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 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.

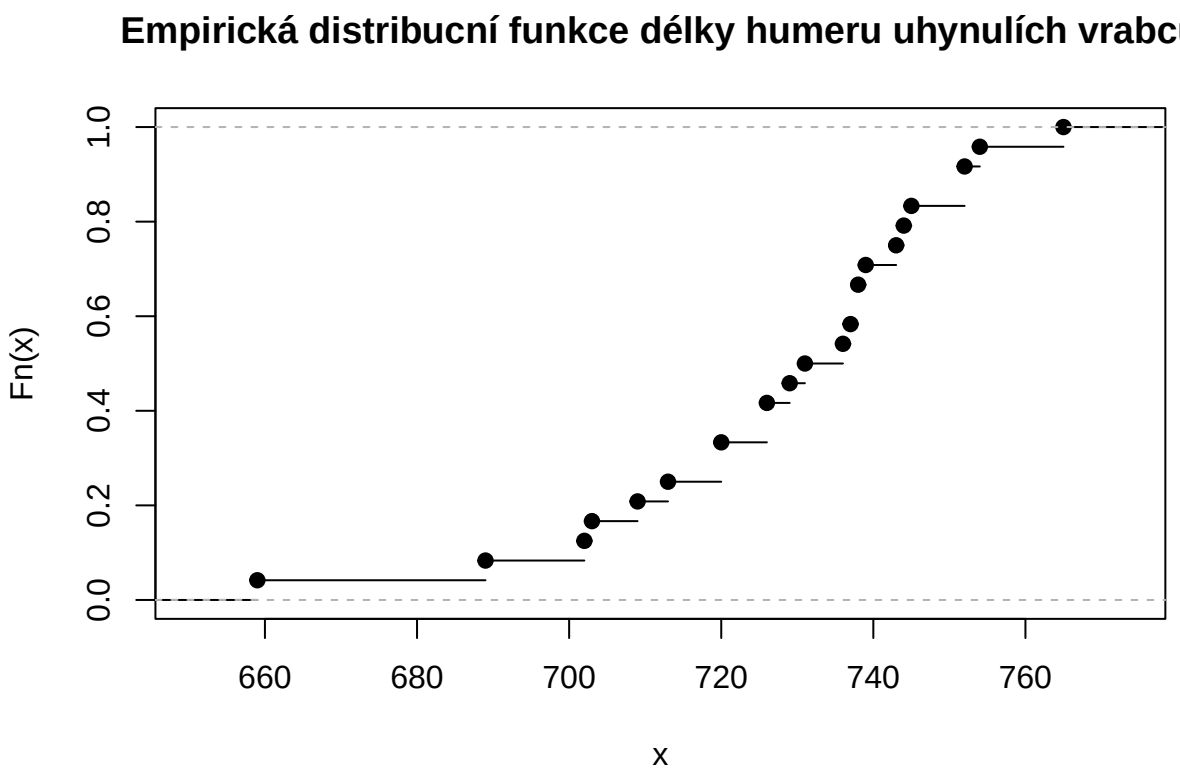
Histogram je sloupcový graf, kde každý sloupec má zvolenou nějakou vhodnou šířku. Výška daných sloupců se získá ze vztahu

