L = 4
  M = ((K+L)*47)%%11+1.
```

```
## [1] 3
```

Úloha 1

(1) Načtěte datový soubor a rozdělte sledovanou proměnnou na příslušné dvě pozorované skupiny. Stručně popište data a zkoumaný problém. Pro každu skupinu zvlášť odhadněte střední hodnotu, rozptyl a medián příslušného rozdělení.

Dataset, který budeme v tomto úkolu zpracovávat, je case0201. Tento dataset obsahuje 59 záznamů dvou proměnných

- Humerus - délka kosti pažní vrabců (v mili-palcích)
- Status - zda vrabec přežil(“survived”), či zahynul(“Perished”)

Data nasbíral H. Bumpus. Zkoumal, zda uhynulí vrabci postrádají některé fyzické vlastnosti oproti těm, kteří přežili a tím chtěl podpořit teorii přirozeného výběru.

Proměnnou Humerus rozdělíme do dvou skupin podle stavu.

```
library(Sleuth2)
perished <- subset(case0201, Status=="Perished")$Humerus
survived <- subset(case0201, Status=="Survived")$Humerus
```

Ve skupině Uhynulích máme 24 hodnot, ve skupině přeživších 35.

```
str(perished)
```

```
##  num [1:24] 659 689 703 702 709 713 720 729 726 726 ...
```

```
str(survived)
```

```
##  num [1:35] 687 703 709 715 728 721 729 723 728 723 ...
```

Vzorce pro výběrový průměr, rozptyl a pro medián jsou popořadě

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

,

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

,

$$\text{med}(X) = \begin{cases} a_{\lfloor \frac{n}{2} \rfloor} & \text{pokud } n\%2 = 1 \\ \frac{a_{\lfloor \frac{n}{2} \rfloor} + a_{\lceil \frac{n}{2} \rceil}}{2} & \text{pokud } n\%2 = 0 \end{cases}$$

,

my použijeme následující funkce.

```
mean_per <- mean(perished)
var_per <- var(perished)
med_per <- median(perished)