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

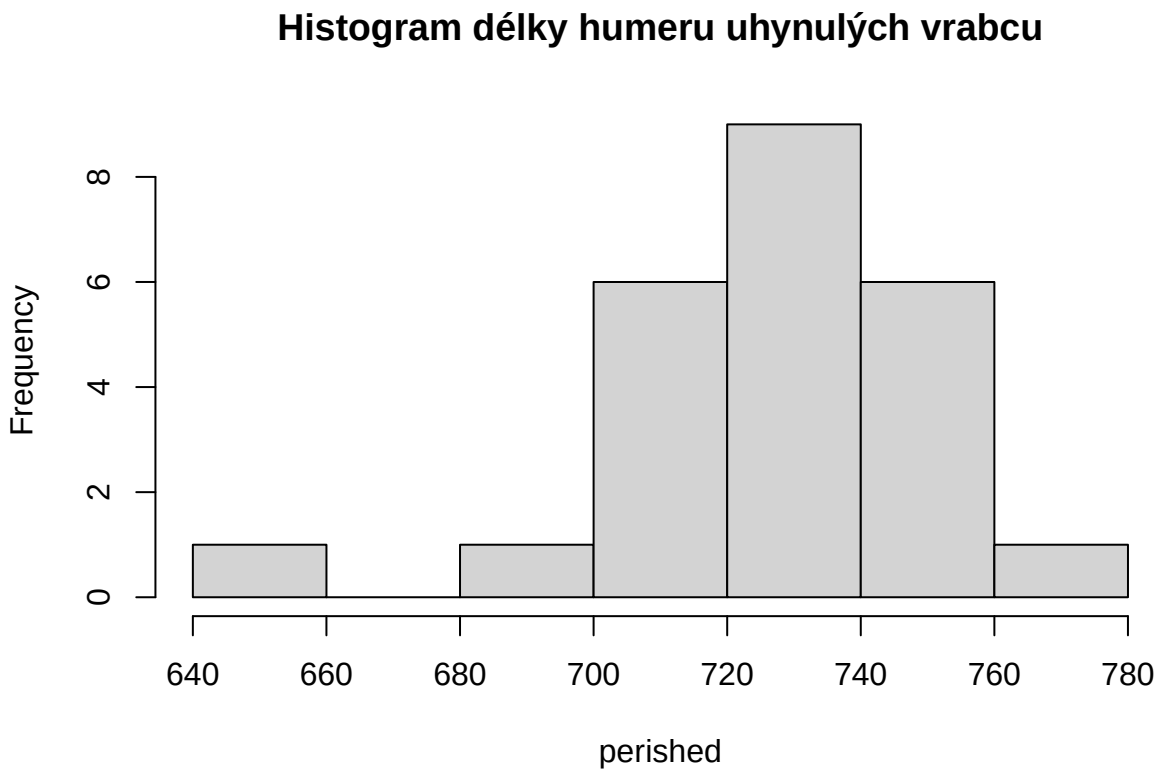
Funkce hist z daných hodnot zvládne odhadnout nejlepší šířku sloupce a rovnou histogram vykreslí.

Empirická distribuční funkce a histogram skupiny uhynulích

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



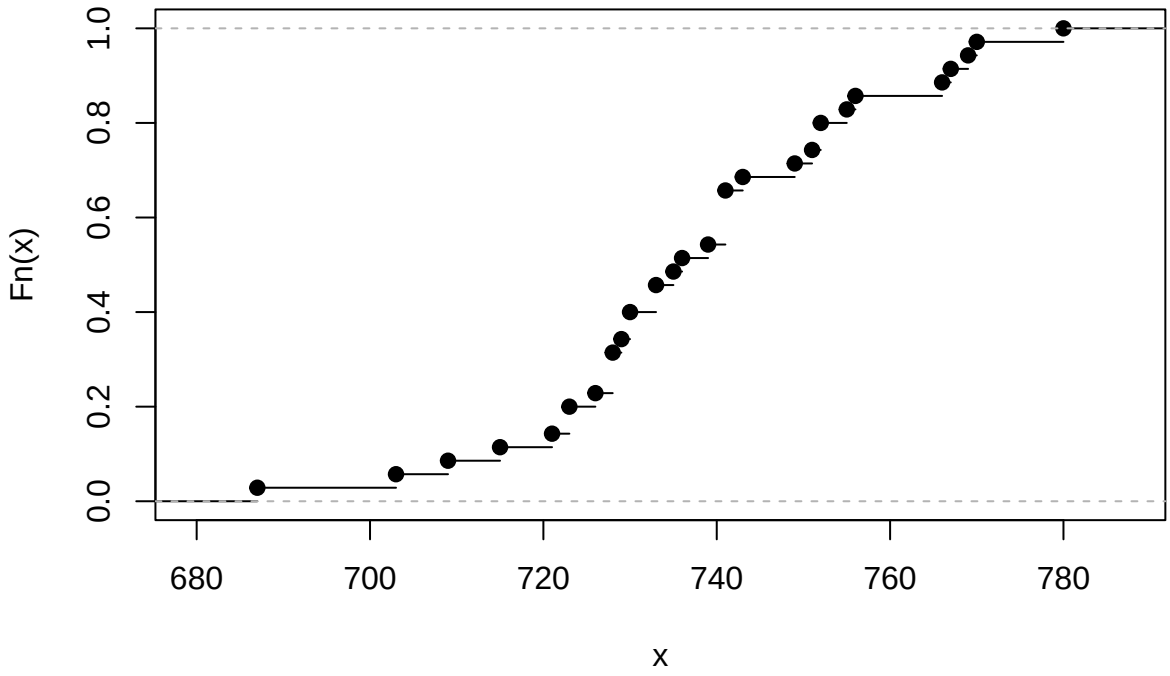

```
hist(perished, main = "Histogram délky humeru uhynulých vrabců" )
```