mean_sur <- mean(survived)
var_sur <- var(survived)
med_sur <- median(survived)
```

Výsledné hodnoty jsou v této tabulce.

	Přeživší	Uhynulí
Výběrový průměr	727.9166667	738
Výběrový rozptyl	554.2536232	393.5882353
Medián	733.5	736

Úloha 2

(1b) Pro každou skupinu zvlášť odhadněte hustotu a distribuční funkci pomocí histogramu a empirické distribuční funkce.

Empirická distribuční funkce je definována jako

$$F_n(x) = F_n(x, X_1, ..., X_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}}$$

Tedy pro reálnou proměnnou x zjistíme počet hodnot x_i, které jsou menší nebo rovny x a podělíme je počtem všech záznamů n dané skupiny. V jazyce R se pro výpočet hodnot používá funkce ecdf, výstup vykreslí funkce plot.

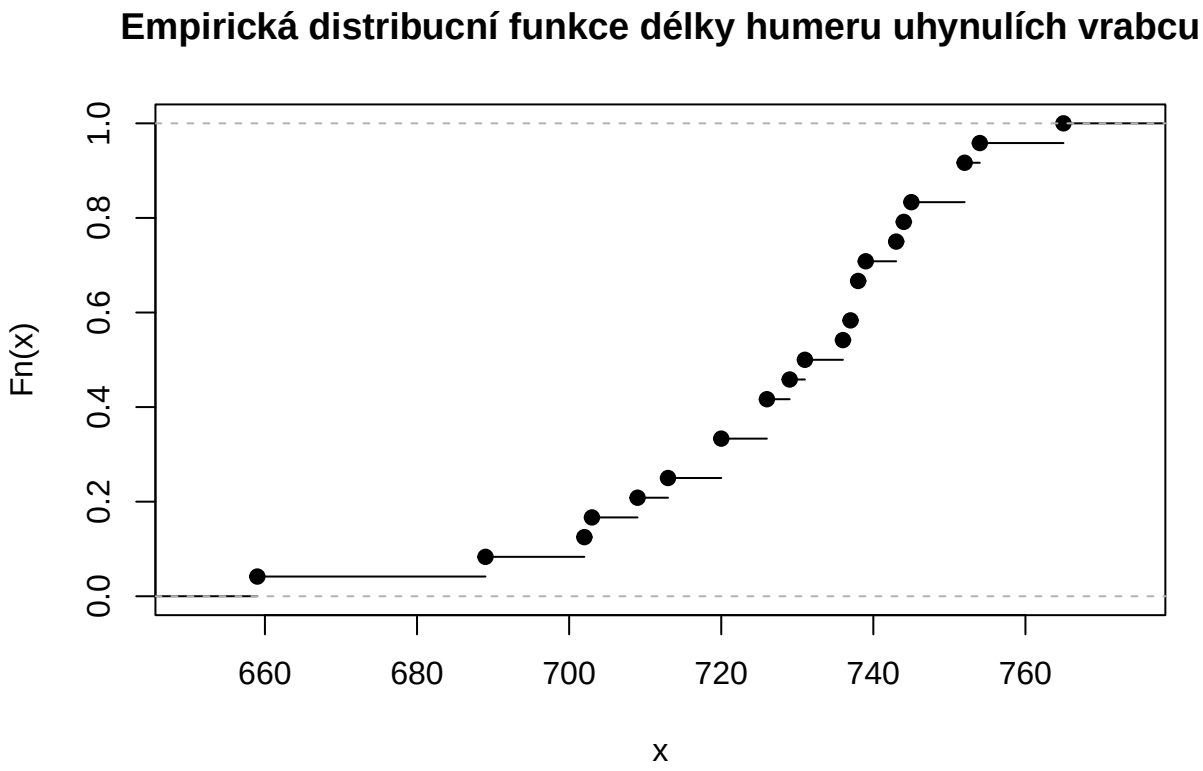
Histogram je sloupcový graf, kde každý sloupec má zvolenou nějakou vhodnou šířku. Výška daných sloupců se získá ze vztahu

$$\frac{m_i}{n \cdot h} = \frac{\text{počet hodnot uvnitř sloupce}}{\text{počet všech hodnot} \cdot \text{šířka sloupce}}$$

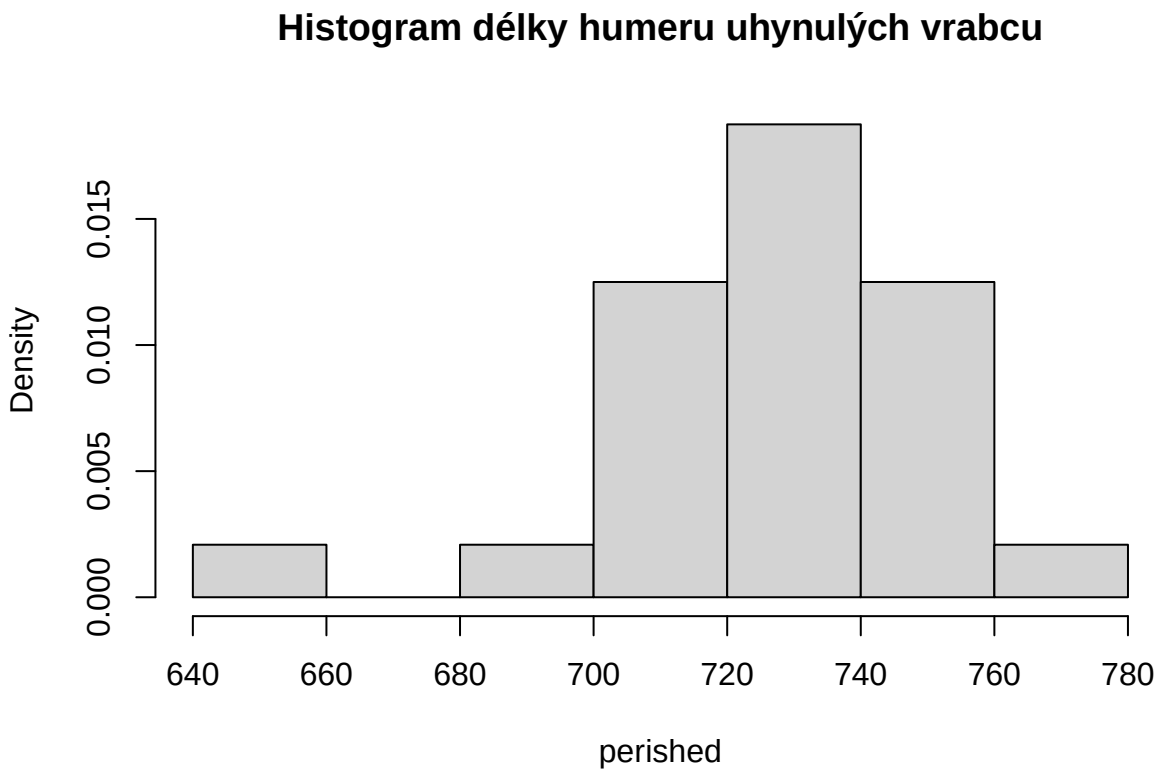
Funkce hist z daných hodnot zvládne odhadnout nejlepší šířku sloupce a rovnou histogram vykreslí.

Empirická distribuční funkce a histogram skupiny uhynulích

```
plot(ecdf(perished), main = "Empirická distribuční funkce délky humeru uhynulích vrabců")
```



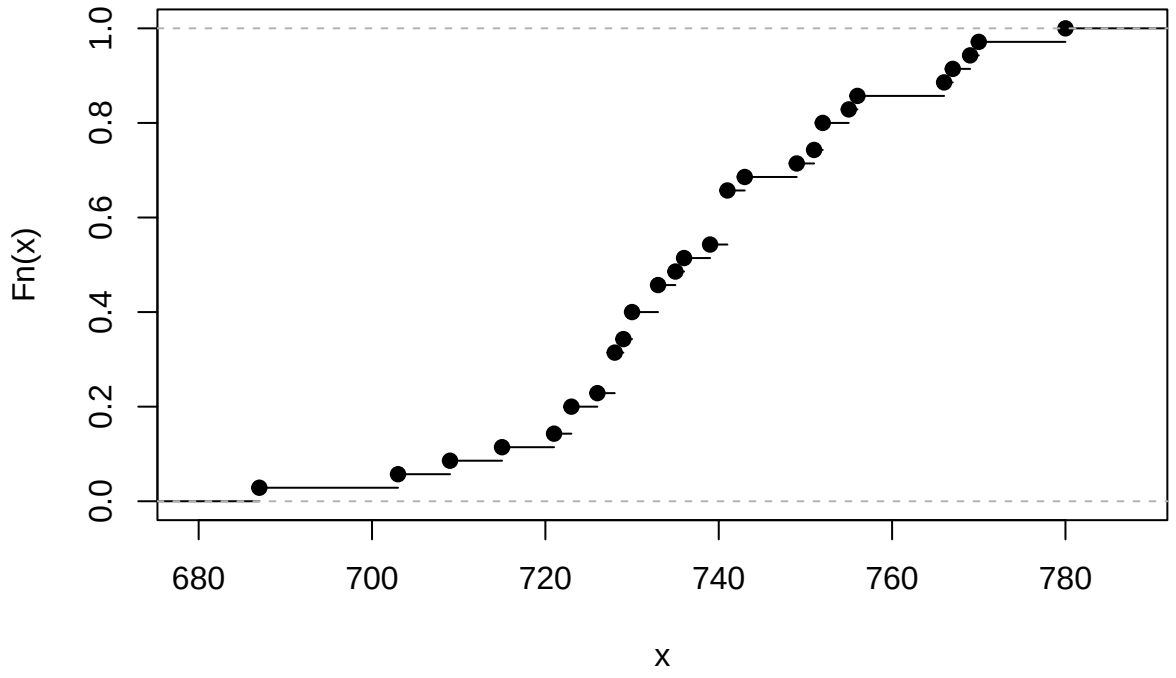