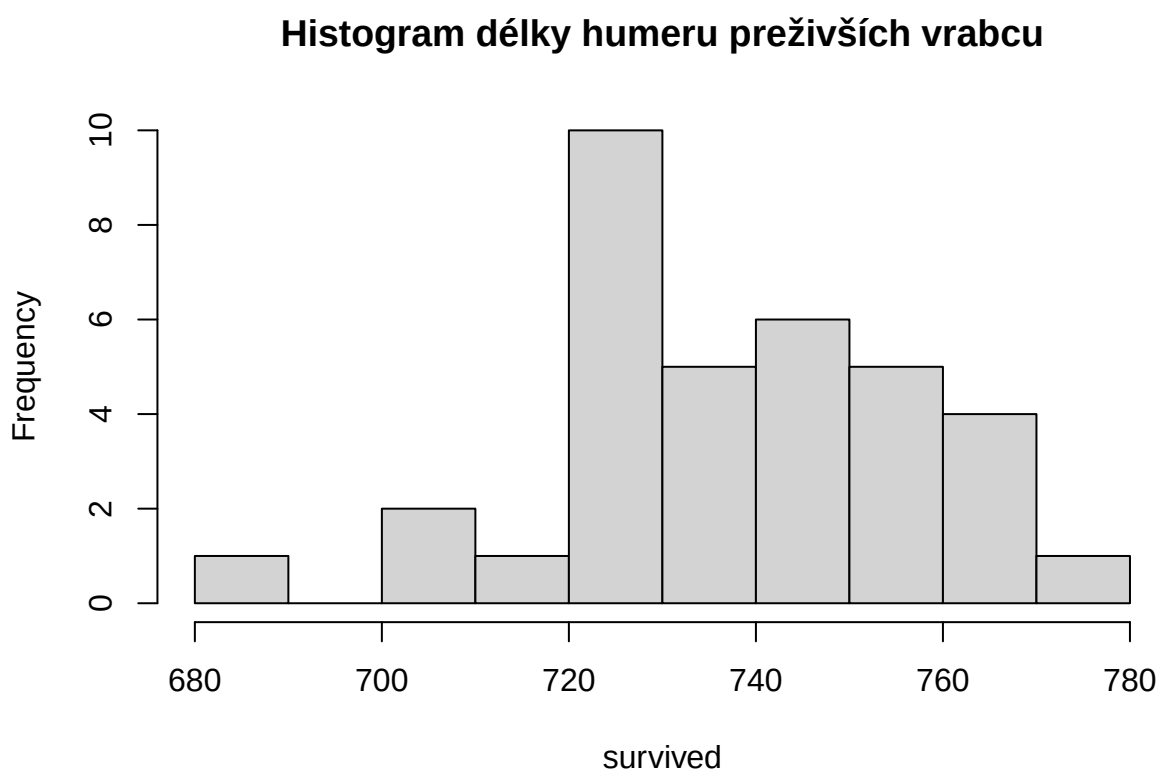
Lze tvrdit, že délka humeru uhynulých vrabců se řídí normálním rozdělením.

Empirická distribuční funkce a histogram skupiny přeživších

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




```
hist(survived, 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.

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.

Úloha 5

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

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.

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.