```
hist(perished, freq = FALSE, main = "Histogram délky humeru uhynulých vrabců" )
```



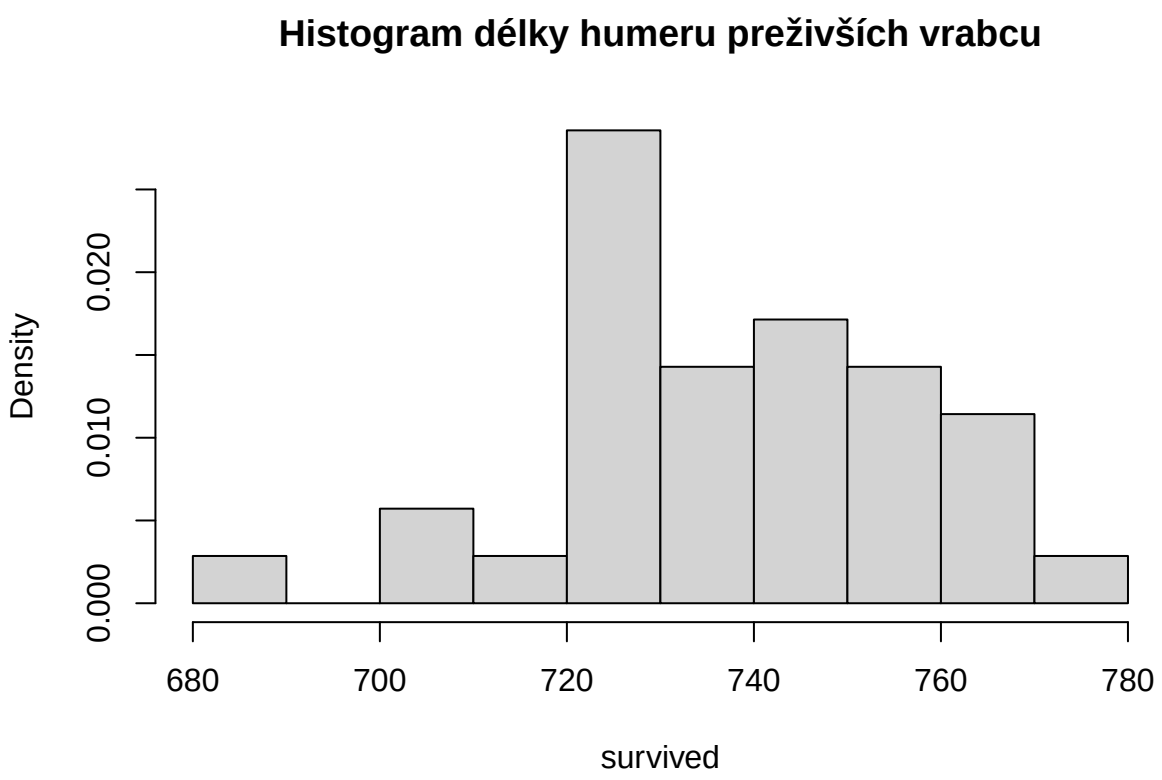
Lze tvrdit, že délka humeru uhynulých vrabců se řídí normálním rozdělením.

Empirická distribuční funkce a histogram skupiny přeživších

```
plot(ecdf(survived), main = "Empirická distribuční funkce délky humeru přeživších vrabců")
```




```
hist(survived, freq = FALSE, main = "Histogram délky humeru přeživších vrabců")
```



Lze tvrdit, že i délka humeru přeživších vrabců se řídí normálním rozdělením.

Úloha 3

> (3b) Pro každou skupinu zvlášť najděte nejbližší rozdělení: Odhadněte parametry normálního, exponenciálního a rovnoměrného rozdělení. Zaneste příslušné hustoty s odhadnutými parametry do grafů histogramu. Diskutujte, které z rozdělení odpovídá pozorovaným datům nejlépe.

Pro odhad parametrů lze použít balíček `EnvStats`.

```
library(EnvStats, warn.conflicts=F, quietly=T)
```

Normální rozdělení

Exponenciální rozdělení

Exponenciální rozdělení je takové rozdělení, při kterém události mají nezávislé exponenciální časy mezi sebou. Pro hustotu platí

$$f_n(x) = \begin{cases} \lambda e^{-\lambda x} & \text{pro } x \in (0, \infty) \\ 0 & \text{jinde} \end{cases}$$

Hledáme tedy odhad parametru λ. Použijeme funkce `eexp`, u které zvolíme momentovou metodu a metodu maximální věrohodnosti. Získané parametry jsou prvky pole parameters, λ na indexu 1.

```r
exp_perished <- eexp(perished, method="mle/mme")$parameters
exp_survived <- eexp(survived, method="mle/mme")$parameters

lambda_per <- exp_perished[1]
lambda_sur <- exp_survived[1]
```

Výsledné odhady:

	Přeživší	Uhynulí
λ	0.001355	0.0013738

Rovnoměrné rozdělení

Rovnoměrné rozdělení je takové rozdělení, které má v nějakém intervalu (a, b) konstatní pravděpodobnost, mimo něj je pravděpodobnost rovna 0. Pro hustotu platí

$$f_n(x) = \begin{cases} \frac{1}{b-a} & \text{pro } x \in (a, b) \\ 0 & \text{pro } x \notin (a, b) \end{cases}$$

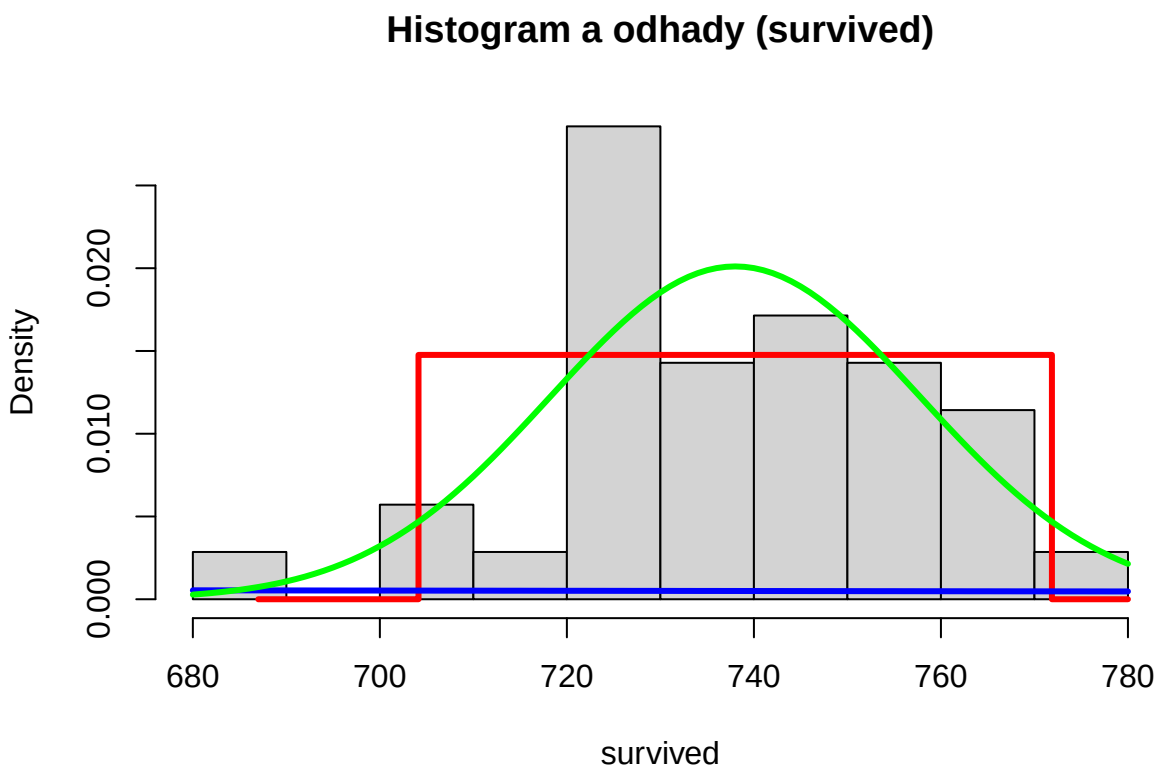
Hledáme tedy odhad parametrů a a b. Použijeme funkci `eunif` u které zvolíme momentovou metodu. Získané parametry jsou prvky pole parameters, a na indexu 1, b na indexu 2.

```r
unif_perished <- eunif( perished, method="mme")$parameters
unif_survived <- eunif( survived, method="mme")$parameters

a_per <- unif_perished[1]
b_per <- unif_perished[2]

a_sur <- unif_survived[1]
b_sur <- unif_survived[2]
```

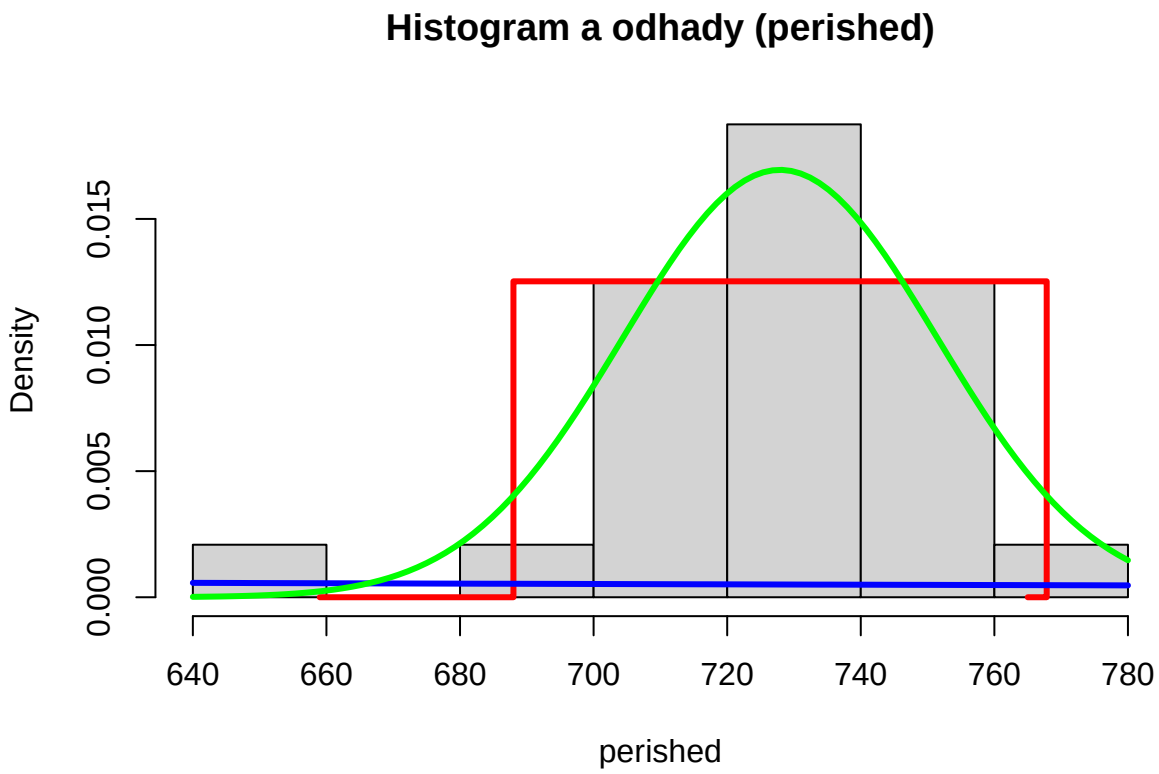
Výsledné odhady:

	Přeživší	Uhynulí
a	704.1321897	687.9982603
b	771.8678103	767.835073

```r
hist(survived, freq = FALSE, main = "Histogram a odhady (survived)")
min <- min(survived)
max <- max(survived)
x <- c( min, a_sur, a_sur, b_sur, b_sur, max )
p <- dunif(a_sur, min=a_sur, max=b_sur)
y <- c( 0, 0, p, p, 0, 0)
lines(x, y, col="red", lwd=3)

x <- survived
curve(dexp(x, rate = lambda_sur), lwd= 3, col = "blue", add = TRUE)

curve(dnorm(x, mean = mean_sur, sd = sqrt(var_sur)), col="green", add = TRUE, lwd=3)
```



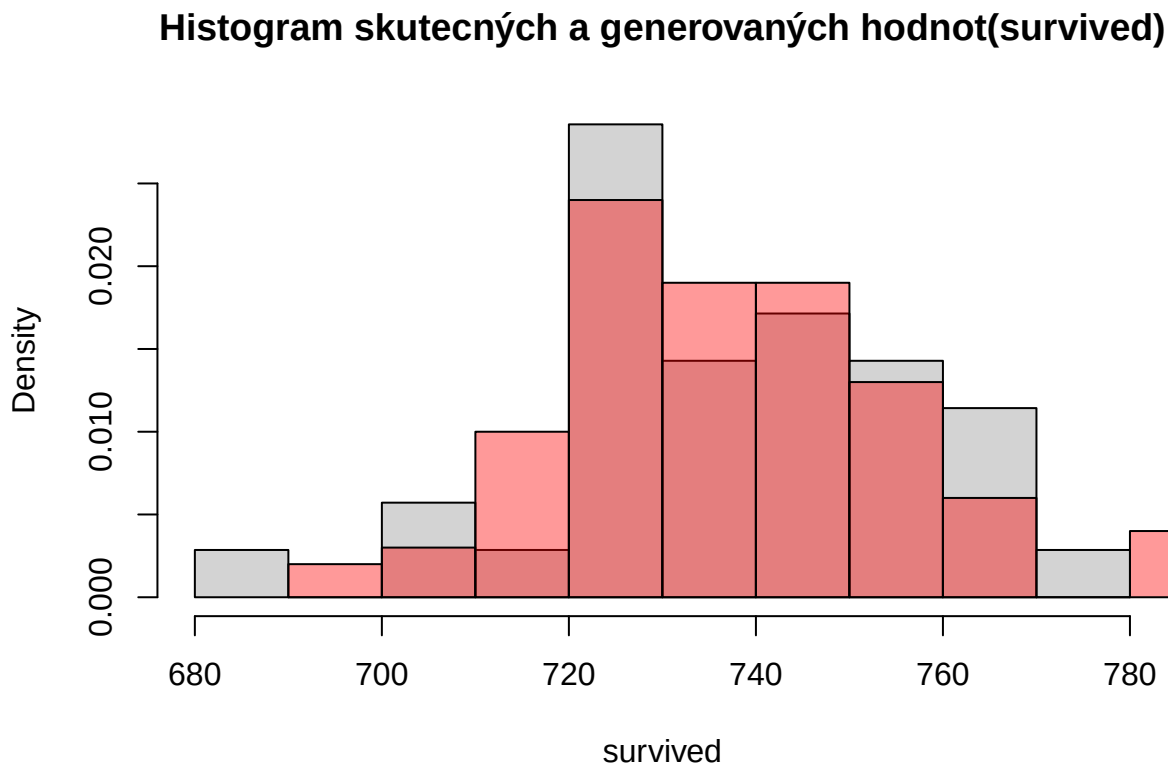

```
min <- min(perished)
max <- max(perished)
hist(perished, freq = FALSE, main = "Histogram a odhady (perished)")
x <- c( min , a_per, a_per, b_per, b_per, max )
p <- dunif(a_per, min=a_per, max=b_per)
y <- c(0, 0, p, p, 0, 0 )
lines(x, y, col="red", lwd=3)

x <- perished
curve(dexp(x, rate = lambda_per), lwd= 3, col = "blue", add = TRUE)

curve(dnorm(x, mean = mean_per, sd = sqrt(var_per)), col="green", add = TRUE, lwd=3)
```



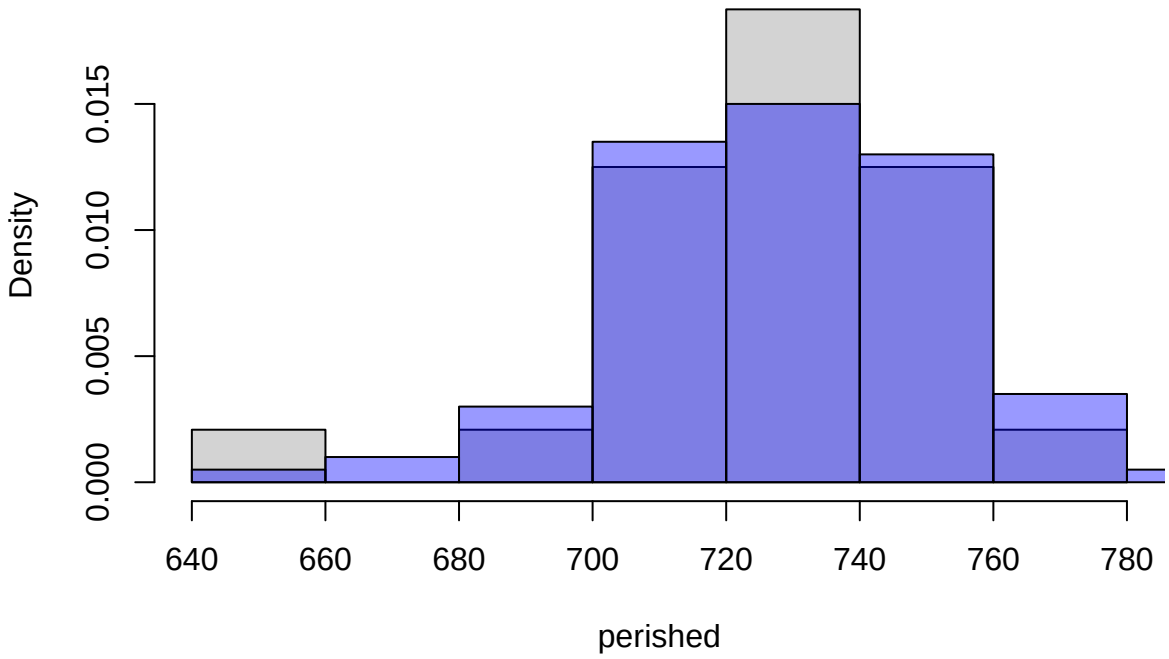
Úloha 4

(1b) Pro každou skupinu zvlášť vygenerujte náhodný výběr o 100 hodnotách z rozdělení, které jste zvolili jako nejbližší, s parametry odhadnutými v předchozím bodě. Porovnejte histogram simulovaných hodnot s pozorovanými daty.

```
data <- rnorm( 100, mean = mean_sur, sd=sqrt(var_sur))
hist(survived, freq = FALSE, main = "Histogram skutečných a generovaných hodnot(survived)")
hist(data, freq=FALSE, add=TRUE, col = rgb( 1, 0, 0, 0.4))
```




```
data <- rnorm( 100, mean = mean_per, sd=sqrt(var_per))
hist(perished, freq = FALSE, main = "Histogram skutečných a generovaných hodnot (perished)")
hist(data, freq=FALSE, add=TRUE, col = rgb( 0, 0, 1, 0.4))
```



Úloha 5

(1b) Pro každou skupinu zvlášť spočítejte oboustranný 95% konfidenční interval pro střední hodnotu.

```
model <- lm(survived ~ 1)
interval <- confint(model, level=0.95)

L_bound_sur = interval[1]
U_bound_sur = interval[2]
```

```
model <- lm(perished ~ 1)
interval <- confint(model, level=0.95)

L_bound_per = interval[1]
U_bound_per = interval[2]
```

	dolní mez	horní mez
Přeživší	731.185045	744.814955
Uhynulí	717.9755021	737.8578313

Úloha 6

(1b) Pro každou skupinu zvlášť otestujte na hladině významnosti 5 % hypotézu, zda je střední hodnota rovná hodnotě K (parametr úlohy), proti oboustranné alternativě. Můžete použít buď výsledek z předešlého bodu, nebo výstup z příslušné vestavěné funkce vašeho softwaru.

```
t.test( survived, mu = K, conf.level = 0.95, alternative = "two.sided" )
```

```
##
##  One Sample t-test
##
## data:  survived
## t = 218.88, df = 34, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 4
## 95 percent confidence interval:
##  731.185 744.815
## sample estimates:
## mean of x
##       738
```

```
t.test( perished, mu = K, conf.level = 0.95, alternative = "two.sided" )
```

```
##
##  One Sample t-test
##
## data:  perished
## t = 150.64, df = 23, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 4
## 95 percent confidence interval:
##  717.9755 737.8578
## sample estimates:
## mean of x
##  727.9167
```

Úloha 7

(2b) Na hladině významnosti 5 % otestujte, jestli mají pozorované skupiny stejnou střední hodnotu. Typ testu a alternativy stanovte tak, aby vaše volba nejlépe korespondovala s povahou zkoumaného problému.

Použijeme oboustranný dvouvýběrový t-test.

```
t.test( survived, perished, paired=FALSE, conf.level = 0.95,
        alternative = "two.sided", var.equal = TRUE )
```

```
##
##  Two Sample t-test
##
## data:  survived and perished
## t = 1.777, df = 57, p-value = 0.0809
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
##  -1.279386 21.446053
## sample estimates:
## mean of x mean of y
##  738.0000  727.9167
```

Z výstupu funkce vidíme, že 95% konfidenční interval rozdílu středních hodnot obou skupin je (-1.279386, 21.446053) a p-hodnota (Minimální hladina významnosti, na které lze hypotézu H_0 při dané realizaci náhodného výběru zamítnout) rovná 0.0809 (8%). Hladina významnosti byla stanovena na 5%, rovnost středních hodnot tedy nezamítáme.